Essays on statistical inference in non-regular semiparametric models

Adam Lee

TESI DOCTORAL UPF / Year 2022

THESIS SUPERVISOR

Professor Geert Mesters

Department of Economics and Business

upf. Universitat Pompeu Fabra Barcelona

Acknowledgements

First and foremost I would like to take this opportunity to thank my advisor, Geert Mesters, for his continual advice and support throughout the PhD. I particularly benefitted from the freedom he gave me in choosing topics and the relaxed style of his supervision. I hope that we will be able to continue to work together in the future.

Secondly, I am indebted to the (past and present) Econometrics and Statistics groups at UPF for all they have taught me and their valuable advice along the way. I have benefitted enormously from classes given by, discussions with and advice from: Majid Al Sadoon, Christian Brownlees, Kirill Evdokimov, Gábor Lugosi, Katerina Petrova, David Rossell, Barbara Rossi and Piotr Zwiernik.

Thirdly, I owe a debt of graditude to my officemates: Bjarni G. Einarsson, Rui Guan, Lukas Hoesch, André B. M. Souza and Philipp Tiozzo. I thank them for making it a pleasure to come into work each and every day, in addition to our many valuable discussions, whether academic or otherwise.

Finally, I would like to thank Marta Araque and Laura Agustí for their expert guidance through labyrinthine bureaucracy and their unparalled ability to solve any problem they were confronted with on my behalf.

Barcelona, 2022

Abstract

This thesis consists of three chapters which relate to problems of statistical inference in (potentially) non-regular semiparametric models. Chapter 1 considers hypothesis testing problems in semiparametric models which may be non-regular for certain values of a potentially infinite dimensional nuisance parameter. I establish that, under mild regularity conditions, tests based on the efficient score function provide locally uniform size control and enjoy minimax optimality properties. Two examples are studied in some detail. Chapter 2 applies the methodology of Chapter 1 to the case of (static) linear simultaneous equations models. Existing inference methods that exploit non-Gaussianity to identify structural parameters in such models suffer from size distortions when the structural shocks are close to Gaussian. The approach proposed herein yields valid inference for the structural parameters of interest regardless of the distance to Gaussianity. An application to production function estimation is presented. Chapter 3 develops a semi-parametric approach to conduct inference in non-Gaussian SVAR models robust to "weak" non-Gaussianity based on the ideas in Chapter 1. The method exploits non-Gaussianity when it is present, while yielding correct coverage regardless of the distribution of the structural errors. Two empirical applications are presented.

Resumen

Esta tesis consta de tres capítulos que se relacionan con problemas de inferencia estadística en modelos semi-paramétricos potencialmente irregulares. El capítulo 1 considera problemas con hipótesis en modelos semi-paramétricos que podrían ser irregulares para ciertos valores de un parámetro de molestia de dimensional infinita. Establezco que, en condiciones de regularidad leve, pruebas basadas en la función de puntuación eficiente proporcionan un control de tamaño localmente uniforme y son óptimas en un sentido minimax. Dos ejemplos se estudian en detalle. El capítulo 2 aplica la metodología del Capítulo 1 al caso de modelos de ecuaciones lineales simultáneas estáticas. Los métodos de inferencia existentes que explotan la no Gaussianidad para identificar parámetros estructurales en tales modelos sufren distorsiones de tamaño cuando los choques estructurales están cerca de Gaussian. El enfoque propuesto en este capítulo produce una inferencia válida para los parámetros estructurales de interés, independientemente de su distancia a la Gaussianidad. Se presenta una aplicación para la estimación de funciones de producción. El capítulo 3 desarrolla un enfoque semi-paramétrico para realizar inferencias en modelos SVAR no Gaussianos robustos a la no Gaussianidad "débil" basada en las ideas del Capítulo 1. El método explota la no Gaussianidad cuando está presente y a su vez que brinda una cobertura correcta independientemente de la distribución de errores estructurales. Se presentan dos aplicaciones empíricas.

Preface

This thesis consists of three interdependent chapters.

The first chapter, titled “Robust and efficient inference for non-regular semiparametric models”, considers hypothesis testing problems in semiparametric models which may be non-regular for certain values of a potentially infinite dimensional nuisance parameter. I establish that, under mild regularity conditions, tests based on the efficient score function provide locally uniform size control and enjoy minimax optimality properties. This approach is applicable to situations with (i) identification failures, (ii) boundary problems and (iii) distortions induced by the use of regularised estimators. Full details are worked out for two examples: a single index model where the link function may be relatively flat and a linear simultaneous equations model that is (weakly) identified by non-Gaussian errors. In practice the tests are easy to implement and rely on χ^2 critical values.

The second chapter, titled “Robust inference for non-Gaussian linear simultaneous equations models”, expands on the potential weak identification problem in non-Gaussian (static) simultaneous equations models. In particular, all parameters in linear simultaneous equations models can be identified (up to permutation and scale) if the underlying structural shocks are independent and if at most one of them is Gaussian. Unfortunately, existing inference methods that exploit such identifying assumptions suffer from size distortions when the true distributions of the shocks are close to Gaussian. To address this weak non-Gaussianity problem, the chapter develops a robust semi-parametric inference method (based on the methodology outlined in Chapter 1) that yields valid confidence intervals for the structural parameters of interest regardless of the distance to Gaussianity. The densities of the structural shocks are treated non-parametrically and construct identification robust tests based on the efficient score function. The finite sample properties of the methodology are illustrated in a large simulation study and an empirical study for production function estimation.

The third chapter, titled “Robust inference in structural VAR models identified by non-Gaussianity” considers a dynamic version of the model considered in Chapter 2. As

in the static case, standard methods that exploit non-Gaussian distributions to identify structural functions in SVAR models are not robust when deviations from Gaussianity are small, leading to confidence bands with incorrect coverage. A robust semi-parametric approach to conduct hypothesis tests and compute confidence bands in the SVAR model is proposed. The method exploits non-Gaussianity when it is present, but yields correct coverage regardless of the distance to the Gaussian distribution. The performance of the method is evaluated in a simulation study and two empirical studies are revisited.

Contents

List of Figures

List of Tables

Chapter 1

Robust and efficient inference for non-regular semiparametric models

1.1. Introduction

In many econometric models, the behaviour of commonly used inference procedures can depend crucially on the value of nuisance parameters. There are many cases where the asymptotic distributions of test statistics derived using standard (fixed parameter) arguments provide poor approximations to the finite sample distribution for certain values of nuisance parameters. When this occurs, the corresponding tests justified by such asymptotic arguments may have (finite sample) size far in excess of the nominal level.

In this paper I develop a general framework for conducting inference on a finite dimensional parameter in a semiparametric model, robust to (sequences of) values of a possibly infinite dimensional nuisance parameter which may invalidate standard inference methods. In particular, the main contribution of this paper is to show that semiparametric score tests based on the efficient score function (e.g. Bickel et al., 1998; van der Vaart, 2002) are robust under mild assumptions which allow for, among others, (i) identification failure, (ii) nuisance parameters on the boundary and (iii) the use of regularised estimates of nuisance parameters.

Importantly – and unlike other general approaches put forward in the robust inference literature (e.g. Andrews and Guggenberger, 2009, McCloskey, 2017 and Elliott et al., 2015) – this approach permits the nuisance parameter which causes standard inferential approaches to break down to be *infinite dimensional*.

A key benefit of this approach is that this efficient score test does not sacrifice power in order to obtain this robustness: when classical regularity conditions hold, the test enjoys classical optimality properties. Additionally, I demonstrate that the test is minimax optimal in some cases which fall in-between classical regularity conditions and the weaker conditions under which the robustness results of this paper are obtained. Such results apply, for example, when the parameter of interest is underidentified. Moreover these tests are often easy to compute and require only χ^2 critical values.

The semiparametric models I consider are parametrised by a pair $\gamma = (\theta, \eta)$ where θ is the parameter of interest and η collects all nuisance parameters (and is therefore typically infinite dimensional). γ fully parametrises the distribution of the observed data and I write the corresponding probability law as P_γ. This setup permits a large range of models regularly used in practice and includes both traditional parametric models and models defined by moment conditions as special cases.

The theoretical results of this paper are derived under a few high level conditions, for which some more primitive conditions are given subsequently. The main condition is local asymptotic normality (LAN) of the model, which implicitly defines score functions for θ and η. LAN specifies that the logarithms of certain likelihood ratios posses a local quadratic approximation and – in the i.i.d. case considered in this paper – can be demonstrated to hold under an L_2-differentiability condition known as "differentiability in quadratic mean" (DQM).[1] Such conditions are common in the semiparametric statistical theory as expounded by e.g. Bickel et al. (1998) or van der Vaart (2002).[2] This literature usually complements LAN (or DQM) with additional regularity conditions, such as (a) the non-singularity of information matrices and (b) all parameters lying in the interior of the parameter space.[3] These conditions rule out a number of cases of interest in econometrics. For example, (a) the non-singularity of the information matrix is often violated when the parameter of interest is under- or un-identified; (b) many model specifications permit nuisance parameters to lie on the boundary. Fortunately, as I show in this paper, valid inference can be conducted without these additional conditions.[4]

With the LAN condition in hand, the *efficient score function* (for the parameter of interest) can be defined as the orthogonal projection (in L_2) of the score function for θ on the orthocomplement of the set of score functions for η. This efficient score function is the basis of the robust inferential theory put forward in this paper. The main test statistic I consider, the *efficient score statistic*, is the quadratic form of an estimate of the efficient score function, weighted by a (pseudo-)inverse of its (estimated) variance matrix.[5] The

[1] See e.g. Le Cam and Yang (2000, Chapters 6 and 7).

[2] Similar quadratic expansions of an objective function have also been previously used to analyse nonstandard models in econometrics. See, for instance, Andrews (2001); Andrews and Cheng (2012).

[3] Cf. e.g. Definitions 2.1.1, 2.1.2 and 3.1.1 of Bickel et al. (1998).

[4] Cf. section 6.9 of Le Cam and Yang (2000) where the authors explicitly discuss a number of simplifying assumptions which are often made but are not essential. Their point (v), that "the points ... are interior points of $\Theta \in \mathbb{R}^k$" is clearly directly relevant to the case (b) with parameters potentially on the boundary. For (a), where un- or under-identification of the parameter of interest may cause singularity of the information matrix, cf. Le Cam and Yang, 2000, example (a), pp. 56 - 57.

[5] When the variance matrix is non-singular, the corresponding *efficient score test* is the same as the "effective score test" of Choi et al. (1996). Additionally, the efficient score statistic can be viewed as the semiparametric analogue of Neyman's C(α) statistic (Neyman, 1959, 1979).

key insight I exploit is that – under the null – the limiting distribution of the efficient score function is the same regardless of the (local) nuisance parameter sequence along which the limit is taken. This directly leads to robustness of the efficient score test against such sequences and consequently that such tests control size in a (locally) uniform manner over certain compact subsets. In contrast, there are many models in which this property fails to hold for commonly used test statistics: different sequences of nuisance parameters consistent with the null hypothesis result in different limiting distributions.

Moving from size to power, the efficient score test has attractive optimality properties if the possible local nuisance parameter values are indexed by a linear space.[6] Firstly, if the covariance matrix of the efficient score function is non-singular then the efficient score test is asymptotically uniformly most powerful within the class of asymptotically invariant tests as defined and demonstrated by Choi et al. (1996).[7] Moreover, if the covariance matrix of the efficient score function has positive rank, I establish that the test enjoys a local asymptotic minimax optimality property. In addition to the standard full rank case, this situation may arise when the parameter of interest is underidentified.

I work out the details of the application of the general theory to two econometric models: a single index model where the link function may be relatively flat compared to sampling variation and a linear simultaneous equations model where identification may be weak when an identifying assumption of non-Gaussianity is close to failing. In each case, the models have nonstandard features which can invalidate some standard approaches to inference. For each model I give primitive conditions that allow (i) derivation of the efficient score function and (ii) a demonstration that the high level conditions required for the application of the previously developed theory are satisfied. Crucially, the assumptions imposed do not carve out parts of the parameter space which cause problems for other testing approaches.

Firstly, I consider a single index model (SIM). The SIM is a popular model in econometrics as it retains a large amount of flexibility whilst successfully combating the curse of dimensionality. Identification of parameters in the index function requires a number of assumptions, including the non-constancy of the link function. As is usual with points of identification failure, if the link function is sufficiently close to constancy relative to the sample size, a weak identification problem obtains. Importantly, the identification status of the parameter of interest in this model depends on the link function, an infinite dimensional nuisance parameter. Additionally regularised estimation is required to perform inference in this model. I demonstrate that the efficient score test provides (locally uniformly) valid size control in spite of these issues.

Secondly, I examine a semiparametric linear simultaneous equations model (LSEM). The LSEM is a foundational model in econometrics, used to analyse equilibrium relationships.

[6] This is often – but not always – the case. It fails, for example, at boundary points of the parameter space. See Rieder (2014) for a discussion and some optimality results in such cases.

[7] For scalar parameters the asymptotic invariance can be replaced by asymptotic unbiasedness for two-sided tests; for one-sided tests the asymptotic optimality holds over all tests of correct asymptotic level.

As is well known, the simultaneity problem precludes the identification of all structural parameters from observed data without further restrictions, leading researchers to adopt alternative methods (e.g. analysing only one equation with the help of instrumental variable techniques); see Dhrymes (1994) for an in-depth review.

In fact, the identification status of the structural parameters of interest depends on the true error distribution (an infinite dimensional nuisance parameter). In particular, if no more than one of the (mutually independent) error components is Gaussian the structural parameters are identified as a consequence of the Darmois-Skitovich Theorem (Comon, 1994).[8] If multiple components are Gaussian the structural parameters may be under- or un-identified and standard inferential approaches may fail to control size. As is typical in models with points of identification failure, such behaviour is also observed if the true error distributions are sufficiently to close to Gaussianity, relative to sampling variation. In addition to these potential identification problems, regularised estimation is required to handle the non-parametric part of the model, leading to regularisation bias. I demonstrate that despite the presence of these non-standard features, the efficient score test provides (locally uniformly) valid and efficient inference in the LSEM model, providing researchers with a direct approach to conduct inference on structural parameters in linear simultaneous systems without needing to employ, for example, instrumental variables approaches.

I conduct a large scale simulation study based on each example. The results verify that the asymptotic size results obtained provide a good guide to finite sample size, with the efficient score test always being correctly sized, including in cases where alternative procedures fail to correctly control size. The simulation studies also highlight the power of this testing approach and suggest that the asymptotic approximations provide a good guide to finite sample power, with finite sample power curves and surfaces matching the predictions of the asymptotic theory.

1.1.1. Relation to the literature

This paper is primarily a contribution to the literature on general approaches to robust inference methods for statistical and econometric models with non-standard asymptotic behaviour in part of the parameter space.

A number of papers analyse size-correction methods to provide inference valid uniformly over nuisance parameter values. For instance, Andrews and Guggenberger (2009, 2010a,b) analyse the use of resampling methods and data-dependent critical values to provide uniformly correct size control over the parameter space; McCloskey (2017) provides alternative size correction approaches based on Bonferroni bounds, which can improve the power of such size corrected tests. The approaches proposed in the cited papers are

[8] Strictly speaking the identification result is up to column permutations and sign changes of the matrix which transforms the structural shocks into reduced form shocks.

designed for models in which a statistic has a limiting distribution which is discontinuous in a finite-dimensional nuisance parameter.[9] This setup is very general but differs from the one considered in the present paper on a number of key points: (i) in this paper, the parameter which may cause standard inferential approaches to suffer from size distortions can be infinite dimensional; (ii) rather than size-correcting tests based on a specific test statistics which have parameter discontinuous asymptotic distributions, I suggest the use of the the efficient score statistic which always has a χ^2 distribution and hence the tests always use χ^2 critical values. There is not complete overlap between the class of models considered in this paper and those to which the methods in these papers are applicable: the efficient score test remains valid in cases where the asymptotic distribution of (other) test statistics may depend on the particular local sequence of infinite dimensional nuisance parameters. Conversely, the example of an autoregressive model with a root which may be local to unity studied in Andrews and Guggenberger (2009) does not satisfy the high-level conditions I impose as such models are locally asymptotically quadratic (LAQ) but not LAN (Jeganathan, 1995; Jansson, 2008).

Romano and Shaikh (2012) provide high level conditions under which bootstrap and subsampling procedures yield tests and confidence sets with (uniformly) correct size and coverage probabilities in a very general class of models. Their approach differs substantially from the approach in this paper, using resampling schemes to provide appropriate quantiles to conduct tests and construct confidence sets for the values of general parameters of interest defined on the model. As a result, their approach can deal with more general parameters of interest than are considered in this paper. On the other hand, there are cases in which the procedure outlined in this paper correctly controls size, but subsampling and bootstrapping approaches fail to do so, for example, subsampling TSLS t-type statistics in IV regression models with weak instruments (Andrews and Guggenberger, 2010a) and subsampling Wald-type statistics in models with nuisance parameters near the boundary (Andrews and Guggenberger, 2010b).

Elliott et al. (2015) provide nearly optimal tests for models which have a Gaussian shift limit experiment (locally to the true parameter) with part of the shift vector being a nuisance parameter. Their tests correctly control size and (approximately) maximise weighted average power given a weighting function (over the nonstandard region of the parameter space). Their approach requires the nuisance parameter to be finite dimensional and is quite different from the one proposed in this paper, though it shares some common threads, being based on a least favourable approach in a Gaussian shift limit experiment.[10]

For numerous classes of nonstandard inference problems a large literature exists analysing

[9] In related work, Andrews et al. (2020) provide some general results to establish the (uniform) size of tests and (uniform) coverage probabilities of confidence sets based on (pointwise) asymptotic distributions which are discontinuous in some function of a parameter.

[10] I do not consider least favourable distributions explicitly, however the efficient score function can be considered to correspond to an approximately least favourable submodel; see §25.11 in van der Vaart (1998).

the problem at hand and providing particular solutions. There are too many such examples to provide a full account here; instead I provide a selective summary of the literature pertaining to those non-standard features relevant to the examples I consider in detail in this paper, comprising (a) identification robust inference, (b) inference in models with boundary constraints and (c) inference post a model selection or regularisation step.

Inference robust to identification problems has been considered in various settings by, inter alia, Stock and Wright (2000); Kleibergen (2005); Andrews and Cheng (2012, 2013); Andrews and Mikusheva (2015, 2016a,b, 2022); Han and McCloskey (2019); Andrews and Guggenberger (2019).[11] Dufour (1997) provides some impossibility results. Chen et al. (2018) consider semiparametric models in which parameters may be only partially identified and suggest inferential procedures based on a Monte Carlo simulation approach. Kaji (2021) puts forward a general theory of weak identification in semiparametric models and focusses on efficient estimation rather than robust inference.

A long considered problem is inference in models with boundary constraints, which has been studied by, amongst others, Chernoff (1954); Geyer (1994); Andrews (2000, 2001); Andrews and Guggenberger (2010a,b); Chen et al. (2017); Ketz (2018); Cavaliere et al. (2020). An antecedent to the approach of this paper in the case of nuisance parameters potentially on (or close to) the boundary can be found in Andrews (2001, p. 698) where the nuisance parameters are split into those which satisfy a block diagonality condition with respect to the other parameters and those which do not. The author of that paper then notes that those which satisfy the block diagonality condition "may or may not lie on the boundary of the parameter space". I exploit a similar idea, as the efficient score function is orthogonal to *all* nuisance scores by construction.

Inference post model selection or regularisation is also problem with a long history, which has become increasingly important in recent years due to the increasing availability of "big data". Leeb and Pötscher (2005) analyse in detail some of the difficulties associated with inference post model selection; additional demonstrations along with applications of some of the size correction approaches previously mentioned can be found in Andrews and Guggenberger (2010a); McCloskey (2020). Chernozhukov et al. (2015) outline an approach to post model selection / post regularisation inference which uses an approach similar to the one proposed in this paper with their class of "Neyman orthogonalised" statistics also being a generalisation of the C(α) approach of Neyman (1959, 1979).[12] The development in their paper is framed somewhat differently and focusses on post-regularisation inference in problems defined by a finite vector of known moment conditions with a larger class of test statistics, whereas I consider a more general class of inference problems with potentially

[11] There is also a large literature on robust inference in models defined by moment inequalities (and partially identified models more generally). Additionally a further sub-literature exists on subvector inference for weakly identified parameters. I do not consider subvector inference in this sense in this paper, though I note here that Chaudhuri and Zivot (2011) used the efficient score corresponding to a GMM model as a way to improve power in projection-based subvector inference with weak identification.

[12] See also Belloni et al., 2017 and Chernozhukov et al., 2018.

non-standard features but only one test statistic.[13]

The general approach to inference outlined in this paper is based on the efficient score function which, along with its variance matrix (the "efficient information matrix"), is a key quantity in the literature on semiparametric efficiency. Textbook treatments of this framework can be found in Bickel et al. (1998); van der Vaart (2002) and van der Vaart (1998, Chapter 25). The efficient score test was shown to be optimal (in certain classes of tests) by Choi et al. (1996). These ideas have been widely used in statistics and econometrics since their introduction, particularly to determine efficiency bounds in semiparametric models and construct estimators which attain them.

I now briefly turn to the specific examples I consider. The first – inference in the single index model with potential identification failure – is related to the (previously summarised) literatures on inference with potential identification problems and inference post-regularisation as well as the literature on single index models and extensions thereof. Such models have been widely studied by, amongst others, Ichimura (1993); Newey and Stoker (1993); Ma and Zhu (2013).

The second example I consider, the LSEM, is related to the (previously summarised) literatures on inference with potential identification problems and inference post-regularisation as well as the statistical literature on independent components analysis (ICA) modelling. The ICA model has long been used in a number of fields as an approach to the analysis of data forming systems of simultaneous equations; see Hyvärinen et al. (2001) for many examples.[14] By adding covariates to the ICA model a class of linear simultaneous equations models is obtained. Such systems of equations have a long history in econometrics; see the introduction of Lee and Mesters (2022a) for a summary.[15] A semiparametric approach to the ICA model was considered in Amari and Cardoso (1997); Chen and Bickel (2006). Lee and Mesters (2022a) consider a semiparametric approach to the LSEM which uses the approach discussed in this paper to conduct tests robust to potential identification failure. Concretely, they consider testing when the (fixed) distribution of the error terms may be arbitrarily close to Gaussianity but this distribution is not permitted to change with the sample size. They provide simulation evidence of a weak identification problem when the error distribution is sufficiently close to Gaussianity (relative to the sample size), but their theoretical work assumes a fixed error distribution and consequently does not cover weak identification. In contrast, in this paper, I explicitly model weak identification and obtain size results which are valid locally uniformly over

[13] In many models, the test statistic considered in this paper would belong to the general class they consider.

[14] The ICA model relates observables Y and errors ϵ according to $Y = A^{-1}\epsilon, \quad \mathbb{E}\epsilon = 0, \ \mathbb{V}\epsilon = I$ where ϵ has independent components.

[15] More recently such models have also been adopted in econometrics as an approach to SVAR modelling, with an assumption of non-Gaussianity imposed to identify the matrix required to obtain the structural shocks from the reduced form shocks. A recent summary of this approach is given by Montiel Olea et al. (2022). Also see, inter alia, Gouriéroux et al. (2017, 2019); Lanne and Lütkepohl (2010); Lanne et al. (2017); Lanne and Luoto (2021); Bekaert et al. (2019, 2020); Fiorentini and Sentana (2022, 2021); Davis and Ng (2021). Velasco (2020) considers the more general SVARMA case. In this paper I do not consider dynamics for simplicity.

subsets of the parameter space.

1.1.2. Outline

The remainder of this paper is organised as follows. Section 1.2 describes the setting of the paper, explains the intuition underlying the testing approach and introduces a number of examples. Section 1.3 formalises the heuristic definitions given previously, develops the theoretical contributions of this paper under high level conditions and provides some lower-level conditions and constructions sufficient for their validity. Two examples are worked out in detail in sections 1.4 and 1.5; these sections also discuss the results from several simulation studies. Section 1.6 concludes and discusses possible extensions.

1.2. Heuristic explanation and examples

I now provide a heuristic discussion of the efficient score test, focussing on the underlying intuition, and provide a number of examples to demonstrate the breadth of applicability of my framework. I purposely omit all formal definitions and assumptions, which are provided in section 1.3 below.

The parameter of interest is $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and the goal is to construct (asymptotically) correctly sized tests for the hypothesis $H_0 : \theta = \theta_0$ or confidence sets for θ which have correct (asymptotic) coverage probability over a range of data generating processes (DGPs) consistent with the null hypothesis.

I suppose that the researcher observes a random sample $(W_i)_{i=1}^n$. The considered probability model for the distribution of each such observation W_i is given by

$$\mathcal{P} = \{P_\gamma : \gamma \in \Gamma\}, \quad \Gamma = \Theta \times \mathcal{H}, \tag{1.1}$$

where $\gamma = (\theta, \eta)$ with η collecting all the remaining parameters required to fully describe the distribution of the data (given θ). In the classical parametric setting η is finite dimensional; in the semiparametric models which are the focus of this paper it may be infinite dimensional.

Analogously to the parametric case, it is possible to define *score functions* for all of the parameters in semiparametric models (see section 1.3 for the details). Let $\dot{\ell}_\gamma$ be the (vector of) score functions for θ and $\mathscr{H}_\gamma = \{B_\gamma h : h \in H\}$ a collection of score functions for η.[16] All score functions are mean zero and have finite variance. The *efficient score function*

[16] The score functions are indexed by elements h in a set H. In the parametric case this set could be taken as the integers from 1 to the (finite) number of elements in η. In the case where η is infinite dimensional, the indexing set H will typically also be infinite dimensional.

is defined as the orthogonal projection (in L_2) of the scores for θ onto the orthogonal complement of the scores for η:

$$\tilde{\ell}_\gamma = \dot{\ell}_\gamma - \Pi\left(\dot{\ell}_\gamma \middle| \overline{\text{lin}}\, \mathscr{H}_\gamma\right), \tag{1.2}$$

where $\overline{\text{lin}}\, \mathscr{H}_\gamma$ denotes the closed linear span of the set $\mathscr{H}_\gamma$.[17] This operation removes from $\dot{\ell}_\gamma$ that part which can explained by score functions in $\mathscr{H}_\gamma$. The corresponding variance matrix, the *efficient information matrix* is

$$\tilde{\mathcal{I}}_\gamma = \int \tilde{\ell}_\gamma \tilde{\ell}_\gamma' \, \mathrm{d}P_\gamma.$$

Analytical derivation of the efficient score function for specific models can be complex, however due to the central role of the efficient score function in the literature on semi parametrically efficient estimation the efficient score function has already been derived for a large number of popular models.[18]

As a direct consequence of the definition in (1.2), $\int \tilde{\ell}_\gamma \, \mathrm{d}P_\gamma = 0$ and hence the efficient score function provides a d_θ-dimensional vector of moment condition on which one can base inference about θ. In general, constructing estimators and tests based on the efficient score function is attractive as these have well established optimality properties (e.g. Bickel et al., 1998; van der Vaart, 2002; Choi et al., 1996). In some of the examples considered in this paper, the conditions which are required to obtain such results may fail. For instance, if θ is unidentified, no consistent estimator of θ can exist, let alone asymptotically efficient estimators. Nevertheless, I will show that in such situations tests based on the efficient score function can be used to conduct valid inference provided some mild conditions are satisfied.

To introduce the test statistic, let $\hat{\ell}_{n,\theta}$ and $\hat{\mathcal{I}}_{n,\theta}$ denote estimates of $\tilde{\ell}_\gamma$ and $\tilde{\mathcal{I}}_\gamma$ respectively. The efficient score statistic (for a given θ) is given by

$$\hat{S}_{n,\theta} = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\ell}_{n,\theta}(W_i)\right)' \hat{\mathcal{I}}_{n,\theta}^{\dagger} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\ell}_{n,\theta}(W_i)\right),$$

where "†" denotes the Moore-Penrose pseudo-inverse. Supposing that mild assumptions hold, I show that, under $H_0 : \theta = \theta_0$, $\hat{S}_{n,\theta_0}$ converges in distribution to a χ^2_r random variable where $r = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$. Importantly (i) this convergence holds under any local sequence of nuisance parameters and (ii) the assumptions imposed do not require θ to be identified, allow η to be on the boundary of the parameter space and allow for the estimates to depend on regularised estimators of η. Based on this convergence, the efficient score test is performed by comparison of $\hat{S}_{n,\theta_0}$ to the appropriate quantile of the $\chi^2_{r_n}$ distribution where $r_n = \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_0})$ and confidence sets for θ can be constructed by inverting the test.

[17] The projection in the preceding display should be understood componentwise.

[18] Additionally guidance and a large number of examples can be found in Newey (1990), Bickel et al. (1998) and van der Vaart (1998, Chapter 25).

Intuitively there are two features of the efficient score statistic which are responsible for this result. The first is that the null value θ_0 is imposed in the construction of the statistic which precludes the need for θ to be identifiable or consistently estimable. This is key in models with potential identification failures, where such requirements can fail. Second, the orthogonal projection in the definition of the efficient score function ensures that

$$\int \tilde{\ell}_\gamma \, B_\gamma h \, \mathrm{d}P_\gamma = 0 \;\text{ for all }\; B_\gamma h \in \mathscr{H}_\gamma, \tag{1.3}$$

i.e. the efficient score function is uncorrelated with the scores $B_\gamma h$ for the nuisance parameters (in each direction h). Similar properties have been shown to alleviate size distortions in a number of settings, including those caused by identification issues (Kleibergen, 2005), boundary effects (Andrews, 2001) and regularised estimation of nuisance parameters (Chernozhukov et al., 2015, 2018). Property (1.3) has a fundamentally important role more generally in models with nuisance parameters in order to obtain the same limiting distribution regardless of the local sequence of nuisance parameters under which the limit is taken (cf. Hall and Mathiason, 1990; Choi et al., 1996).[19]

In addition to the robustness properties that (1.3) gives the efficient score test, (1.3) is also important for its power optimality properties – reflecting the original development of the C(α) test by Neyman (1959). If the efficient information matrix has full rank – as is usually the case in well identified models – and local perturbations to the nuisance parameters are indexed by a linear space, the efficient score test belongs to the class of asymptotically uniformly most powerful invariant tests (AUMPI) as described and demonstrated in Choi et al. (1996). Moreover, if the efficient information matrix has positive rank, there are directions against which non-trivial local power can be attained. I demonstrate that the efficient score test is minimax optimal in this scenario, in that there is no alternative test which provides higher power in a minimax sense.

To illustrate the broad applicability of these results, I now present two different examples to show (i) how commonly used econometric models can be placed into the framework required by (1.1) and (ii) how certain (local) sequences of nuisance parameters η can cause problems for commonly used inferential procedures. Following this I briefly discuss a number of other important examples in econometrics for which the inferential approach in this paper could be useful.

Example 1 (Single-index model)**.** *Consider the single-index regression model (e.g Ichimura, 1993; Horowitz, 2009)*

$$Y = f(X_1 + X_2\theta) + \epsilon, \quad \mathbb{E}(\epsilon|X) = 0,$$

where $f : \mathbb{R} \to \mathbb{R}$ *belongs to some function class* $\mathscr{F}$*,* X_1 *and* X_2 *are continuously*

[19] See also the discussions comparing Rao's score test and Neyman's C(α) test on page 133 of Andrews and Mikusheva (2015) and page 492 of Kocherlakota and Kocherlakota (1991).

distributed random variables and ϵ is an unobserved error term. $(\epsilon, X) \sim \zeta$ for some Lebesgue density function ζ which ensures that the conditional mean restriction indicated above is satisfied. Such single-index models are popular as they relax the commonly imposed linear structure of linear regression models but avoid the curse of dimensionality by ensuring the argument of f is a scalar. The density of an observation $W = (Y, X) \in \mathbb{R}^3$ is

$$p_\gamma(W) = \zeta(Y - f(X_1 + X_2\theta), X),$$

and the corresponding model is given by $\mathcal{P} = \{P_\gamma : \gamma \in \Theta \times \mathcal{H}\}$ for some open $\Theta \subset \mathbb{R}$ and $\mathcal{H} = (f, \zeta) \in \mathscr{F} \times \mathscr{Z}$, where the latter set restricts the possible distribution of (ϵ, X).

As discussed in Horowitz (2009), θ is unidentified when f is a constant function. Weak identification can therefore occur when f is sufficiently close to constancy (relative to the sample size). The potential identification failure here is due to an infinite dimensional *nuisance parameter and therefore robust approaches to inference designed for cases where identification failure is caused by a finite dimensional nuisance parameter do not apply. Derivations of the efficient score function for the model above (and various extensions) can been found in the literature, see e.g. Newey and Stoker (1993); Ma and Zhu (2013); Kuchibhotla and Patra (2020). The efficient score test permits inference on θ to be performed which is robust to potential identification failure; full details are given in section 1.4.* △

Example 2 (Simple linear simultaneous equations model)**.** *Suppose that the $K \times 1$ vector W satisfies*

$$W = A(\theta)^{-1}\epsilon,$$

where $A(\theta)$ is a rotation matrix parametrised by $\theta \in \Theta$ and ϵ a $K \times 1$ vector of independent structural shocks each with mean zero and unit variance. Let $\eta = (\eta_1, \ldots, \eta_K) \in \mathcal{H}$ denote the densities of the components of ϵ. This yields the model

$$\mathcal{P} = \{P_\gamma : \gamma = (\theta, \eta) \in \Gamma = \Theta \times \mathcal{H}\},$$

where P_γ has Lebesgue density $p_\gamma(W) = \prod_{k=1}^{K} \eta_k \left(A_k(\theta)W\right)$.[20]

If all ϵ_k are Gaussian, $A(\theta)$ is not identified and hence the same is true of θ. In contrast, if (at least) $K - 1$ of the components of ϵ have non-Gaussian distributions, $A(\theta)$ is identified up to sign changes and column permutations (Comon, 1994). Appropriate restrictions on the signs and labelling of the elements then result in identification of θ. However, if the non-Gaussian distributions of the ϵ_k are sufficiently close to Gaussian, θ is only weakly identified and inference methods which assume non-Gaussianity can suffer from size distortions.

The efficient score test avoids these size distortions by fixing $\theta = \theta_0$ under the null and orthogonalising with respect to (the scores for) η. In section 1.5, I show that the conclusions

[20] $A_k(\theta)$ is the k-th row of $A(\theta)$.

of these heuristic arguments hold formally in a considerably richer class of LSEMs. I also show that inference based on the efficient score test is minimax optimal in these models, including in cases where θ is underidentified.

The identification problem in this example is caused by an infinite dimensional *nuisance parameter and therefore robust approaches to inference designed for cases where identification failure is caused by a finite dimensional nuisance parameter do not apply.* △

Other examples

In addition to the preceding examples, robust inference on a large variety of other models of interest in econometrics can be conducted using the approach in this paper, pending verification of the high-level conditions in the next section. I briefly discuss four such cases here.

Firstly, consider inference on the slope parameters θ associated with the endogenous variables in an instrumental variables regression model. As is well known, many standard tests are unreliable in instrumental variable regression models if the instruments are weak (Andrews et al., 2019). In contrast, the efficient score test could be used to provide valid inference in this model. In this model – unlike examples 1 or 2 – the lack of identification is caused by a finite dimensional parameter. Nevertheless, due to potential heteroskedasticity, the efficient score in this model depends on an infinite dimensional object, the heteroskedastic function. The resulting test does not coincide with any of the "standard" weak-IV robust tests, such as the AR, LM and CLR statistics (e.g. Anderson and Rubin, 1949; Staiger and Stock, 1997; Moreira, 2003; Kleibergen, 2002, 2007).

Secondly, consider the classical linear errors-in-variables model (as in, for example, equation (1.1) of Bickel and Ritov, 1987 or equation (1) of Ben-Moshe, 2020). As discussed by numerous authors (e.g. Reiersøl, 1950; Willassen, 1979; Bickel and Ritov, 1987; Ben-Moshe, 2020), identification of the regression coefficients may depend on (joint) distributional properties of the covariates, structural errors and measurement errors. These can include, for example, independence restrictions and non-Gaussianity assumptions on the latent covariates (Reiersøl, 1950; Willassen, 1979). Similarly to example 2, on verification of the high-level conditions in the next section, the inferential framework in this paper could be used to perform inference which will remain valid if, for instance, the distribution of the latent covariates is sufficiently close to Gaussianity that the regression coefficients become weakly identified. As in examples 1 and 2, this is a case of non-regularity caused by an infinite dimensional parameter.

As a third example, consider the mixed proportional hazard model, a common model used in duration analysis which allows for unobserved heterogeneity (see van den Berg, 2001, for a review). As was demonstrated by Hahn (1994), in the case where the baseline hazard function is Weibull, the efficient information matrix (for the Euclidean parameters) is singular, and no regular estimator sequence for these parameters can exist.[21] Pending verification of the high-level conditions in the next section, the inferential framework outlined in this paper could be used to perform inference which will remain valid if the baseline hazard function is (close to) Weibull. As in examples 1 and 2, this is a case of non-regularity caused by an infinite dimensional parameter.

Finally, as is well known, models with nuisance parameters on or close to the boundary can cause standard testing approaches to be unreliable (Andrews, 2001; Elliott et al., 2015; Ketz, 2018). Similar problems may arise in models where nuisance functions are estimated with shape restrictions imposed (cf. Chetverikov et al., 2018, section 3). Due to the orthogonality between the scores for the parameter of interest and the nuisance scores, these restrictions do not affect the limiting distribution of the efficient score statistic and hence inferential approach in this paper will remain valid in these models – pending the verification of the high-level conditions in the next section. Depending on the model and the restriction under consideration, this case of non-regularity may be caused by either a finite-dimensional parameter or an infinite-dimensional parameter.

The next section describes the high level theory and provides a set of mild assumptions under which the efficient score test provides robust inference and has power optimality properties. Thereafter I revisit and generalise examples 1, 2 and work out the details for implementation.

1.3. Theory

In this section I formalise inference based on the efficient score statistic. First I set out the high-level assumptions which will be required throughout and formally define the efficient score test and associated confidence sets. Second, I perform an asymptotic analysis of the size properties of this test and the coverage of the associated confidence sets. Third, I demonstrate that this test has power optimality properties in a number of scenarios. Finally I provide a number of conditions and constructions which are sufficient for the high-level assumptions and often simpler to verify. In what follows I will often use operator notation for integrals, e.g. for a function f and a probability measure P, $Pf := \int f \, \mathrm{d}P$. $\mathbb{P}_n$ denotes the empirical measure of the sample $(W_i)_{i=1}^n$, so $\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(W_i)$.[22]

1.3.1. Model setup and maintained assumptions

The first assumption that I impose merely formalises the model of interest as discussed in section 1.2 and stipulates that the observed data form a random sample.

[21] Hahn (1994) also derives the efficient score function for this model.

[22] See appendix section A for additional details and notational conventions.

Assumption M (Model and sampling). *Let $(W_i)_{i=1}^n$ be independent copies of a $\mathcal{W}$-valued random element W, with $\mathcal{W}$ a Polish space, all defined on an underlying probability space $(\Omega, \mathcal{F}, \mathrm{P})$.*[23] *The considered model for the law of W on $(\mathcal{W}, \mathcal{B}(\mathcal{W}))$ is*

$$\mathcal{P} \coloneqq \{P_\gamma : \gamma \in \Gamma\},$$

where Γ has the product form $\Gamma = \Theta \times \mathcal{H}$ for Θ an open subset of $\mathbb{R}^{d_\theta}$ and $\mathcal{H}$ a metric space. A typical value $\gamma \in \Gamma$ will be written as $\gamma = (\theta, \eta)$ where $\theta \in \Theta$ and $\eta \in \mathcal{H}$. Each $P_\gamma \in \mathcal{P}$ is dominated by a common σ-finite measure ν. ◇

The next assumption is the key requirement. It imposes that the model satisfies a LAN condition (e.g. van der Vaart, 1998, Chapter 7; Le Cam and Yang, 2000, Chapter 6), where the parameter $\gamma = \gamma_n$ can change with the sample size n. In order to state this assumption, some notation is required. For any $P_\gamma \in \mathcal{P}$ I write p_γ for its density with respect to ν and for any two points $\gamma_1, \gamma_2 \in \Gamma$, $\Lambda_n(\gamma_1, \gamma_2)$ denotes the log-likelihood ratio:

$$\Lambda_n(\gamma_1, \gamma_2) \coloneqq \log \prod_{i=1}^n \frac{p_{\gamma_1}}{p_{\gamma_2}}. \tag{1.4}$$

The LAN requirement is imposed as follows.

Assumption LAN (Local asymptotic normality). *Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in Γ which converges to a point $\gamma \in \Gamma$ and H_η a subset of a Banach space, H, which includes 0.*

For any sequence $\tau_n \to \tau$ with each $\tau_n, \tau \in \mathbb{R}^{d_\theta}$, any sequence $h_n \to h$ with $h_n, h \in H_\eta$, a convergent sequence of $d_\theta \times d_\theta$ matrices δ_n and sequences $\eta_n(h_n) \to \eta$ with each $\eta_n(h_n) \in \mathcal{H}$, define

$$\gamma_n(\tau_n, h_n) \coloneqq (\theta_n + \delta_n \tau_n, \eta_n(h_n)),$$

and suppose that

1. *the sequence $(P_{\gamma_n(\tau_n,h_n)})_{n\geq 1}$ is (eventually) in $\mathcal{P}$,*
2. *the associated log-likelihood ratio satisfies*

$$\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right] - \frac{1}{2} P_{\gamma_n} \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right]^2 + o_{P_{\gamma_n}}(1), \tag{1.5}$$

for a sequence of functions $(\dot{\ell}_{\gamma_n})_{n\in\mathbb{N}}$ with each $\dot{\ell}_{\gamma_n} \in L_2^0(P_{\gamma_n})$ and a sequence of linear maps $(B_{\gamma_n})_{n\in\mathbb{N}}$ with each $B_{\gamma_n} : H_\eta \to L_2^0(P_{\gamma_n})$ such that $\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ is uniformly square P_{γ_n}-integrable. ◇

In what follows I use the notation P_{γ_n,τ_n,h_n} for $P_{\gamma_n(\tau_n,h_n)}$. The functions $\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ will

[23] A Polish space is a separable completely metrisable topological space. Let d be a metric such that $(\mathcal{W}, d)$ is a complete (separable) metric space. $\mathcal{B}(\mathcal{W})$ is the Borel σ-algebra on $(\mathcal{W}, d)$.

(collectively) be called "score functions", as will the vector $\dot{\ell}_{\gamma_n}$ (the "score functions for θ") and the functions $B_{\gamma_n}h$ (the "score functions for η"). Such functions play the same role as score functions in classical parametric models in which – under regularity conditions – a similar LAN condition holds (e.g. van der Vaart, 1998, Theorem 7.2).

Assumption LAN stipulates that the likelihood ratios $\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n)$ admit a local quadratic approximation with a particular form. It is important to clarify the roles of the different sequences of parameters present in these likelihood ratios. I refer to $(\gamma_n)_{n\in\mathbb{N}}$ as the "base sequence" and the components $\delta_n\tau_n$ and $\eta_n(h_n) - \eta_n$ as "local perturbations" to the elements of this base sequence respectively:

$$\begin{aligned}\gamma_n(\tau_n, h_n) &= \gamma_n + (\delta_n\tau_n,\ \eta_n(h_n) - \eta_n)\\ &= \big(\,\theta_n + \underbrace{\delta_n\tau_n}_{\text{local perturbation of }\theta_n},\quad \eta_n + \underbrace{\eta_n(h_n) - \eta_n}_{\text{local perturbation of }\eta_n}\,\big).\end{aligned}$$

That γ_n is permitted to vary with n has two important implications. Firstly, replacing a fixed θ with a convergent sequence $\theta_n \to \theta$ permits the demonstration that confidence sets constructed by inverting the efficient score test are *uniformly* valid over compact subsets of Θ. Secondly, this permits local power analysis in situations where the rate of information accumulation is non-standard.[24]

The separation of the local perturbation of θ_n into a "rate" term δ_n and a "direction" term τ_n is not strictly necessary but clarifies the role each plays in the subsequent power results. Due to the (possible) infinite dimensionality of the nuisance parameters η_n, the form of the local perturbation may be complex and generally will be model dependent, but the role of h_n is analogous to that of τ_n, i.e. it is the "direction" term in the perturbation.

Assumption LAN requires that for any permitted sequence of local perturbations, the measures P_{γ_n,τ_n,h_n} eventually belong to the model and (1.5) holds. That these hold over all such local sequences is key for the size results below which demonstrate that the efficient score test controls size *locally uniformly*, i.e. over any compact set of local perturbation directions consistent with the null. I emphasise that in the size and power results below LAN is only assumed to hold along certain specified base sequences $(\gamma_n)_{n\in\mathbb{N}}$ which are defined in the relevant results.

It is also important to note that assumption LAN concerns only the model $\mathcal{P}$ and perturbation spaces H_η, both of which are chosen by the researcher. This includes the choice

[24] For instance, one key feature of weak or semi-strong identification (in the terminology of Andrews and Cheng, 2012) is that the information that can be learned about the parameter of interest accrues at a rate slower than the "usual" $\sqrt{n}$; robust tests can then often be built on top of "rescaling" arguments: some part of γ_n changes with the sample size, causing a slower rate of information acquisition, which can be compensated for by a "slower" rate sequence δ_n — i.e. the local alternatives are "closer" than in the "usual" $\sqrt{n}$ case (Cf. Antoine and Renault, 2009, 2011; Andrews and Mikusheva, 2015). The prototypical "weak identification" case is usually the limiting case of this argument, where $\delta_n \not\to 0$ and the "local" alternatives are, in a sense, "fixed" alternatives.

of the metric on $\mathcal{H}$, which – particularly in the infinite dimensional case – has implications for the uniformity results obtained below, which hold over compact sets. Specifically, choosing a stronger metric on $\mathcal{H}$ will often simplify the demonstration that assumption LAN holds, but leads to "fewer" compact sets and therefore weaker uniformity results.[25]

Finally, rather than establishing LAN directly, one may establish that the relevant submodels are differentiable in quadratic mean (see assumption DQM below), which then implies assumption LAN (under assumption M; see proposition 1.3.10). A detailed analysis of the relationship between conditions of these types is given by Le Cam (1986, Chapter 17, section 3); see also Strasser (1985, Theorem 75.9).

I now introduce the next assumption, which concerns the limits of the scores.

Assumption CM(i) (Convergence of moments (i))**.** *In the setting of assumption LAN suppose that there exists a vector of functions* $\dot{\ell}_\gamma \in L_2^0(P_\gamma)$ *and a bounded linear map* $B_\gamma : H_\eta \to L_2^0(P_\gamma)$ *such that for each* $(\tau, h) \in \mathbb{R}^{d_\theta} \times H_\eta$

$$\lim_{n\to\infty} P_{\gamma_n}\left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h\right]^2 = P_\gamma\left[\tau'\dot{\ell}_\gamma + B_\gamma h\right]^2.$$

◇

The uniform integrability required by assumption LAN may directly imply that assumption CM(i) holds; see subsection 1.3.4 for some sufficient conditions.

With the quantities introduced in the preceding assumptions, the efficient score function can be formally defined. First define the *tangent sets* for η as

$$\mathscr{H}_\gamma := \{B_\gamma h : h \in H_\eta\}, \quad \text{for} \quad \gamma \in \{\gamma\} \cup \{\gamma_n : n \in \mathbb{N}\}.$$

The efficient score functions are defined as the orthogonal projections of the score functions for θ, i.e. the $\dot{\ell}_{\gamma_n}$ and $\dot{\ell}_\gamma$ onto the orthocomplement of $\mathscr{H}_{\gamma_n}$ and $\mathscr{H}_\gamma$ respectively. The corresponding efficient information matrices are the expectations of the outer products of these (vectors of) functions:

$$\tilde{\ell}_\gamma := \dot{\ell}_\gamma - \Pi_\gamma\left(\dot{\ell}_\gamma \mid \overline{\text{lin}}\,\mathscr{H}_\gamma\right), \quad \tilde{\mathcal{I}}_\gamma := P_\gamma\left[\tilde{\ell}_\gamma\tilde{\ell}_\gamma'\right], \quad \text{for} \quad \gamma \in \{\gamma\} \cup \{\gamma_n : n \in \mathbb{N}\},$$

where $\Pi_\gamma(\cdot|\mathcal{S})$ is the orthogonal projection on $\mathcal{S} \subset L_2(P_\gamma)$.

I assume the same uniform integrability moment convergence conditions on the efficient scores that have been imposed on the scores for θ and η.

Assumption CM(ii) (Convergence of moments (ii))**.** *Suppose that assumption CM(i) holds and moreover that* $\|\tilde{\ell}_{\gamma_n}\|_2^2$ *is uniformly* P_{γ_n}*-integrable and* $\lim_{n\to\infty}\tilde{\mathcal{I}}_{\gamma_n} = \tilde{\mathcal{I}}_\gamma$. ◇

[25] More formally, if d_1 and d_2 are metrics on $\mathcal{H}$ with d_1 stronger than d_2 (i.e. every open subset of $\mathcal{H}$ with respect to d_2 is also open with respect to d_1), then if a set $H' \subset \mathcal{H}$ is compact with respect to d_1, then it is compact with respect to d_2.

The definition of the efficient score function ensures that $P_\gamma \tilde{\ell}_\gamma = 0$, since both $\dot{\ell}_\gamma$ and the elements of $\overline{\text{lin}}\ \mathscr{H}_\gamma$ are mean zero by assumption LAN. In other words, the efficient score function provides d_θ moment conditions on which inference about θ can be based.

In many cases, the efficient score function will not be formed only of observed or known quantities, but will need to be estimated. The following two conditions impose what is required of these estimates and complete the collection of high-level assumptions.

Assumption E (Estimation). *Let $(\gamma_n)_{n\in\mathbb{N}}$ be as in assumption LAN and suppose that for an estimator $\hat{\ell}_{n,\theta_n}$*

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] = o_{P_{\gamma_n}}(1), \tag{1.6}$$

and for an estimator $\hat{\mathcal{I}}_{n,\theta_n}$

$$\left\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_\gamma\right\|_2 = o_{P_{\gamma_n}}(1). \tag{1.7}$$

◇

Assumption R (Rank convergence). *Let $(\gamma_n)_{n\in\mathbb{N}}$ be as in assumption LAN and suppose that the estimator $\hat{\mathcal{I}}_{n,\theta_n}$ of assumption E satisfies*

$$P_{\gamma_n}\left(\text{rank}(\hat{\mathcal{I}}_{n,\theta_n}) = \text{rank}(\tilde{\mathcal{I}}_\gamma)\right) \to 1. \tag{1.8}$$

◇

That the first condition of assumption E, equation (1.6), can hold is often related to the specific structure of the efficient score function, particularly the fact that it is orthogonalised with respect to the nuisance scores. The second condition (1.7) requires consistency of an estimator of the efficient information matrix $\tilde{\mathcal{I}}_\gamma$. If the latter is non-singular and (1.7) holds, then (1.8) holds automatically.[26] If $\tilde{\mathcal{I}}_\gamma$ is rank deficient, (1.8) must be established separately. A construction which can ensure this holds, given an initial estimator with known convergence rate is given in subsection 1.3.4.

The fact that assumption R is required is due to the fact that the Moore-Penrose pseudo-inverse (which I denote by $M^\dagger$ for an arbitrary matrix M) is not continuous. However, if $E_n \to 0$ such that $M + E_n$ has the same rank as M, then $(M + E_n)^\dagger \to M^\dagger$.[27]

Verification of equations (1.6) and (1.7) is model specific and typically requires the application of various stochastic limit theorems. Incorporating estimates of Euclidean parts of the nuisance parameter can typically be achieved relatively simply via discretisation arguments if a $\sqrt{n}$-consistent estimator is available; see the example in section 1.5 below. For nonparametric parts, sample splitting can often be used to provide estimators for which the verification of the required conditions is relatively straightforward.

[26] See Lemma C.7.

[27] See e.g. Ben-Israel and Greville (2003, Section 6.6) and Cf. Andrews (1987).

1.3.2. The efficient score test

In this section, I define the efficient score test, which forms the basis of the inferential approach suggested in this paper. Two different definitions are required: one for a (scalar) one-sided hypothesis and one for a two-sided hypothesis.

For the purposes of testing a two-sided hypothesis at level $\alpha \in (0, 1)$, the efficient score statistic at θ is defined as

$$\hat{S}_{n,\theta} := \left(\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta}\right)' \hat{\mathcal{I}}_{n,\theta}^{\dagger} \left(\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta}\right). \tag{1.9}$$

The efficient score test can then be defined as

$$\phi_{n,\theta} := \mathbf{1}\left\{\hat{S}_{n,\theta} > c_n\right\}, \tag{1.10}$$

where c_n is the $1-\alpha$ quantile of the $\chi^2_{r_n}$ distribution, with $r_n := \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta})$. The confidence set corresponding to the efficient score test is denoted by $\hat{C}_n$ and defined as

$$\hat{C}_n := \{\theta \in \Theta : \phi_{n,\theta} = 0\} = \left\{\theta \in \Theta : \hat{S}_{n,\theta} \leq c_n\right\}. \tag{1.11}$$

For the purposes of testing a one-sided hypothesis for a scalar parameter, i.e. when $d_\theta = 1$ and $\alpha \in (0, 1/2]$, I instead define the efficient score statistic at θ as

$$\hat{S}_{n,\theta} := \left(\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta}\right) \sqrt{\hat{\mathcal{I}}_{n,\theta}^{\dagger}}, \tag{1.12}$$

and define the corresponding test as

$$\phi_{n,\theta} := \mathbf{1}\left\{\hat{S}_{n,\theta} > z_\alpha\right\}, \tag{1.13}$$

where z_α is the $1 - \alpha$ quantile of the $\mathcal{N}(0, 1)$ distribution. A confidence set can again be constructed by test inversion as

$$\hat{C}_n := \{\theta \in \Theta : \phi_{n,\theta} = 0\} = \left\{\theta \in \Theta : \hat{S}_{n,\theta} \leq z_\alpha\right\}. \tag{1.14}$$

The use of the same notation for these different objects should not cause any confusion as only one of the two is applicable to any given testing problem and hence which is meant will be clear from context.

1.3.3. Asymptotic properties

I now derive the asymptotic properties of the efficient score test and test inversion confidence sets. I first state a weak convergence result along local alternatives, which follows directly from standard stochastic limit theorems and Le Cam's third lemma. Following this size results are given in section 1.3.3 and power results in section 1.3.3.[28]

Proposition 1.3.1. *Suppose that assumptions M, LAN and CM(i) hold. Then, the sequences of product measures* $\left(P_{\gamma_n}^n\right)_{n\in\mathbb{N}}$ *and* $\left(P_{\gamma_n,\tau_n,h_n}^n\right)_{n\in\mathbb{N}}$ *are mutually contiguous. If also assumption CM(ii) holds, then under* P_{γ_n,τ_n,h_n}

$$\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma).$$

If, additionally, (1.6) *of assumption E holds, then also under* P_{γ_n,τ_n,h_n}

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_n} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma).$$

The key takeaway from the preceding proposition is that the limiting distributions depend on τ but not on h (or $(h_n)_{n\in\mathbb{N}}$): by its construction the efficient score function has an invariance property with regard to the local nuisance perturbations.

Size results

The invariance property discussed in the preceding paragraph is precisely what ensures that the size of the efficient score test does not depend on the particular local nuisance perturbation along which the limit is taken.[29]

Proposition 1.3.2. *Suppose that assumptions M, LAN, CM(ii), E and R hold for a sequence* $(\gamma_n)_{n\in\mathbb{N}} \subset \Gamma$ *with limit* $\gamma \in \Gamma$ *and where* $\theta_n = \theta_0$ *for all* $n \in \mathbb{N}$*. Then, for any compact subset* H_η' *of* H_η*,*

$$\lim_{n\to\infty} \sup_{h\in H_\eta'} P_{\gamma_n,0,h}^n \phi_{n,\theta_0} \leq \alpha.$$

The preceding proposition demonstrates that the efficient score test is correctly sized uniformly over local perturbations consistent with the null. Note that this result specifies that the high-level conditions need hold only along the specified base sequence with $\gamma_n = (\theta_0, \eta_n) \to (\theta_0, \eta) = \gamma$. This result immediately implies that the efficient score test is correctly sized along any sequence of local perturbations of $\gamma_n = (\theta_0, \eta_n)$ with $\tau_n = 0$ and $h_n \to h$ in H_η.[30]

An analogous result holds for confidence sets constructed by test inversion, provided the

[28] Readers primarily interested in the robustness results may safely skip section 1.3.3.

[29] In fact this property can be shown to hold rather more generally, for $\breve{\ell}_{\gamma_n}$ in place of $\tilde{\ell}_{\gamma_n}$ as long as $P_{\gamma_n}[\breve{\ell}_{\gamma_n} B_{\gamma_n} h] = 0$ for all $h \in H_\eta$. If $\breve{\ell}_{\gamma_n} \neq \tilde{\ell}_{\gamma_n}$ this would typically result in a less powerful test and hence I do not explicitly consider this case in the theoretical results. Nevertheless this observation can be particularly useful in cases when the efficient score function is hard to estimate. See e.g. the treatment of heteroskedasticity in section 1.4 below.

[30] In a metric space the union of a convergent sequence and its limit is compact.

high level conditions hold along sequences of the form $\gamma_n = (\theta_n, \eta_n) \to (\theta, \eta) = \gamma$, for any convergent sequence $\theta_n \to \theta$ (in a compact subset of Θ) and a specified $\eta_n \to \eta$.

Proposition 1.3.3. *Let Θ' be a compact subset of Θ. Fix a convergent sequence $(\eta_n)_{n\in\mathbb{N}}$ and denote its limit by η. Suppose that assumptions M, LAN, CM(ii), E and R hold for any sequence $(\gamma_n)_{n\in\mathbb{N}}$ where each $\gamma_n := (\theta_n, \eta_n)_{n\in\mathbb{N}} \subset \Theta' \times \mathcal{H}$ with $\theta_n \to \theta \in \Theta'$. Then, for any compact subset H'_η of H_η,*

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta'} \inf_{h\in H'_\eta} P^n_{(\theta,\eta_n),0,h}(\theta \in \hat{C}_n) \geq 1 - \alpha.$$

Power results

In the scalar case I consider both one-sided tests of the form $H_0 : \theta > \theta_0$ against $H_1 : \theta \leq \theta_0$ and two-sided tests, i.e. $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. These results are essentially standard (Cf. Choi et al., 1996), with the key difference being that here they are stated with γ_n potentially changing with n. Whilst this is a potentially useful strengthening, it simply reflects the corresponding change in the assumptions – i.e. assumption LAN is assumed to hold along such sequences – with the arguments following in the usual way.[31] The first result concerns the power of one-sided tests.

Proposition 1.3.4. *Suppose that assumptions M, LAN, and CM(i) hold. Additionally suppose that H_η is a linear subspace of H and $\tilde{\mathcal{I}}_\gamma > 0$. Then, for any $\alpha \in (0,1)$, any sequence of asymptotically level-α tests $(\psi_n)_{n\in\mathbb{N}}$ for $H_0 : \tau \leq 0$ against $H_1 : \tau > 0$, i.e. any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ such that*

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau,h}\psi_n \leq \alpha \quad \textit{for all } \tau \leq 0,\ h \in H_\eta$$

is subject to the power bound

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\psi_n \leq 1 - \Phi\left(z_\alpha - \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right), \tag{1.15}$$

for all $\tau_n \to \tau > 0$ and $h_n \to h \in H_\eta$ where z_α is the $1-\alpha$ quantile of the standard normal distribution and Φ is the standard normal CDF.

Any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ of asymptotic level α which attains the power bound (1.15) is called "asymptotically locally uniformly most powerful of level-α". The efficient score test attains this bound under the assumptions of section 1.3.1, provided that H_η is a linear subspace and $\tilde{\mathcal{I}}_\gamma > 0$.

Corollary 1.3.5. *Suppose that assumptions M, LAN, CM(ii), E hold, with $\gamma_n = (\theta_0, \eta_n) \to (\theta_0, \eta) = \gamma$. Additionally suppose that H_η is a linear subspace of H, $\tilde{\mathcal{I}}_\gamma > 0$ and $\alpha \in (0,1)$.*

[31] In particular the proofs are based on convergence of a particular sequence of experiments to a Gaussian shift limit experiment. The construction of the relevant sequence of experiments is given in section B.

Then the sequence of tests $(\phi_{n,\theta_0})_{n\in\mathbb{N}}$ is asymptotically locally uniformly most powerful of level-α for the hypothesis $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, i.e. it is asymptotically level$-\alpha$ and achieves the power bound in (1.15) *for any $\tau_n \to \tau > 0$ and any $h_n \to h \in H_\eta$.*

A similar result holds for two-sided tests, with the claim of optimality holding in the class of tests which are (asymptotically) unbiased and of level-α.

Proposition 1.3.6. *Suppose that assumptions M, LAN, CM(i) hold. Additionally suppose that H_η is a linear subspace of H and $\tilde{\mathcal{I}}_\gamma > 0$. Then, for any $\alpha \in (0, 1)$, any sequence of asymptotically unbiased, level-α tests $(\psi_n)_{n\in\mathbb{N}}$ for $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$, i.e. any sequence of tests $\psi_n : \mathcal{W}^n \to [0, 1]$ such that*

$$\limsup_{n\to\infty} P^n_{\gamma_n,0,h}\psi_n \leq \alpha \quad \textit{for all } h \in \mathfrak{H}_\gamma,$$

and

$$\liminf_{n\to\infty} P^n_{\gamma_n,\tau,h}\psi_n \geq \alpha \quad \textit{for all } \tau \neq 0,\ h \in H_\eta$$

is subject to the power bound

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\psi_n \leq 1 - \Phi\left(z_{\alpha/2} - \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right) + 1 - \Phi\left(z_{\alpha/2} + \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right) \tag{1.16}$$

for all $\tau_n \to \tau \neq 0$ and $h_n \to h \in H_\eta$, where z_α is the $1 - \alpha$ quantile of the standard normal distribution and Φ is the standard normal CDF.

Any asymptotically unbiased sequence of tests $\psi_n : \mathcal{W}^n \to [0, 1]$ of asymptotic level α which attains the power bound (1.15) is called "asymptotically locally uniformly most powerful unbiased of level-α". The efficient score test attains this bound under the same assumptions as for the one-sided case.

Corollary 1.3.7. *Suppose that assumptions M, LAN, CM(ii) and E hold, with $\gamma_n = (\theta_0, \eta_n) \to (\theta_0, \eta) = \gamma$. Additionally suppose that H_η is a linear subspace of H, $\tilde{\mathcal{I}}_\gamma > 0$ and $\alpha \in (0, 1)$. Then the sequence of tests $(\phi_{n,\theta_0})_{n\in\mathbb{N}}$ is asymptotically locally uniformly most powerful unbiased of level-α for the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, i.e. it is asymptotically unbiased and of level-α and achieves the power bound in* (1.16) *for any $\tau_n \to \tau \neq 0$ and any $h_n \to h \in H_\eta$.*

For multivariate hypotheses I consider maximin optimality.[32] The difference between the power bound given here and what might be called the "usual" case (Cf. Theorem 13.5.4 of Lehmann and Romano (2005) for the parametric case) is that I do not require the efficient information matrix to be positive definite. Rather I consider a restricted class of directions along which θ may be approached. Specifically, letting $N(\tilde{\mathcal{I}}_\gamma)$ denote the nullspace of $\tilde{\mathcal{I}}_\gamma$, the permitted directions are $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ rather than $\tau \in \mathbb{R}^{d_\theta}$. Note that these coincide

[32] For an alternative approach which restricts the class of tests to those satisfying a rotation invariance condition see Choi et al. (1996).

if (and only if) $\tilde{\mathcal{I}}_\gamma \succ 0$ and hence the "usual" case is a special case of this result. The generalisation given here is useful for models in which the parameter of interest may be underidentified.[33]

Proposition 1.3.8. *Suppose that assumptions M, LAN and CM(i) hold. Additionally suppose that H_η is a linear subspace of H and $r := \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$. Then, for any $\alpha \in (0,1)$, any sequence of asymptotically level-α tests $(\psi_n)_{n\in\mathbb{N}}$ for $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$, i.e. any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ such that*

$$\limsup_{n\to\infty} P^n_{\gamma_n,0,h}\psi_n \leq \alpha \quad \textit{for all } h \in H_\eta$$

is subject to the power bound

$$\limsup_{n\to\infty} \inf_{(\tau,h)\in M_a} P^n_{\gamma_n,\tau,h}\psi_n \leq 1 - \mathrm{P}\left(\chi^2_r(a) \leq c_{r,\alpha}\right), \tag{1.17}$$

for all $a > 0$, where $M_a := \{(\tau,h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta : \tau'\tilde{\mathcal{I}}_\gamma\tau \geq a\}$, $c_{r,\alpha}$ is the $1-\alpha$ quantile of the χ^2_r distribution and $\chi^2_r(a)$ denotes a non-central χ^2 random variable with r degrees of freedom and non-centrality a.

Any sequence of tests $\psi_n : \mathcal{W}^n \to [0,1]$ of asymptotic level α which attains the power bound (1.15) over all compact subsets of M_a is called "asymptotically maximin of level-α".[34] The efficient score test is asymptotically maximin of level-α under the assumptions in section 1.3.1, provided that H_η is a linear subspace and $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$.

Corollary 1.3.9. *Suppose that assumptions M, LAN, CM(ii), E and R hold, with $\gamma_n = (\theta_0, \eta_n) \to (\theta_0, \eta) = \gamma$. Additionally suppose that H_η is a linear subspace of H, $r := \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$ and $\alpha \in (0,1)$. Then the sequence of tests $(\phi_{n,\theta_0})_{n\in\mathbb{N}}$ is asymptotically maximin of level-α for the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ over all compacts, in the sense that for any compact $K_a \subset M_a$*

$$\lim_{n\to\infty} \inf_{(\tau,h)\in K_a} P^n_{\gamma_n,\tau,h}\phi_{n,\theta_0} = 1 - \mathrm{P}\left(\chi^2_r(a) \leq c_{r,\alpha}\right). \tag{1.18}$$

There are two key takeaways from this result. Firstly, when the efficient information matrix is rank deficient, the efficient score test continues to enjoy non-trivial power in certain directions.[35] Secondly the power it achieves is – in a certain sense – optimal.[36]

[33] For details of the construction of the sequence of experiments used to establish this result see appendix section B.

[34] Cf. Section 13.5.3 of Lehmann and Romano (2005) for the terminology

[35] This is demonstrated in a specific example in section 1.5.5.

[36] Nevertheless, if one has a particular direction against which one wishes to direct power, or – more generally – a weighting function over alternatives, a criterion based on weighted average power would seem more appropriate. Cf. e.g. Elliott et al. (2015); Montiel Olea (2020).

1.3.4. Sufficient conditions for the assumptions

In the i.i.d. setting it is well known that differentiability in quadratic mean (e.g. van der Vaart, 2002, Definition 1.6) is a sufficient condition for a LAN expansion like that in equation (1.5) with a fixed $\gamma \in \Gamma$ (e.g. Bickel et al., 1998; Le Cam and Yang, 2000; van der Vaart, 2002). In the setting of interest here, a suitably adapted version of this condition also suffices for assumption LAN.[37]

Assumption DQM (Differentiability in quadratic mean). *Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in Γ which converges to a point $\gamma \in \Gamma$ and H_η a subset of a Banach space, H, which includes 0.*

For any sequence $\tau_n \to \tau$ with each $\tau_n, \tau \in \mathbb{R}^{d_\theta}$, any sequence $h_n \to h$ with $h_n, h \in H_\eta$, a convergent sequence of $d_\theta \times d_\theta$ matrices δ_n and sequences $\eta_n(h_n) \to \eta$ with each $\eta_n(h_n) \in \mathcal{H}$, define $\gamma_n(\tau_n, h_n)$ as in assumption LAN and suppose that

1. *the sequence $(P_{\gamma_n(\tau_n,h_n)})_{n\geq 1}$ is (eventually) in $\mathcal{P}$,*
2. *for some sequence of measurable functions $(g_n)_{n\in\mathbb{N}}$ such that $(g_n^2)_{n\in\mathbb{N}}$ are uniformly P_{γ_n}-integrable and $P_{\gamma_n} g_n = o(n^{-1/2})$,*

$$\int \left[\sqrt{n}\left(\sqrt{p_{\gamma_n(\tau_n,h_n)}} - \sqrt{p_{\gamma_n}}\right) - \frac{1}{2} g_n \sqrt{p_{\gamma_n}}\right]^2 \mathrm{d}\nu \to 0. \tag{1.19}$$

◇

Proposition 1.3.10. *Suppose assumptions M and DQM hold. Moreover suppose that for a sequence of functions $(\dot{\ell}_{\gamma_n})_{n\in\mathbb{N}}$ with each $\dot{\ell}_{\gamma_n} \in L_2^0(P_{\gamma_n})$ and a sequence of linear maps $(B_{\gamma_n})_{n\in\mathbb{N}}$ with each $B_{\gamma_n} : H_\eta \to L_2^0(P_{\gamma_n})$,*

$$P_{\gamma_n}\left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h - g_n\right]^2 \to 0.$$

Then assumption LAN holds.

The addtional condition in the display in proposition 1.3.10 allows DQM to be shown with any sequence g_n such that the L_2 distance between g_n and the scores $\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ vanishes as $n \to \infty$.

I next record two conditions useful for checking the integral convergence required in CM(ii), once the uniform square P_{γ_n}-integrability has been established. The first can be obtained as an immediate corollary of a (stronger) result of Feinberg et al. (2016), who establish a uniform (over Borel sets) version of the integral convergence. The second is effectively the standard result that weak convergence and uniform integrability imply convergence of

[37] Results of this nature are known to hold see e.g. Strasser (1985, Chapter 74) or van der Vaart (1988b, A.2). I provide this formulation to facilitate the demonstration of the version of LAN assumed in this paper.

moments, where the condition of continuous convergence is imposed to ensure the weak convergence of the appropriate laws.

Lemma 1.3.11. *Suppose that $(P_n)_{n\in\mathbb{N}}$ is a sequence of probability measures which converges in total variation to P.*[38] *If $(f_n)_{n\in\mathbb{N}}$ is a sequence of functions in $L_1(P_n)$ such that (a) $f_n \xrightarrow{P} f \in L_1(P)$ and (b) $(f_n)_{n\in\mathbb{N}}$ is uniformly P_n-integrable, then $P_n f_n \to Pf$.*

Lemma 1.3.12. *Let S be a metric space and suppose that $(P_n)_{n\in\mathbb{N}}$ is a sequence of measures on $(S, \mathcal{B}(S))$ which converge weakly to P. Suppose that $(f_n)_{n\in\mathbb{N}}$ is a sequence of real-valued functions with each $f_n \in L_1(P_n)$ which (a) converge continuously to $f \in L_1(P)$ and (b) are uniformly P_n-integrable.*[39] *Then $P_n f_n \to Pf$.*

Assumption R requires the estimate of the efficient information matrix, $\hat{\mathcal{I}}_{n,\theta_n}$, to have the same rank as $\tilde{\mathcal{I}}_\gamma$ with P_{γ_n}-probability approaching one. The following construction is sufficient to guarantee this; it requires knowledge of the rate of convergence to zero of the difference (in the spectral norm) of an estimator $\check{\mathcal{I}}_{n,\theta_n}$ and a matrix $\mathcal{I}_n$ where $\mathcal{I}_n \to \tilde{\mathcal{I}}_\gamma$ and $\operatorname{rank}(\mathcal{I}_n) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ for all sufficiently large n. As there is nothing special about the limit being the efficient information matrix here, the construction is given more generally.[40]

In particular, suppose that the sequence of (random) positive semi-definite (symmetric) matrices $(\check{M}_n)_{n\in\mathbb{N}}$ (of fixed dimension $L \times L$) satisfy

$$P_n\left(\|\check{M}_n - M_n\|_2 < \nu_n\right) \to 1, \tag{1.20}$$

for a sequence $(P_n)_{n\in\mathbb{N}}$ of probability measures, a known non-negative sequence $\nu_n \to 0$ and a sequence of deterministic matrices $M_n \to M$ with $\operatorname{rank}(M_n) = \operatorname{rank}(M)$ for all sufficiently large n.[41] Let $\check{M}_n = \check{U}_n \check{\Lambda}_n \check{U}_n'$ be the corresponding eigendecompositions and define

$$\hat{M}_n := \check{U}_n \Lambda_n(\nu_n) \check{U}_n' \ , \tag{1.21}$$

where $\Lambda_n(\nu_n)$ is a diagonal matrix with the ν_n-truncated eigenvalues of $\check{M}_n$ on the main diagonal and $\check{U}_n$ is the matrix of corresponding orthonormal eigenvectors. That is, if $(\check{\lambda}_{n,i})_{i=1}^L$ denote the non-increasing eigenvalues of $\check{M}_n$, then the (i,i)-th element of $\Lambda_n(\nu_n)$ is $\check{\lambda}_{n,i}\mathbf{1}(\check{\lambda}_{n,i} \geq \nu_n)$.

Proposition 1.3.13. *If* (1.20) *holds, $M_n \to M$ and for all n greater than some $N \in \mathbb{N}$*

[38] Each P_n and P are defined on a common measurable space $(S, \mathcal{B}(S))$.

[39] Continuous convergence requires $f_n(s_n) \to f(s)$ for all $(s_n)_{n\in\mathbb{N}} \subset S$ with $s_n \to s \in S$. Here this is equivalent to compact convergence of the f_n to a continuous limit f (cf. Remmert, 1991, Chapter 3, §1, Section 5).

[40] A similar construction appears as part of Theorem 2 in Lee and Mesters (2022a). If the (non-zero) eigenvalues of $\tilde{\mathcal{I}}_\gamma$ can be computed, a simpler truncation approach can be utilised, cf. Proposition 2 in Lütkepohl and Burda (1997).

[41] (1.20) is implied by $\|\check{M}_n - M_n\| = o_{P_{\gamma_n}}(\nu_n)$ for any matrix norm. Moreover, the existence of such a sequence $(\nu_n)_{n\in\mathbb{N}}$ is guaranteed if $\|\check{M}_n - M_n\|_2 \to 0$ in P_n-probability, however its explicit knowledge is necessary to perform the subsequent construction.

$\mathrm{rank}(M_n) = \mathrm{rank}(M)$, *then* $\hat{M}_n \xrightarrow{P_n} M$ *and*

$$P_n\left(\mathrm{rank}(\hat{M}_n) = \mathrm{rank}(M)\right) \to 1, \tag{1.22}$$

where $\hat{M}_n$ *is defined as in* (1.21)*.*

Assumption T. *Let* $(\gamma_n)_{n\in\mathbb{N}}$ *be a sequence in* Γ *with a limit* $\gamma \in \Gamma$, $(\tilde{\mathcal{I}}_n)_{n\in\mathbb{N}}$ *a deterministic sequence of matrices with* $\tilde{\mathcal{I}}_n \to \tilde{\mathcal{I}}_\gamma$ *and* $\mathrm{rank}(\tilde{\mathcal{I}}_n) = \mathrm{rank}(\tilde{\mathcal{I}}_\gamma)$ *for all* n *exceeding some* $N \in \mathbb{N}$ *and suppose that the sequence* $(\check{\mathcal{I}}_{n,\theta_n})_{n\in\mathbb{N}}$ *satisfies*

$$P_{\gamma_n}\left(\|\check{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_n\|_2 < \nu_n\right) \to 1. \tag{1.23}$$

◇

Corollary 1.3.14. *If assumption T holds, the estimate* $\hat{\mathcal{I}}_{n,\theta_n}$ *formed by truncating the eigendecompositions of* $\check{\mathcal{I}}_{n,\theta_n}$ *at* ν_n*, as in* (1.21)*, satisfies equation* (1.7) *and assumption R.*

In practice equation (1.23) is likely to be established by demonstrating that $\|\check{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_n\| = o_{P_{\gamma_n}}(\nu_n)$.[42] As this condition concerns only asymptotic behaviour, there is wide scope for different possible sequences which have the same asymptotic behaviour but rather different behaviour in finite samples. Simulation experiments designed to replicate various possible DGPs for the case under consideration may provide some guidance.

1.4. Single index model

In this section I provide details of the application of the theory of section 1.3 to a more general version of the single index model in example 1.

Consider the single index (regression) model (SIM), where $W = (Y, X)$ with

$$Y = f(X_1 + X_2'\theta) + \epsilon, \quad \mathbb{E}[\epsilon|X] = 0, \tag{1.24}$$

for $X = (X_1, X_2) \in \mathbb{R}^K$ a vector of covariates such that $(\epsilon, X) \sim \zeta$ for some Lebesgue density ζ and some unknown link function f.[43] As recorded in Theorem 2.1 of Horowitz (2009), f and θ are identified in this model if f is differentiable, not constant on the support of $X_1 + X_2'\theta$ and the support of X is not contained in a proper linear subspace of $\mathbb{R}^K$. By utilising the inferential approach developed in section 1.3, this section provides an

[42] For any matrix norm $\|\cdot\|$.

[43] This particular specification of the single index model is relatively simple. More complex versions of this model (e.g. with a more general index specification or a linear component $Z'\xi$) could be analysed using similar techniques. The form used here is deliberately chosen to retain only the key aspect of the model relevant to this paper: that θ may be unidentified or weakly identified for certain values of f, an infinite dimensional nuisance parameter.

inferential approach for θ in model (1.24) which is robust to failure of these assumptions, and – perhaps more importantly – robust in a setting where f is relatively flat when compared with sampling variation, leading to weak identification of θ.

The first step of the analysis is to formally specify the model under consideration and establish some primitive assumptions under which the results will be obtained. The basic model setup is given by the following assumption.

Assumption SIM. *Suppose that $W = (Y, X) \in \mathbb{R}^{1+K}$ satisfies* (1.24) *and*

1. $\Theta \subset \mathbb{R}^{d_\theta}$ *is open,*
2. $(\epsilon, X) \sim \zeta$ *where* $\zeta \in \mathscr{Z}$,
3. $f \in \mathscr{F}$,

where $\mathscr{Z}$ and $\mathscr{F}$ are defined as follows. Let $\mathscr{X} \subset \mathbb{R}^K$ be closed, $\phi(\epsilon, X) := \frac{\partial \log \zeta(e,X)}{\partial e}(\epsilon, X)$ the log-density score in the first argument of ζ and $\rho > 0$. Then $\mathscr{Z}$ is the collection:

$$\mathscr{Z} := \left\{ \zeta \in L_1(\mathbb{R}^{1+K}) : \zeta \geq 0, \int_{\mathbb{R}\times\mathscr{X}} \zeta \, \mathrm{d}\lambda = 1, \text{ if } (e, Z) \sim \zeta \text{ then } (1.26), \ \zeta \text{ satisfies } (1.25) \right\},$$

where $L_1(\mathbb{R}^{1+K})$ is the space of Lebesgue integrable functions on $\mathbb{R}^{1+K}$ and

$$e \mapsto \sqrt{\zeta(e, X)} \text{ is continuously differentiable } \lambda - a.e., \tag{1.25}$$

$$\mathbb{E}[\epsilon|X] = 0, \quad \mathbb{E}[(\phi(\epsilon, X)^{2+\rho} + 1)\|X\|_2^{2+\rho}] < \infty, \quad \mathbb{E}[XX'] \succ 0. \tag{1.26}$$

$\mathscr{F} := C_b^1(\mathscr{D})$ is the class of functions which are bounded and continuously differentiable with bounded derivative λ-a.e. on $\mathscr{D} := \{X_1 + X_2'\theta : \theta \in \Theta, x \in \mathscr{X}\}$.

The model is given by $\mathcal{P} = \{P_\gamma : \gamma \in \Gamma\}$ for $\Gamma = \Theta \times \mathcal{H}$ with $\mathcal{H} = \mathscr{F} \times \mathscr{Z}$ where each P_γ is the probability measure on $\mathbb{R}^{1+K}$ corresponding to the Lebesgue density $p_\gamma(W) = \zeta(Y - f(X_1 + X_2'\theta), X)$. ◇

Part 2 of the preceding assumption restricts the class of density functions which govern the distribution of the error term and covariates in (1.24). The key restrictions it imposes are (a) the required conditional mean restriction $\mathbb{E}[\epsilon|X] = 0$, (b) the existence of some moments of specific functions of the data, and (c) a smoothness condition on the density function. Part 3 restricts the link function f to belong to a specified class of functions; the restrictions imposed on f by this assumption are relatively weak and common in the literature on single index models.[44] Note that these restrictions do not rule out f being constant on $\mathscr{D}$: if $f(v) = c$ for all $v \in \mathscr{D}$ and some $c \in \mathbb{R}$, $f \in \mathscr{F}$.

[44] Cf. Assumption 4.1 in Newey and Stoker (1993); Assumptions A0 – A2 in Kuchibhotla and Patra (2020).

1.4.1. Verification of the modelling assumptions

Given a random sample $(W_i)_{i=1}^n$ satisfying assumption SIM, assumption M holds. To establish assumptions LAN and CM(ii) I first need to specify the local perturbations to the nuisance parameter η for which the quadratic approximation will hold.

The considered local perturbations to the nuisance parameters take the form

$$\eta_n(h) := (f + t_n h_1,\ \zeta(1 + t_n h_2)), \quad t_n = n^{-1/2}, \tag{1.27}$$

with $h_1 \in \dot{\mathscr{F}} := C_b^1(\mathscr{D})$, the set of real valued functions on $\mathbb{R}$ which are continuously differentiable and bounded λ-a.e. on $\mathscr{D}$, and $h_2 \in \dot{\mathscr{Z}}_\eta$ where

$$\dot{\mathscr{Z}}_\eta := \left\{ h_2 \in C_b^{1|1}(\mathbb{R}^{1+K}) : \mathbb{E}[h_2(\epsilon, Z)] = 0,\ \mathbb{E}[\epsilon h_2(\epsilon, X)|X] = 0 \text{ if } (\epsilon, X) \sim \zeta \right\},$$

for $C_b^{1|1}(\mathbb{R}^{1+K})$ is the space of functions $h_2 : \mathbb{R}^{1+K} \to \mathbb{R}$ which are bounded λ-a.e. and such that $e \mapsto h_2(e, X)$ is continuously differentiable with bounded derivative λ-a.e.. The perturbation directions for η are $H_\eta := \dot{\mathscr{F}} \times \dot{\mathscr{Z}}_\eta$ which is a linear subspace of $L_\infty(\lambda) \times L_\infty(G) =: H$, for λ the Lebesgue measure on $\mathbb{R}$. Equip H with the norm $\|h\| = \|h_1\|_{\lambda,\infty} + \|h_2\|_{G,\infty}$.

I now establish that the model is differentiable in quadratic mean and hence (by Proposition 1.3.10) locally asymptotically normal.

Proposition 1.4.1. *Suppose that assumption SIM holds, $\theta_n \to \theta \in \Theta$ and $\eta \in \mathcal{H}$ and consider the sequence defined by $\gamma_n = (\theta_n, \eta) \in \Gamma$. Let $\delta_n = I/\sqrt{n}$, $\tau_n \to \tau$, $h_n \in H_\eta$ with $h_n \to h \in H_\eta$ and define $\eta_n : H_\eta \to \mathcal{H}$ as in* (1.27)*. Then assumption DQM holds with score functions $g_n = \tau \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ where for $V_{\theta_n} := X_1 + X_2'\theta_n$, $e_n := Y - f(V_{\theta_n})$,*

$$\begin{aligned} \dot{\ell}_{\gamma_n}(W) &:= -\phi(e_n, X) f'(V_{\theta_n}) X_2 \\ [B_{\gamma_n} h](W) &:= -\phi(e_n, X) h_1(V_{\theta_n}) + h_2(e_n, X). \end{aligned}$$

The efficient score function for this model was derived by Newey and Stoker (1993) and is given in the following Proposition.

Proposition 1.4.2. *Consider the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of Proposition 1.4.1, suppose that assumption SIM holds and*

$$\mathbb{E}[\epsilon\phi(\epsilon, X)|X] = -1, \quad \mathbb{E}[\phi(\epsilon, X)^2|X] < C < \infty, \quad 0 < c < \mathbb{E}[\epsilon^2|X] < C < \infty. \tag{1.28}$$

Additionally suppose there exists a function $\tilde{m} : \mathbb{R} \to \mathbb{R}$ which is bounded and continuously differentiable with bounded derivative such that $\mathbb{E}[\epsilon\tilde{m}(\epsilon)|X]$ is bounded away from zero uniformly in X. Then assumption CM(ii) holds and for $\omega(X) := \mathbb{E}[\epsilon^2|X]^{-1}$ the efficient *score function is*

$$\tilde{\ell}_{\gamma_n} := \omega(X)(Y - f(V_{\theta_n}))f'(V_{\theta_n}) \left[X_2 - \frac{\mathbb{E}\left[\omega(X)X_2|V_{\theta_n}\right]}{\mathbb{E}\left[\omega(X)|V_{\theta_n}\right]} \right].$$

The (conditional) moment conditions in (1.28) are standard. The first is a particular case of the (conditional) generalised information equality; it will hold provided differentiation and integration can be interchanged appropriately. The second and third provide uniform bounds on some conditional expectation functions. Existence of the function $\tilde{m}$ is a weak condition; see Assumption 4.2 and the subsequent discussion in Newey and Stoker (1993, p. 1210).

1.4.2. Implementation of the efficient score test

I now consider estimation of the efficient score function just described in order to satisfy assumptions E and R. Estimation in the (conditionally) heteroskedastic case introduces technical difficulties which are essentially unrelated to the problem studied in this paper and therefore I initially focus on the (conditionally) homoskedastic case and subsequently note that this belongs to a more general class of statistics which remain robust under heteroskedasticity though are typically not power optimal.[45]

Suppose that $\sigma^2 := \mathbb{E}[\epsilon^2|X] = \mathbb{E}[\epsilon^2] > 0$. Under this simplification, the efficient score function is:

$$\tilde{\ell}_{\gamma_n} := \sigma^{-2}(Y - f(V_{\theta_n})f'(V_{\theta_n}) \left[X_2 - Z(V_{\theta_n})\right],$$

where $Z(V_{\theta_n}) := \mathbb{E}\left[X_2|V_{\theta_n}\right]$.

To estimate the nonparametric parts of the efficient score function I will use split-sample estimators. Let $N^{(1)} = \{1, \ldots, \lfloor n/2 \rfloor\}$ and $N^{(2)} = [n] \setminus N^{(1)}$. For $i \in [n]$ let N_{-i} denote whichever of $N^{(1)}$ or $N^{(2)}$ that does *not* contain i. The class of estimators considered have the following form:

$$\begin{aligned}
\hat{f}_{n,i} &:= \hat{f}_n(V_{\theta_n,i}) := \check{f}_n(V_{\theta_n,i}, \hat{\xi}_{1,n,i}) & \hat{\xi}_{1,n,i} &:= \xi_{1,n}((W_j)_{j \in N_{-i}}), \\
\widehat{f'}_{n,i} &:= \widehat{f'}_n(V_{\theta_n,i}) := \check{f'}_n(V_{\theta_n,i}, \hat{\xi}_{2,n,i}) & \hat{\xi}_{2,n,i} &:= \xi_{2,n}((W_j)_{j \in N_{-i}}), \\
\hat{Z}_{n,i} &:= \hat{Z}_n(V_{\theta_n,i}) := \check{Z}_n(V_{\theta_n,i}, \hat{\xi}_{3,n,i}) & \hat{\xi}_{3,n,i} &:= \xi_{3,n}((W_j)_{j \in N_{-i}}),
\end{aligned} \tag{1.29}$$

where each $\hat{\xi}_{j,n,i}$ is a (random) vector whose dimension may increase with the sample size. This class of estimators includes, for example, series estimators (of conditional moment functions and their derivatives) as considered by e.g. Newey (1997); Belloni et al. (2015);

[45] The class contains a member which achieves the power bound under appropriate conditions but is not feasible as it requires knowledge of the optimal weighting function $\omega(X)$. Cf. the approach taken for estimation in Ichimura (1993).

Chen and Christensen (2015); Cattaneo et al. (2020).[46] In this case, e.g. $f(V_{\theta_n})$ is the conditional expectation of Y given V_{θ_n} and estimates of $f(V_{\theta_n,i})$ and $\widehat{f'}(V_{\theta_n,i})$ can be given as

$$\hat{f}_n(V_{\theta_n,i}) = \check{f}_n(v, \hat{\xi}_{1,n,i}) = q_n(V_{\theta_n,i})'\hat{\xi}_{1,n,i}, \quad \widehat{f'}_n(V_{\theta_n,i}) = \check{f'}_n(V_{\theta_n,i}, \hat{\xi}_{2,n,i}) = \left[q_n'(V_{\theta_n,i})\right]' \hat{\xi}_{2,n,i},$$

where q_n is a K_n-vector of basis functions from $\mathbb{R} \to \mathbb{R}$, q_n' their derivatives and

$$\hat{\xi}_{1,n,i} = \hat{\xi}_{2,n,i} = \left(\sum_{j \in N_{-i}} q_n(V_{\theta_n,j}) q_n(V_{\theta_n,j})' \right)^{-1} \left(\sum_{j \in N_{-i}} q_n(V_{\theta_n,j}) Y_j \right).$$

Similar estimators can be constructed for $Z(V_{\theta_n})$ which is the conditional expectation of X_2 given V_{θ_n}.

Given such estimators I form an estimate of σ^2 as

$$\hat{\sigma}_n^2 \coloneqq \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_{n,i})^2,$$

and the estimates

$$\hat{\ell}_{n,\theta_n}(W_i) \coloneqq \hat{\sigma}_n^{-2} \left(Y_i - \hat{f}_{n,i} \right) \widehat{f'}_{n,i} \left[X_{2,i} - \hat{Z}_{n,i} \right], \quad \check{\mathcal{I}}_{n,\theta_n} \coloneqq \frac{1}{n} \sum_{i=1}^n \hat{\ell}_{n,\theta_n}(W_i) \hat{\ell}_{n,\theta_n}(W_i)'. \tag{1.30}$$

Let $\hat{\mathcal{I}}_{n,\theta_n}$ be the eigendecomposition-truncated version of $\check{\mathcal{I}}_{n,\theta_n}$ at ν_n (analogously to (1.21)), where $(\nu_n)_{n \in \mathbb{N}}$ is a non-negative sequence converging to zero. With these estimators assumptions E and R can be shown to hold under conditions on the sequence $(\nu_n)_{n \in \mathbb{N}}$ and the following high-level condition which assumes certain (probabilistic) rates of convergence hold for

$$\begin{aligned}
\mathcal{R}_{1,n,i} &\coloneqq \left(\int \left[\check{f}_n(v, \hat{\xi}_{1,n,i}) - f(v) \right]^2 \mathrm{d}\mathcal{V}_n(v) \right)^{1/2}, \\
\mathcal{R}_{2,n,i} &\coloneqq \left(\int \left[\check{f'}_n(v, \hat{\xi}_{2,n,i}) - f'(v) \right]^2 \mathrm{d}\mathcal{V}_n(v) \right)^{1/2}, \\
\mathcal{R}_{3,n,i} &\coloneqq \left(\int \left\| \check{Z}_n(v, \hat{\xi}_{3,n,i}) - Z(v)) \right\|_2^2 \mathrm{d}\mathcal{V}_n(v) \right)^{1/2},
\end{aligned}$$

where $\mathcal{V}_n$ is the distribution of V_{θ_n}.

Assumption SIM-NP(i). *Suppose that $\mathscr{X}$ is a compact set, equation* (1.28) *holds,* $\sigma^2 \coloneqq \mathbb{E}[\epsilon^2|X] = \mathbb{E}[\epsilon^2]$, $\mathbb{E}[\epsilon^4] < \infty$ *and with P_{γ_n}-probability approaching one for $l \in$* $[3]$ *and each $i \in [n]$, $\mathcal{R}_{l,n,i} \leq r_n = o(n^{-1/4})$.* ◇

The rates in assumption SIM-NP(i) are attainable under reasonable regularity conditions.

[46] This class of estimators also includes, for example, kernel estimators.

For example, series (linear sieve) estimators of f, f' and Z can attain these rates given sufficient smoothness of the target function and other regularity conditions. See, inter alia, Belloni et al. (2015); Chen and Christensen (2015); Cattaneo et al. (2020); Huang and Su (2021). This assumption is sufficient for the estimator of σ^{-2} to be $\sqrt{n}$-consistent.

Lemma 1.4.3. *Suppose that assumption SIM holds and $\sigma^2 := \mathbb{E}[\epsilon^2|X] = \mathbb{E}[\epsilon^2] \in (0, \infty)$ and let $(\gamma_n)_{n\in\mathbb{N}}$ be as in Proposition 1.4.1. If $\mathbb{E}[\epsilon^4] < \infty$ and with P_{γ_n}-probability approaching one, $\mathcal{R}_{1,n,i} \leq r_n = o(n^{-1/4})$, then $\sqrt{n}(\hat{\sigma}_n^{-2} - \sigma^{-2}) = O_{P_{\gamma_n}}(1)$.*

In the general, heteroskedastic, case I consider a related estimator, where – as in Ichimura (1993) – a *known* weighting function $\breve{\omega}(X)$ is utilised in place of the unknown $\omega(X)$. In particular, I estimate the function

$$\breve{\ell}_{\gamma_n}(W) := \breve{\omega}(X)(Y - f(V_{\theta_n}))f'(V_{\theta_n}) \left[X_2 - \frac{\mathbb{E}\left[\breve{\omega}(X)X_2|V_{\theta_n}\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_{\theta_n}\right]}\right].$$

Clearly if $\breve{\omega} = \omega$, $\breve{\ell}_{\gamma_n}$ coincides with the efficient score function and hence power optimality results are available if the conditions outlined in section 1.3 hold. In the case where $\breve{\omega} \neq \omega$ the resulting statistic will not be power optimal, but will retain the locally uniform size control properties of the efficient score statistic.

In the heteroskedastic case, I replace the function $Z(V_{\theta_n}) := \mathbb{E}[X_2|V_{\theta_n}]$ with $Z_1(V_{\theta_n})/Z_2(V_{\theta_n})$ where $Z_1(V_{\theta_n}) := \mathbb{E}[\breve{\omega}(X)X_2|V_{\theta_n}]$ and $Z_2(V_{\theta_n}) := \mathbb{E}[\breve{\omega}(X)|V_{\theta_n}]$. Let $\hat{f}_{n,i}$ and $\widehat{f'}_{n,i}$ be as in (1.29) and similarly define

$$\begin{aligned} \hat{Z}_{1,n,i} &:= \hat{Z}_{1,n}(V_{\theta_n,i}) := \check{Z}_{1,n}(V_{\theta_n,i}, \hat{\xi}_{3,n,i}) \qquad \hat{\xi}_{3,n,i} := \xi_{3,n}((W_j)_{j\in N_{-i}}) \\ \hat{Z}_{2,n,i} &:= \hat{Z}_{2,n}(V_{\theta_n,i}) := \check{Z}_{2,n}(V_{\theta_n,i}, \hat{\xi}_{4,n,i}) \qquad \hat{\xi}_{4,n,i} := \xi_{4,n}((W_j)_{j\in N_{-i}}). \end{aligned} \tag{1.31}$$

With these estimates I can form an estimate of $\breve{\ell}_{\gamma_n}$ and $\Upsilon_{\gamma_n} := P_{\gamma_n}\breve{\ell}_{\gamma_n}\breve{\ell}'_{\gamma_n}$ according to

$$\breve{\ell}_{n,\theta_n}(W_i) := \breve{\omega}(X_i)\left(Y_i - \hat{f}_{n,i}\right)\widehat{f'}_{n,i}\left[X_{2,i} - \frac{\hat{Z}_{1,n,i}}{\hat{Z}_{2,n,i}}\right], \quad \hat{\Upsilon}_{n,\theta_n} := \frac{1}{n}\sum_{i=1}^{n}\breve{\ell}_{n,\theta_n}(W_i)\breve{\ell}_{n,\theta_n}(W_i)'. \tag{1.32}$$

Let $\check{\Upsilon}_{n,\theta_n}$ be the eigendecomposition-truncated version of $\hat{\Upsilon}_{n,\theta_n}$ at ν_n (analogously to (1.21)). The test statistic that will be used in this case (for testing a two-sided hypothesis) is

$$\check{S}_{n,\theta} := \left(\sqrt{n}\mathbb{P}_n\breve{\ell}_{n,\theta}\right)' \check{\Upsilon}^{\dagger}_{n,\theta}\left(\sqrt{n}\mathbb{P}_n\breve{\ell}_{n,\theta}\right), \tag{1.33}$$

with the test and confidence then being defined analogously to (1.10) and (1.11) with $\check{S}_{n,\theta}$ in place of $\hat{S}_{n,\theta}$. Denote these respectively by $\check{\phi}_{n,\theta_0}$ and $\check{C}_n$. This test will be called the "pseudo efficient score test" in what follows. Let $\breve{\mathcal{R}}_{l,n,i} := \mathcal{R}_{l,n,i}$ for $l = 1, 2$ and define

$$\breve{\mathcal{R}}_{3,n,i} := \left(\int \left\| \breve{Z}_{1,n}(v, \hat{\xi}_{3,n,i}) - Z_1(v)) \right\|_2^2 \, \mathrm{d}\mathcal{V}_n(v) \right)^{1/2}$$
$$\breve{\mathcal{R}}_{4,n,i} := \left(\int \left(\breve{Z}_{2,n}(v, \hat{\xi}_{4,n,i}) - Z_2(v)) \right)^2 \, \mathrm{d}\mathcal{V}_n(v) \right)^{1/2} .$$

In the heteroskedastic case, assumption SIM-NP(i) is replaced by the following assumption:

Assumption SIM-NP(ii). *Suppose that $\mathscr{X}$ is a compact set, equation* (1.28) *holds,* $\mathbb{E}[\epsilon^4] < \infty$, $\breve{\omega} : \mathbb{R}^K \to (\underline{\omega}, \overline{\omega})$ *is a known function and with P_{γ_n}-probability approaching one for $l \in [4]$ and each $i \in [n]$, $\breve{\mathcal{R}}_{l,n,i} \leq r_n = o(n^{-1/4})$.* $\diamond$

The rates required by this assumption are attainable under reasonable regularity conditions; cf. the discussion following assumption SIM-NP(i).

1.4.3. Asymptotic properties

I start by detailing the asymptotic properties of the efficient score statistic in the homoskedastic case.

Proposition 1.4.4. *Suppose that assumptions SIM, SIM-NP(i) hold and there exists a function $\tilde{m}$ as in Proposition 1.4.2. Consider the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of proposition 1.4.1, suppose the observations form an i.i.d. sample and $\hat{\ell}_{n,\theta_n}$ and $\hat{\mathcal{I}}_{n,\theta_n}$ are as in* (1.30)*, with $0 \leq \nu_n \to 0$ such that $r_n + n^{-1/2}\log(n)^{1/2+\kappa} = o(\nu_n)$ for some $\kappa > 0$. Then assumptions M, LAN, CM(ii), E and R hold.*

With the estimators $\hat{\ell}_{n,\theta_n}$ and $\hat{\mathcal{I}}_{n,\theta_n}$ the efficient score statistic, test and confidence set can be defined as in section 1.3.2. The following results demonstrate that the efficient score test is optimal under strong-identification asymptotics and provides robust size control and the corresponding confidence sets robust coverage, including under asymptotics in which the function f is local to a constant (function) at rate $\sqrt{n}$, corresponding to a setting where θ is weakly identified.

Corollary 1.4.5. *In the setting of Proposition 1.4.4, let H_η' be a compact subset of H_η. Then, the efficient score test satisfies*

$$\limsup_{n\to\infty} \sup_{h\in H_\eta'} P^n_{\gamma_n,0,h}\phi_{n,\theta_0} \leq \alpha,$$

and, for any compact $\Theta' \subset \Theta$, the corresponding test inversion confidence sets satisfy

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \inf_{h\in H_\eta'} P^n_{\gamma_n,0,h}(\theta \in \hat{C}_n) \geq 1 - \alpha.$$

Corollary 1.4.6. *In the setting of Proposition 1.4.4, suppose additionally that* $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$*. If* $d_\theta = 1$*, then the efficient score test is locally asymptotically uniformly most powerful unbiased. If* $d_\theta > 1$*, then the efficient score test is locally asymptotically maximin.*

I now establish a similar uniform size control result for the heteroskedastic case, with the psuedo efficient score test defined immediately following (1.33).

Proposition 1.4.7. *Suppose that that assumptions SIM, SIM-NP(ii) hold and there exists a function* $\tilde{m}$ *as in Proposition 1.4.2. Consider the sequence* $(\gamma_n)_{n\in\mathbb{N}}$ *of proposition 1.4.1, suppose the observations form an i.i.d. sample and* $\check{\ell}_{n,\theta_n}$ *and* $\check{\Upsilon}_{n,\theta_n}$ *are as in* (1.32)*, with* $0 \leq \nu_n \to 0$ *such that* $r_n + n^{-1/2}\log(n)^{1/2+\kappa} = o(\nu_n)$ *for some* $\kappa > 0$*. Let* H'_η *be a compact subset of* H_η*. Then, the psuedo efficient score test satisfies*

$$\limsup_{n\to\infty} \sup_{h\in H'_\eta} P^n_{\gamma_n,0,h} \check{\phi}_{n,\theta_0} \leq \alpha,$$

and, for any compact $\Theta' \subset \Theta$*, the corresponding test inversion confidence sets satisfy*

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \inf_{h\in H'_\eta} P^n_{\gamma_n,0,h}(\theta \in \check{C}_n) \geq 1 - \alpha.$$

I remark here that if $\breve{\omega} = \omega$ then each $\check{\ell}_{\gamma_n} = \tilde{\ell}_{\gamma_n}$. In this situation, if the rank of $\Upsilon_\gamma = \tilde{\mathcal{I}}_\gamma$ is positive, then in the setting of Proposition 1.4.7 the (pseudo) efficient score test is is locally asymptotically uniformly most powerful unbiased if $d_\theta = 1$ and locally asymptotically maximin if $d_\theta > 1$. However, as this is infeasible in the heteroskedastic case, I do not state a formal power result.

1.4.4. Simulation study

I conduct a simulation study to examine the finite sample properties of the efficient score test. I draw $n \in \{200, 400, 600, 800\}$ samples from model (1.24) for a number of different functions f and distributions ζ. I set $K = 1$ throughout and examine finite sample size using 5000 Monte Carlo replications, at a nominal level of 5%. In each case I test the null $H_0 : \theta = 1$.

Overall the simulation experiments suggest the asymptotic results of section 1.4.3 provide a good guide to the performance of the efficient score test (and psuedo efficient score test) in finite samples.

Homoskedastic case

Initially I consider the homoskedastic case. The error term is taken as either (1) $\epsilon \sim \mathcal{N}(0, 1)$ or (2) $\epsilon|\xi \sim \sqrt{5}(-1)^\xi \operatorname{Beta}(2, 3)$, $\xi \sim \operatorname{Bernoulli}(1/2)$. In both cases $\mathbb{E}\epsilon = 0$ and $\mathbb{V}\epsilon = 1$.

The covariates are drawn as either (a) $X_k = Z_k$ or (b) $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$ where $Z_k \sim U(-1, 1)$ for $k = 1, 2$. The link functions considered take the form $f(v) = \delta f^\star(v)$ for $f^\star \in \{v \mapsto c_1(1 + \exp(-v))^{-1}, v \mapsto c_2 \exp(-v^2), v \mapsto c_3 v^2\}$, $\delta \in (0, 1)$.[47] Each of these functions has a different shape; the scalars c_i $(i = 1, 2, 3)$ vary across the functions $f^\star$ and distributions for X and are chosen so that the variance of $f^\star(V_\theta)$ equals 4 under H_0: $\theta = 1$, whilst δ is taken the same for all functions and used to scale this variance.[48]

To examine the finite sample size of the proposed test, the efficient score function and efficient information matrix are estimated as in (1.30), with split-sample (penalised) smoothing cubic B-splines used to estimate each of $\hat{f}$, $\widehat{f'}$ and $\hat{Z}$.[49] I truncate the efficient information matrix at machine precision. Additionally I consider a Wald statistic estimated using an Ichimura (1993) style estimator, which uses the same estimates of $\hat{f}$, $\widehat{f'}$ and $\hat{Z}$ as the efficient score statistic.[50] The finite sample empirical rejection frequencies are reported in tables D.1 - D.4. In all specifications considered the efficient score provides good size control, whereas the Wald statistic based on the Ichimura (1993) type estimator described above displays substantial over-rejection, particularly for small δ.

To analyse the finite sample power of the efficient score test I consider the finite sample rejection frequency of the efficient score test of $\theta = 1$ for a grid of values around θ. Specifically, I take 21 equally spaced values between 0.875 and 1.125 and all other parameters are the same as for the simulations used to investigate finite sample size. Figures D.5 - D.8 plot the finite sample power function of the efficient score test, which demonstrate that – as expected – higher δ leads to higher power for the same distance from the null.

Heteroskedastic case

I now consider the heteroskedastic case. I consider two specifications for the error term: (1) $\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$ and (2) $\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_2)^2))$ where the constants s_i $(i = 1, 2)$ are chosen such that in each case $\mathbb{V}(\epsilon) = 1$ (unconditionally) under $H_0 : \theta = 1$.[51] The distributions for the covariates and the link functions used are the same as in the homoskedastic case.

To examine the finite sample size of the proposed test, the pseudo-efficient score function and its variance matrix are estimated as in (1.32) with split-sample (penalised) smoothing cubic B-splines used to estimate each of $\hat{f}$, $\widehat{f'}$, $\hat{Z}_1$ and $\hat{Z}_2$.[52] As in the homoskedastic case

[47] The first of these is the standard Logistic CDF.

[48] The scaling constants c are calculated in closed form for the case (a) with $X = (Z_1, Z_2)$. In the correlated case (b), evaluation of the integrals becomes substantially more complex and so simulated values are used, based on 10,000,000 draws.

[49] In particular I use the `smooth.spline` function in R with its default knot choice and penalty settings.

[50] This approach estimates θ by minimising the criterion $\theta \mapsto \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,i}(V_\theta))^2$; the estimates of $\widehat{f'}$ and $\hat{Z}$ are necessary to construct the asymptotic variance.

[51] These are determined by simulation with 10,000,000 draws.

[52] See footnote 49.

I truncate the variance matrix at machine precision. Additionally I consider a Wald statistic estimated using an Ichimura (1993) style estimator, which uses the same nonparametric estimates as the psuedo-efficient score statistic.[53]

The finite sample rejection frequencies with $\breve{\omega}(X)$ is taken as the infeasible truth $\omega(X)$ are reported in tables D.5 - D.8, whilst tables D.9 - D.12 report the finite sample size where $\breve{\omega}(X) = 1$. The results demonstrate qualitatively the same conclusions as the homoskedastic case, with the pseudo efficient score statistic always providing good size control, unlike the Wald statistic, which displays large over-rejection, particularly for small δ.

As in the homoskedastic case, to analyse the finite sample power of the pseudo efficient score test I consider the finite sample rejection frequency of the efficient score test of $\theta = 1$ for a grid of values around θ. As in the homoskedastic case, I consider 21 equally spaced values between 0.875 and 1.125 with all other parameters the same as for the simulations used to investigate finite sample size. Figures D.9 - D.12 plot the finite sample power curves. Similar observations apply as in the homoskedastic case, with higher δ leading to higher power for a given distance from the null. Moreover, as expected, the optimal (but infeasible) weighting scheme delivers higher power, though the difference seems to be relatively small for the designs considered.

1.5. Linear simultaneous equations models

In this section, I work out the details of the application of the theory developed in section 1.3 to a class of linear simultaneous equations models (LSEMs) where identification is based on an assumption of mutually independent and non-Gaussian errors. Under this assumption, no external information (e.g. instrumental variables) is required in order to identify the parameter of interest.

Consider the following linear simultaneous equations model (LSEM)

$$Y = RX + V, \quad V = A(\theta, \sigma)^{-1}\epsilon, \quad \mathbb{E}\epsilon = 0, \mathbb{V}\epsilon = I, \tag{1.34}$$

where the K components of ϵ are mutually independent, $X = (1, \tilde{X}')'$ is a vector of covariates independent of ϵ. R is a $K \times L$ matrix of regression coefficients and $A(\theta, \sigma)$ is a $K \times K$ invertible matrix. For later convenience I collect the Euclidean nuisance parameters R and σ into one vector: $\beta := (\beta_1', \beta_2')' := (\sigma', \operatorname{vec}(R)')'$.

As is well known, in simultaneous equations models of this form the elements of the mixing matrix, $A(\theta, \sigma)$, are not identified without further restrictions. However, if no more than one

[53] This approach estimates θ by minimising the criterion $\theta \mapsto \frac{1}{n}\sum_{i=1}^{n} \breve{\omega}(X_i)(Y_i - \hat{f}_{n,i}(V_\theta))^2$; the estimates of $\widehat{f'}$, $\hat{Z}_1$, $\hat{Z}_2$ are necessary to construct the asymptotic variance.

component of ϵ is Gaussian, the elements of the matrix $A(\theta, \sigma)$ are identified up to column permutation and sign changes (Comon, 1994). Imposition of sign restrictions and labelling of the shocks can then yield identification of the elements of $A(\theta, \sigma)$ which – assuming an identifiable parametrisation – yields that of θ.

Nevetheless, the identifying assumption that no more than one component of ϵ is Gaussian is not innocuous. In particular, depending on the parametrisation of the model, if this assumption fails, θ may be underidentified or completely unidentified. Moreover, as is typical in models with points of identification failure, the impact of the potential identification problem here is not binary. "Weak non-Gaussianity", where the error distribution is sufficiently close to Gaussianity relative to sampling uncertainty, can cause problems for inference methods which assume non-Gaussianity to obtain identification.[54] In this section I extend the analysis of Lee and Mesters (2022a) to demonstrate that inference based on the efficient score test is (i) robust to weak identification (in addition to underidentification and complete unidentification) and (ii) minimax optimal if θ is identified or underidentified.[55]

The first step of the analysis is to formally set up the model under consideration. Let η_0 denote the (Lebesgue) density of $\tilde{X}$ and for each $k = 1, \ldots, K$ let η_k be the (Lebesgue) density of ϵ_k and define ϕ_k as the log-density scores, i.e. $\phi_k(e) := \frac{\mathrm{d} \log \eta_k(s)}{\mathrm{d}s}(e)$. I will require a number of moments of (functions of) ϵ and $\tilde{X}$ to satisfy certain conditions.[56] In particular, for each $k \in [K]$ and some $\delta > 0$

$$\mathbb{E}\epsilon_k = 0, \ \mathbb{E}\epsilon_k^2 = 1, \ \mathbb{E}|\epsilon_k|^{4+\delta} < \infty, \ \mathbb{E}|\phi_k(\epsilon_k)|^{4+\delta} < \infty, \ \mathbb{E}\epsilon_k^4 - 1 > (\mathbb{E}\epsilon_k^3)^2, \tag{1.35}$$

and

$$\mathbb{E}\tilde{X}\tilde{X}' \succ 0, \quad \mathbb{E}\|\tilde{X}\|_2^{4+\delta} < \infty. \tag{1.36}$$

These moment restrictions are used to characterise the DGPs permitted by the model. Specifically, the density functions η_k and η_0 are assumed to belong (respectively) to the sets $\mathscr{G}$ and $\mathscr{Z}$ which are defined as follows:

$$\mathscr{G} := \left\{ g \in L_1(\mathbb{R}) : g \geq 0, \int g \, \mathrm{d}\lambda = 1, \sqrt{g} \in C^1(\mathbb{R}), \text{ if } \epsilon_k \sim g \text{ then (1.35)} \right\}, \tag{1.37}$$

$$\mathscr{Z} := \left\{ g \in L_1(\mathbb{R}^{L-1}) : g \geq 0, \int g \, \mathrm{d}\lambda^{L-1} = 1, \text{ if } \tilde{X} \sim g \text{ then (1.36)} \right\}, \tag{1.38}$$

where $L_1(\mathbb{R}^d)$ denotes the space of integrable functions on $\mathbb{R}^d$ with respect to the Lebesgue measure (which is denoted by λ^d or λ if the dimension is clear from context) and $C^1(\mathbb{R})$

[54] See Lee and Mesters (2022a) for simulation evidence of this phenomenon.

[55] Lee and Mesters (2022a) provide simulation evidence of a weak identification problem in this class of models, but their theoretical work only considers robustness against fixed distributions under which θ may be identified, underidentified or unidentified and does not cover weak identification.

[56] These conditions are the same as imposed in Lee and Mesters (2022a). Additionally I note that such fourth-moment conditions are common for conducting inference on variance parameters (e.g. White, 1980).

denotes the space of functions $\mathbb{R} \to \mathbb{R}$ which are continuously differentiable λ-a.e.. Finally the parameter $\beta = (\sigma', \mathrm{vec}(R)')'$ is assumed to belong to $\mathscr{B} \subset \mathbb{R}^{d_\beta}$. I will consider two restrictions on $\mathscr{B}$. Firstly it will be permitted to be an (otherwise unrestricted) open set. Alternatively – to explicitly handle the case of sign restrictions (or non-negativity restrictions on variances) – it will be permitted to have the form

$$\mathscr{B} = \mathscr{B}_1 \times \mathscr{B}_2, \qquad \mathscr{B}_1 = \prod_{l=1}^{d_\sigma} \mathscr{B}_{1,l}, \tag{1.39}$$

where $\mathscr{B}_2 \subset \mathbb{R}^{KL}$ is open and each $\mathscr{B}_{1,l} \subset \mathbb{R}$ is either open or one of $(-\infty, 0]$ or $[0, \infty)$.

The assumptions imposed on the LSEM model (1.34) are summarised as follows:

Assumption LSEM. *$W = (Y, \tilde{X})$ satisfies* (1.34) *where the K components of ϵ have marginal densities η_k ($k \in [K]$). Let the density of $\tilde{X}$ be η_0.*[57]

1. *$\Theta \subset \mathbb{R}^{d_\theta}$ is an open set and $\mathscr{B} \subset \mathbb{R}^{d_\beta}$ is either open or has the form $\mathscr{B}_1 \times \mathscr{B}_2$ where these factors are as described following* (1.39)*.*

2. *The components of ϵ are mutually independent and ϵ is independent of X.*

3. *$\eta_k \in \mathscr{G}$ for each $k \in [K]$ and $\eta_0 \in \mathscr{Z}$, for $\mathscr{G}$ and $\mathscr{Z}$ defined in* (1.37) *and* (1.38) *respectively.*

4. *The function $(\theta, \sigma) \mapsto A(\theta, \sigma)$ is continuously differentiable with l-th partial derivative $D_{1,l}(\theta, \sigma)$ and the functions $(\theta, \sigma) \mapsto D_{1,l}(\theta, \sigma) A(\theta, \sigma)^{-1}$ are Lipschitz continuous.*

The model is given by $\mathcal{P} = \{P_\gamma : \gamma \in \Gamma = \Theta \times \mathcal{H}\}$ with $\mathcal{H} := \mathscr{B} \times \mathscr{Z} \times \prod_{k=1}^{K} \mathscr{G}$ and where each P_γ has (Lebesgue) density

$$p_\gamma(W) = |\det(A(\theta, \sigma))| \prod_{k=1}^{K} \eta_k(A_k[Y - RX]) \times \eta_0(\tilde{X}). \tag{1.40}$$

◇

The moment and smoothness conditions imposed by part 3 of assumption LSEM are reasonably weak, as are the smoothness conditions in 4. The independence in 2 is, however, restrictive. Mutual independence of the components of ϵ is a testable assumption in applications (Matteson and Tsay, 2017; Amengual et al., 2021). The independence of $\tilde{X}$ and ϵ could be replaced by a conditional moment restriction, for which the general approach outlined in this paper would continue to hold, but the analysis below would need to be redone under this alternative assumption, with the efficient score function taking a different form.

[57] Each η_k is a density with respect to Lebesgue measure on the appropriate Euclidean space.

1.5.1. Verification of the modelling assumptions

Assumption LSEM coupled with the assumption that the observed data comprises an i.i.d. sample $(W_i)_{i=1}^n$ ensures that assumption M holds. I next show that assumption DQM holds, which is sufficient to imply assumption LAN by proposition 1.3.10.

For any $l \in [d_\theta + d_\sigma]$ and any $(k, j) \in [K]^2$, let $\zeta_{l,k,j} := [D_{1,l}(\theta, \sigma)]_k [A^{-1}]'_j$. Additionally write $D_{2,l}$ for the derivative of R with respect to the l-th component of $\beta_2 = \operatorname{vec}(R)$. $C_b^1(\mathbb{R})$ denotes the space of functions $\mathbb{R} \to \mathbb{R}$ which are bounded, continuously differentiable and have bounded derivatives λ-a.e. and $C_b(\mathbb{R}^L)$ denotes the space of functions $\mathbb{R}^L \to \mathbb{R}$ which are bounded and continuous λ^L-a.e.. Define the sets $\dot{\mathscr{G}}_{\eta,k}$ and $\dot{\mathscr{Z}}_\eta$ as:

$$\dot{\mathscr{G}}_{\eta,k} := \left\{ h_k \in C_b^1(\mathbb{R}) : \int h_k \, \mathrm{d}G_k = \int h_k \iota \, \mathrm{d}G_k = \int h_k \kappa \, \mathrm{d}G_k = 0 \right\}, \tag{1.41}$$

$$\dot{\mathscr{Z}}_\eta := \left\{ h_0 \in C_b(\mathbb{R}^{L-1}) : \int h_0 \, \mathrm{d}G_0 = 0 \right\} \tag{1.42}$$

where G_k is the measure on $\mathbb{R}$ corresponding to η_k ($k \in [K]$), G_0 the measure on $\mathbb{R}^{L-1}$ corresponding to η_0, ι denotes the identity function and $\kappa(e) := e^2 - 1$. Let

$$H_\eta := \prod_{l=1}^{d_\sigma} \mathscr{V}_l \times \mathbb{R}^{KL} \times \dot{\mathscr{Z}}_\eta \times \prod_{k=1}^{K} \dot{\mathscr{G}}_{\eta,k} \subset H := \mathbb{R}^{d_\beta} \times L_\infty(\lambda^{L-1}) \times \prod_{k=1}^{K} L_\infty(\lambda), \tag{1.43}$$

where each $\mathscr{V}_l = \mathbb{R}$ if β is an interior point of $\mathscr{B}$ and otherwise (i) $\mathscr{V}_l = [0, \infty)$ if $\mathscr{B}_{1,l} = [0, \infty)$ and $\sigma_l = 0$ or (ii) $\mathscr{V}_l = (-\infty, 0]$ if $\mathscr{B}_{1,l} = (-\infty, 0]$ and $\sigma_l = 0$. H is equipped with the norm $\|h\| := \|b\|_2 + \|h_0\|_{\lambda^{L-1},\infty} + \sum_{k=1}^K \|h_k\|_{\lambda,\infty}$, for $h = (b, h_0, \ldots, h_K) \in H$.[58] H_η is a linear subspace of H whenever β is an interior point of $\mathscr{B}$.

The sequences of base parameters considered are $\gamma_n = (\theta_n, \eta)$, with local perturbations of the form $\theta_n + \tau_n/\sqrt{n} \to \theta$ with $\tau_n \to \tau$ and

$$\eta_n(h_n) := (\beta_1 + t_n b_{1,n}, \beta_2 + t_n b_{2,n}, \eta_0(1 + t_n h_{n,0}), \eta_1(1 + t_n h_{n,1}), \ldots, \eta_K(1 + t_n h_{n,K})) \tag{1.44}$$

with $h_n \to h$ (all in H_η); note that $\eta_n(h_n) \to \eta$.

The following proposition establishes the quadratic mean differentiability of the model and hence LAN in view of Proposition 1.3.10.

Proposition 1.5.1. *Suppose that assumption LSEM holds, $\theta_n \to \theta \in \Theta$ and $\eta \in \mathcal{H}$ and consider the sequence defined by $\gamma_n = (\theta_n, \eta) \in \Gamma$. Let $\delta_n = I/\sqrt{n}$, $t_n := n^{-1/2}$, $\tau_n \to \tau$, $h_n := (b_n, h_{n,0}, h_{n,1}, \ldots, h_{n,K})$ (with $b_n = (b'_{1,n}, b'_{2,n})'$), with $h_n \to h$, and define $\eta_n : H_\eta \to \mathcal{H}$ as in* (1.44)*. Then assumption DQM holds, with $g_n := \tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ where*

[58] Each of the factors defining H is a Banach space (with the corresponding norm as just indicated) and hence the same is true of H when equipped with the indicated norm.

for $l = 1, \ldots, d_\theta$,

$$\dot{\ell}_{\gamma_n,l}(W) := \sum_{k=1}^{K} \left[\zeta_{l,k,k,n}(\phi_k(A_{n,k}V)A_{n,k}V + 1) + \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j,n}\phi_k(A_{n,k}V)A_{n,j}V \right],$$

$$\begin{aligned}[B_{\gamma_n}h](W) := &\sum_{m=d_\theta+1}^{d_\theta+d_{b_1}} b_{1,m} \sum_{k=1}^{K} \left[\zeta_{m,k,k,n}(\phi_k(A_{n,k}V)A_{n,k}V + 1) + \sum_{j=1, j\neq k}^{K} \zeta_{m,k,j,n}\phi_k(A_{n,k}V)A_{n,j}V \right] \\ &+ \sum_{k=1}^{K} \phi_k(A_{n,k}V) \left[-A_{n,k} \sum_{l=1}^{d_\beta} b_{2,l} D_{2,l} X \right] + h_0(\tilde{X}) + \sum_{k=1}^{K} h_k(A_{n,k}V),\end{aligned}$$

with $A_n := A(\theta_n, \sigma)$, $V := Y - RX$.

In order to simplify the expression of the the efficient score function, I suppose the following moment conditions on ϕ_k hold.

$$\mathbb{E}\phi_k(\epsilon_k) = 0,\ \mathbb{E}\phi_k(\epsilon_k)\epsilon_k = -1,\ \mathbb{E}\phi_k(\epsilon_k)\epsilon_k^2 = 0,\ \mathbb{E}\phi_k(\epsilon_k)\epsilon_k^3 = -3. \tag{1.45}$$

These moment conditions are weak; if (1.35) holds then a sufficient condition for (1.45) to hold is that the tails of the densities satisfy $\eta_k(x) = o(x^{-3})$.[59]

Proposition 1.5.2. *Suppose that assumption LSEM and equation* (1.45) *hold and consider the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of Proposition 1.5.1. Then assumption CM(ii) holds and (provided the inverse in the subsequent display exists) the efficient score function, $\tilde{\ell}_{\gamma_n}$, is given by*

$$\tilde{\ell}_{\gamma_n} = \tilde{\ell}_{\gamma_n,1} - \left[P_{\gamma_n}\tilde{\ell}_{\gamma_n,1}\tilde{\ell}_{\gamma_n,2}'\right] \left[P_{\gamma_n}\tilde{\ell}_{\gamma_n,2}\tilde{\ell}_{\gamma_n,2}'\right]^{-1} \tilde{\ell}_{\gamma_n,2}, \tag{1.46}$$

where for $l = 1, \ldots, d_\theta$, $m = 1, \ldots, d_{b_1}$, $s = 1, \ldots, d_{b_2}$, $v := V - RX$ and $\mu := \mathbb{E}X$,

$$\begin{aligned}\tilde{\ell}_{\gamma_n,1,l}(W) &= \sum_{k=1}^{K} \left[\zeta_{l,k,k,n}\left(\tau_{k,1}A_{n,k}V + \tau_{k,2}\kappa(A_{n,k}V)\right) + \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j,n}\phi_k(A_{n,k}V)A_{n,j}V \right] \\ \tilde{\ell}_{\gamma_n,2,m}(W) &= \sum_{k=1}^{K} \left[\zeta_{m,k,k,n}\left(\tau_{k,1}A_{n,k}V + \tau_{k,2}\kappa(A_{n,k}V)\right) + \sum_{j=1, j\neq k}^{K} \zeta_{m,k,j,n}\phi_k(A_{n,k}V)A_{n,j}V \right] \\ \tilde{\ell}_{\gamma_n,2,d_{b_1}+s}(w) &= \sum_{k=1}^{K} [-A_{n,k}D_{2,s}] \left[(x-\mu)\phi_k(A_{n,k}V) - \mu\left(\varsigma_{k,1}A_{n,k}V + \varsigma_{k,2}\kappa(A_{n,k}V)\right)\right],\end{aligned}$$

and

$$\tau_k := M_k^{-1}\begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \varsigma_k := M_k^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{with } M_k := \begin{pmatrix} 1 & P_{\gamma_n}(A_{n,k}V)^3 \\ P_{\gamma_n}(A_{n,k}V)^3 & P_{\gamma_n}(A_{n,k}V)^4 - 1 \end{pmatrix}.$$

[59] See Lemma S8 in Lee and Mesters (2022b). Alternatively, these conditions will hold provided differentiation and integration can be appropriately interchanged.

The preceding proposition requires the inverse of the variance matrix of $\tilde{\ell}_{\gamma_n,2}$ to exist. This is only necessary for the projection to be expressed in this precise form; if the matrix in question is singular, one can drop linearly dependent (in $L_2(P_{\gamma_n})$) elements from $\tilde{\ell}_{\gamma_n,2}$ until it is nonsingular. Additionally note that M_k is not indexed by n; under P_{γ_n}, $A_{n,k}V \sim \eta_k$ and so the moments making up M_k are constant in n.

1.5.2. Implementation of the efficient score test

Next I impose conditions which are sufficient for the construction of estimates of the efficient score function and efficient information matrix which satisfy assumptions E and R. First, I suppose that there is an appropriate estimator of each log density score ϕ_k available.

Assumption DSE. *Suppose that $(\beta_n)_{n\in\mathbb{N}} \subset \mathscr{B}$ is a deterministic sequence with $\sqrt{n}(\beta_n - \beta) = O(1)$. Let $\gamma_n' := (\theta_n, \beta_n, \eta)$, $A_n := A(\theta_n, \beta_{1,n})$ and $V_{n,i} := Y_i - R_n X_i$. The array of estimates $(\hat{\phi}_{n,k}(A_{n,k}V_{n,i}))_{n\in\mathbb{N}, i\leq n}$ satisfies*

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{n,i}) - \phi_k(A_{n,k}V_{n,i})\right]U_{n,i} &= o_{P_{\gamma_n'}}(n^{-1/2}) \\ \frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_{n,k}(A_{n,k}V_{n,i}) - \phi_k(A_{n,k}V_{n,i})\right]U_{n,i}\right)^2 &= o_{P_{\gamma_n'}}(\nu_n^2), \end{aligned} \tag{1.47}$$

for any $(U_{n,i})_{n\in\mathbb{N}, i\leq n}$ such that for each $n \in \mathbb{N}$, under $P_{\gamma_n'}$, the $U_{n,i} \in L_2^0(P_{\gamma_n'})$, are i.i.d. with marginal distribution G_u and are independent of each $A_{n,k}V_{n,j}$, and where $0 \leq \nu_n \to 0$ satisfies $\upnu_n = o(\nu_n)$ with

$$\upnu_n := \begin{cases} n^{-1/2}\log(n)^{1/2+\rho} & \textit{if } \delta \geq 4 \\ n^{(1-p)/(p)} & \textit{otherwise} \end{cases}, \tag{1.48}$$

for $p := \min\{1 + \delta/4, 2\}$ and some $\rho > 0$. ◇

Lee and Mesters (2022a, Appendix B) propose an appropriate estimator of ϕ_k using cubic B-splines – based on the density score estimator of Chen and Bickel (2006) – and demonstrate that it satisfies assumption DSE under assumption LSEM and some mild additional restrictions on η.

Given such an estimator, $\hat{\phi}_{n,k}$, of each ϕ_k and a $\xi_n := (\theta_n, \beta_n)$, the efficient score functions in Proposition 1.5.2 can be estimated by replacing each $\phi_k(A_k v)$ with $\hat{\phi}_{n,k}(A_{n,k}V_{n,k})$ and replacing each τ_k, ς_k and μ by their sample counterparts:

$$\begin{aligned}
\hat{\ell}_{\xi_n,1,l}(W_i) &:= \sum_{k=1}^{K}\left[\zeta_{l,k,k,n}\left(\hat{\tau}_{n,k,1}e_{n,k,i} + \hat{\tau}_{n,k,2}\kappa(e_{n,k,i})\right) + \sum_{j=1,j\neq k}^{K}\zeta_{l,k,j,n}\hat{\phi}_{n,k}(e_{n,k,i})e_{n,j,i}\right]\\
\hat{\ell}_{\xi_n,2,m}(W_i) &:= \sum_{k=1}^{K}\left[\zeta_{m,k,k,n}\left(\hat{\tau}_{n,k,1}e_{n,k,i} + \hat{\tau}_{n,k,2}\kappa(e_{n,k,i})\right) + \sum_{j=1,j\neq k}^{K}\zeta_{m,k,j,n}\hat{\phi}_{n,k}(e_{n,k,i})e_{n,j,i}\right]\\
\hat{\ell}_{\xi_n,2,d_{b_1}+s}(W_i) &:= \sum_{k=1}^{K}[-A_{n,k}D_{2,s}]\left[(X_i - \bar{X}_n)\hat{\phi}_{n,k}(e_{n,k,i}) - \bar{X}_n\left(\hat{\varsigma}_{n,k,1}e_{n,k,i} + \hat{\varsigma}_{n,k,2}\kappa(e_{n,k,i})\right)\right],
\end{aligned} \tag{1.49}$$

where $e_{n,k,i} := A_{n,k}V_{n,i}$, $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ and

$$\hat{\tau}_{n,k} := \hat{M}_{n,k}^{-1}\begin{pmatrix}0\\-2\end{pmatrix},\ \hat{\varsigma}_{n,k} := \hat{M}_{n,k}^{-1}\begin{pmatrix}1\\0\end{pmatrix},\ \text{ with } \hat{M}_{n,k} := \begin{pmatrix}1 & \frac{1}{n}\sum_{i=1}^{n}e_{n,k,i}^{3}\\ \frac{1}{n}\sum_{i=1}^{n}e_{n,k,i}^{3} & \frac{1}{n}\sum_{i=1}^{n}e_{n,k,i}^{4}-1\end{pmatrix}.$$

In practice, β is unknown but estimates can be formed using a discretised version of an estimator for β which is $\sqrt{n}$-consistent under P_{γ_n}. In model (1.34), $\beta_2 = \operatorname{vec}(R)$ can be estimated by OLS. Appropriate estimators of $\sigma = \beta_1$ depend on the parametrisation of the matrix $A(\theta,\sigma)$ but can usually be constructed from the sample analogue of the equality $\mathbb{E}(VV') = A(\theta,\sigma)^{-1}(A(\theta,\sigma)^{-1})'$ for a given θ and estimate of R.[60]

Suppose $\hat{\beta}_n$ is a $\sqrt{n}$-consistent estimate of β and let $\bar{\beta}_n$ be the estimate which replaces $\hat{\beta}_n$ by the closest value in $n^{-1/2}C\mathbb{Z}^{d_\beta}\cap\mathscr{B}$.[61] Let $\bar{\xi}_n := (\theta_n, \bar{\beta}_n)$ and define the estimates

$$\begin{aligned}
\hat{\ell}_{n,\theta_n} &:= \hat{\ell}_{\bar{\xi}_n,1} - \left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,1}\hat{\ell}_{\bar{\xi}_n,2}'\right]\left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,2}\hat{\ell}_{\bar{\xi}_n,2}'\right]^{-1}\hat{\ell}_{\bar{\xi}_n,2}\\
\check{\mathcal{I}}_{n,\theta_n} &:= \mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,1}\hat{\ell}_{\bar{\xi}_n,1}' - \left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,1}\hat{\ell}_{\bar{\xi}_n,2}'\right]\left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,2}\hat{\ell}_{\bar{\xi}_n,2}'\right]^{-1}\left[\mathbb{P}_n\hat{\ell}_{\bar{\xi}_n,2}\hat{\ell}_{\bar{\xi}_n,1}'\right],
\end{aligned} \tag{1.50}$$

and let $\hat{\mathcal{I}}_{n,\theta_n}$ be the eigendecomposition-truncated version of $\check{\mathcal{I}}_{n,\theta_n}$ at ν_n analogously to (1.21) (with ν_n as in assumption DSE).

1.5.3. Asymptotic properties

The following proposition demonstrates that the estimation procedure outlined in the previous subsection satisfies the conditions required for the theory in section 1.3 to apply.

Proposition 1.5.3. *Suppose that assumptions LSEM, DSE and equation* (1.45) *hold and that the observations form an i.i.d. sample. Consider the sequence* $(\gamma_n)_{n\in\mathbb{N}}$ *of Proposition 1.5.1. Suppose the inverse in* (1.46) *exists,* $\theta \mapsto \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ *is locally constant at* γ, $\hat{\beta}_n$ *is a* $\sqrt{n}$*-consistent estimate for* β *under* P_{γ_n} *and* $\hat{\ell}_{n,\theta_n}$, $\hat{\mathcal{I}}_{n,\theta_n}$ *are as in equation* (1.50)*. Then*

[60] Such initial estimators can often be refined by one step updates, see e.g. §25.8 in van der Vaart (1998).

[61] For an abritrary constant $C > 0$.

assumptions M, LAN, CM(ii), E and R hold.[62]

The preceding proposition requires the rank of $\tilde{\mathcal{I}}_\gamma$ to be locally constant in θ at γ. This reflects the situation under study in which the identification status of θ is determined by η. Note that since the rank function is lower semi-continuous and non-negative integer valued, there is always a small enough neighbourhood on which the rank is bounded below by $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$. Therefore the force of the restriction is only that on some neighbourhood the rank cannot strictly exceed $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$, which is evidently the case for full rank $\tilde{\mathcal{I}}_\gamma$. For rank deficient $\tilde{\mathcal{I}}_\gamma$, the assumption has force.[63]

Given the definition of the efficient score and efficient information matrix estimators in (1.50) and supposing the hypothesis of interest is two-sided, the efficient score statistic and test can be defined as in equations (1.9) and (1.10). Since the required conditions have been established above, the results on size and power of the efficient score test – as established in section 1.3 – apply directly.

Corollary 1.5.4. *In the setting of proposition 1.5.3, let H'_η be a compact subset of H_η. Then the efficient score test satisfies*

$$\limsup_{n\to\infty} \sup_{h\in H'_\eta} P^n_{\gamma_n,0,h}\phi_{n,\theta_0} \leq \alpha,$$

and, for any compact $\Theta' \subset \Theta$, the corresponding test inversion confidence sets satisfy

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta'} \inf_{h\in H'_\eta} P^n_{(\theta,\eta),0,h}(\theta \in \hat{C}_n) \geq 1-\alpha.$$

Corollary 1.5.4 is the key results as regards robust inference in the presence of possible weak under- or un-identification of θ, as may occur when the components of η are sufficiently close to Gaussianity relative to the sample size. The results demonstrate that the efficient score has correct asymptotic size uniformly over local perturbations of the nuisance parameters and the corresponding (test inversion) confidence sets are uniformly valid over compact subsets of Θ and local perturbations of the nuisance parameters.

As the perturbation sets H_η are linear spaces whenever $\beta \in \operatorname{int}\mathscr{B}$, if this condition holds the efficient score test has optimality properties in the fully- and under- identified cases

Corollary 1.5.5. *In the setting of proposition 1.5.3 suppose additionally that β is an interior point of $\mathscr{B}$ and $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$. If $d_\theta = 1$, then the efficient score test is locally asymptotically uniformly most powerful unbiased. If $d_\theta > 1$, then the efficient score test is locally asymptotically maximin.*

I next examine the finite sample performance of the efficient score test in two explicit

[62] Where the scores and paths in assumption LAN are as in proposition 1.5.1.

[63] From this discussion it is evident that an alternative way of stating this restriction would be that $\theta \mapsto \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ is upper semi-continuous (or continuous) at γ.

versions of the LSEM via two simulation studies. In the first study I consider a scalar parameter and focus on potential weak identification as may occur under error distributions close to Gaussianity. In the second I consider a two dimensional parameter which is underidentified under Gaussianity.

1.5.4. Simulation study (i)

Consider model (1.34), with $K = 2$, $L = 2$ and let the mixing matrix $A(\theta, \sigma)$ be

$$A(\theta, \sigma) = \begin{bmatrix} \sigma_2^{-1} & 0 \\ 0 & \sigma_3^{-1} \end{bmatrix} \begin{bmatrix} 1 & -\theta \\ -\sigma_1 & 1 \end{bmatrix}.$$

The null hypothesis under consideration is that $H_0 : \theta = 0$. When both ϵ_1 and ϵ_2 are close to Gaussianity, θ in this model will be only weakly identified.

To shed light on the finite sample performance of the efficient score test, I draw 5000 samples from this model for a range of different sample sizes and distributions for the error components ϵ_1 and ϵ_2. The $\tilde{X}$ variables are drawn as independent standard normals and $\beta_1 = \sigma = (0.7, 1.0, 3.0)$, $\beta_2 = \text{vec}(R) = (1, 2, -1, -3/2)'$. Table D.13 tabulates the considered error distributions for ϵ_1 and ϵ_2. 3 different distributions are considered for ϵ_1 and 10 for ϵ_2.[64] In particular, I consider a fixed distribution for ϵ_1 and examine the finite sample behaviour of the efficient score test as the distribution of ϵ_2 approaches Gaussianity, starting from 3 non-Gaussian distributions, each with a different shape.

To implement the efficient score test, I estimate each ϕ_k using the B-spline based estimator described in Appendix B of Lee and Mesters (2022a), which is adapted from a similar estimator proposed by Chen and Bickel (2006).[65] The remaining (Euclidean) nuisance parameters are estimated in two ways: (i) $\beta_2 = \text{vec}(R)$ is estimated by OLS, with an estimate of β_1 recovered from the empirical variance matrix of the residuals $Y_i - \hat{R}X_i$. (ii) These OLS-based estimates are used to estimate the efficient score function for β, and then a one-step update is made based on this preliminary efficient score.[66]

With all the required nuisance parameters estimated, the efficient score function is constructed as in equation (1.50), the efficient score statistic is conducted as in equation (1.9) and the test performed as in equation (1.10) at a nominal level of 5%.[67]

The empirical rejection frequencies for the efficient score test conducted with (i) OLS-based estimates of the Euclidean nuisance parameters and (ii) one-step updates of these estimates

[64] The density functions of these distributions are plotted in figures D.1 - D.3.

[65] In each simulation design, I use 6 cubic B-splines and set the upper and lower knots to be the 95th and 5th percentile of the samples, respectively adjusted up and down by $\log(\log n)$, truncated at the maximum (respectively minimum) sample value.

[66] I note that in the construction of the test θ is fixed throughout and so considered known.

[67] The information matrix eigenvalues are truncated at machine precision.

are recorded in tables D.14 - D.16; each table corresponds to a different distribution for ϵ_1. The table of primary interest is table D.14, with $\epsilon_1 \sim \mathcal{N}(0,1)$ as this corresponds to a potentially weakly identified setting. As this table demonstrates, the efficient score test appears to demonstrate valid size control for all sample sizes and choices of η_2 considered. The version of the efficient score test with one-step updates provides reasonable size control, though demonstrates slight over-rejection in a number of cases. This finding holds also in each tables D.15 - D.16.

Tables D.14 – D.16 also contain size results for a number of alternative testing approaches. Two are Wald and LM tests based on a pseudo-maximum likelihood approach, inspired by the approach in Gouriéroux et al. (2017).[68] Here, a density is chosen for each of the error components and standard psuedo-maximum likelihood tests are performed. Following Gouriéroux et al. (2017) I choose a (normalised) $t(5)$ distribution for both ϵ_1 and ϵ_2 in this simulation experiment. As might be expected, the Wald statistic does not control size at the nominal level, displaying varying degrees of over-rejection (depending on η_2) in table D.14. Its performance in the settings recorded in tables D.15 and D.16 is mixed, demonstrating an ability to control size when at least one psuedo-density is sufficiently close to the truth, and substantial over-rejection otherwise. In contrast, the LM statistic (which imposes the null value of θ) does correctly control size for each choice of η_2 in tables D.14 – D.16.

The final two tests are Wald and LM tests based on a GMM framework in which higher moments of the error terms are used to provide identifying information. The moments used were drawn from Lanne and Luoto (2021).[69] Specifically, the (nine) moment conditions utilised are:

$$\mathbb{E}[\epsilon_1 \tilde{X}] = \mathbb{E}[\epsilon_2 \tilde{X}] = \mathbb{E}[\kappa(\epsilon_1)] = \mathbb{E}[\kappa(\epsilon_2)] = \mathbb{E}[\epsilon_1 \epsilon_2] = \mathbb{E}[\epsilon_1^3 \epsilon_2] = \mathbb{E}[\epsilon_1^2 \epsilon_2^2 - 1] = 0.$$

Neither of these GMM based tests (based on these moments) achieve finite sample size close to nominal in the simulation experiments, as can be seen in tables D.14 – D.16. In the latter two tables, where weak identification is not present, the finite sample sizes of these tests appear to be reducing towards the nominal level as n increases, but remain substantially above the nominal level in each simulation design considered.

I perform a further simulation experiment based on this model to document the failure of size control of the score test based on the score functions for the Euclidean parameters

[68] Gouriéroux et al. (2017) consider a similar problem but in a SVAR setting.

[69] Like Gouriéroux et al. (2017), Lanne and Luoto (2021) consider a SVAR setting.

$(\theta', \beta_1', \beta_2')'$. The relevant scores take the form

$$\dot{\ell}_{\gamma,l}(W) := \sum_{k=1}^{K} \left[\zeta_{l,k,k}(\phi_k(A_k V)A_k + 1) + \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j}\phi_k(A_k V)A_j V \right]$$
$$\dot{\ell}_{\gamma,m}(W) := \sum_{k=1}^{K} [-A_k D_{b,l} X]\phi_k(A_k V),$$

for $l = 1, \ldots, d_\theta, d_\theta + 1, \ldots, d_\theta + d_{\beta_1}$ and $m = d_\theta + d_{\beta_1} + 1, \ldots, d_\theta + d_{\beta_1} + d_{\beta_2}$.[70] Let $\dot{\ell}_\gamma^1$ denote the first d_θ elements, and $\dot{\ell}_\gamma^2$ the remainder. Let $\dot{S}_{n,\theta}$ be the statistic formed analogously to (1.9) but based on an estimated version of $\dot{\ell}_\gamma^1 - \dot{I}_{12}\dot{I}_{22}^{-1}\dot{\ell}_\gamma^2$, with $\dot{I}_\gamma = P_\gamma \dot{\ell}_\gamma \dot{\ell}_\gamma'$, rather than $\tilde{\ell}_\gamma$.

Since score functions have finite second moments,

$$\sqrt{n}\mathbb{P}_n \left[\dot{\ell}_\gamma^1 - \dot{I}_{12}\dot{I}_{22}^{-1}\dot{\ell}_\gamma^2 \right] \rightsquigarrow \mathcal{N}(0, \dot{I}_{\gamma,11} - \dot{I}_{\gamma,12}\dot{I}_{\gamma,22}^{-1}\dot{I}_{\gamma,21}),$$

and hence *if* $\dot{\ell}_\gamma$ and $\dot{I}_\gamma$ could be replaced by estimates with conditions analogous to those in assumption E and R holding, the test based on $\dot{S}_{n,\theta}$ would correctly control size.

Table D.17 demonstrates that this is not the case, with the efficient score based tests controlling size, whilst the analogous tests based on $\dot{\ell}_\gamma$ (with the same estimator of ϕ_k) do not.[71] The key problem here is the bias caused by the regularised estimation of ϕ_k which is present in the estimate of $\dot{\ell}_\gamma$. This bias is removed by the orthogonal projection onto the nuisance score space in the definition of $\tilde{\ell}_\gamma$.

Following the size results, I compared the power of the two efficient score tests to that of the psuedo-ML based LM test which also was able to correctly control size in all designs considered. Figures D.13 - D.15 plot the results, corresponding to $\epsilon_1 \sim \{\mathcal{N}(0,1),\ t'(5),\ \mathcal{SN}'(0,1,4)\}$ respectively where t' and $\mathcal{SN}'$ denote the standardised version of the indicated distribution.

These finite sample power curves show that the power provided by any of the tests considered declines as the density η_2 approaches Gaussianity, particularly in the potentially weakly identified case where $\epsilon_1 \sim \mathcal{N}(0,1)$ (figure D.13) in which available power appears low. In constrast, in figures D.14 and D.15 where there is no (weak) identification issue, the efficient score tests apear to provide good finite sample power, with the version based on one-step updated estimates providing slightly higher power. The pseudo-maximum likelihood LM test also provides good power in cases where the chosen pseudo-densities are close to the truth. In particular, it slightly exceeds the power of the efficient score tests when ϵ_2 has a (standardised) t distribution in figures D.13 and D.14. Nevertheless, the

[70] Cf. proposition 1.5.1.

[71] In this simulation design, ϵ_1 and ϵ_2 have the same distribution, and are at a fixed distance from Gaussianity to focus on the problem of plugging in an estimate of a non-parametric parameter, rather than potential identification problems.

efficient score test is competitive and provides close to identical power in the first row of figure D.14, despite the pseudo density matching the truth in the first panel. Moreover, in cases where the psuedo-density is far from the truth, the power of the efficient score test is substantially higher than that provided by the pseudo-ML LM test (see, in particular, the third row of figure D.14 and each row of figure D.15).

1.5.5. Simulation study (ii)

In this second simulation study I consider the power available in a LSEM where the structural parameter of interest is underidentified. Specifically suppose that the data satisfies (1.34) where for $\theta = (a, b)$ with $a \neq b$ and $\beta_1 = (\sigma_1, \sigma_2) \in (0, \infty)^2$,

$$A(\theta, \beta_1) = \begin{bmatrix} \sigma_1^{-1} & 0 \\ 0 & \sigma_2^{-1} \end{bmatrix} \begin{bmatrix} 1 & -a \\ 1 & -b \end{bmatrix},$$

and there is one, zero-mean, unit variance X variable with coefficients $R = 0$. By explicit calculation, the efficient information matrix in this model takes the form

$$\tilde{\mathcal{I}}_\gamma = \frac{1}{(a-b)^2} \begin{bmatrix} \mathbb{E}[\phi_1(\epsilon_1)^2]c & -1 \\ -1 & \mathbb{E}[\phi_2(\epsilon_2)^2]c^{-1} \end{bmatrix}, \quad c := \left({}^{\sigma_2}/_{\sigma_1}\right)^2. \tag{1.51}$$

I consider three distributions from which to draw each ϵ_k: (i) $\mathcal{N}(0, 1)$, (ii) $t'(5)$ - a (standardised) t distribution with 5 degrees of freedom and (iii) $st'(5, 2)$ a (standardised) skew t distribution constructed as in Fernandez and Steel (1998) with 5 degrees of freedom and skewness parameter 2.[72] These correspond to (i) $\mathbb{E}[\phi_k(\epsilon_k)^2] = 1$, (ii) $\mathbb{E}[\phi_k(\epsilon_k)^2] = 1.25$ and (iii) $\mathbb{E}[\phi_k(\epsilon_k)^2] \approx 2.54$ respectively.

In the standard normal case (i), $\tilde{\mathcal{I}}_\gamma$ has eigenvalues $\lambda_1 = (c + c^{-1})/(a - b)^2$, $\lambda_2 = 0$ and a one-dimensional hyperplane as its nullspace: $N(\tilde{\mathcal{I}}_\gamma) = \{x \in \mathbb{R}^2 : cx_1 = x_2\}$. In cases (ii) and (iii), the matrix is positive definite and so $N(\tilde{\mathcal{I}}_\gamma) = \{0\}$.

Consider testing $\theta = \theta_0 = (a, b) = (1/2, 1/4)$, where $\sigma_1 = \sigma_2 = 1$ and hence the nullspace is the line $x_1 = x_2$. I take $n \in \{600, 1000, 1400\}$ and draw simulation samples according to (1.34) with $\theta = \theta_0 + \tau/\sqrt{n}$ and $X \sim \mathcal{N}(0, 1)$. β_2 is estimated by OLS and β_1 by GMM using the three moment conditions implied by the relationship $\mathbb{E}[VV'] = A(\theta, \beta_1)^{-1}(A(\theta, \beta_1)^{-1})'$. These estimates are used to construct estimates of the efficient score function and information matrix as in (1.50). In each case I truncate at machine precision.

The finite sample and asymptotic power surfaces are plotted in figures D.16 - D.18. Figure D.16 demonstrates the expected trivial power along the hyperplane $N(\tilde{\mathcal{I}}_\gamma)$ in the Gaussian

[72] The density functions of these distributions are plotted in figure D.4.

case, with power otherwise increasing in $\|\tau\|$. In contrast, figures D.17 and D.18 depict the full rank case, in which trivial power is found only at the point $\tau = 0$.[73] In all three figures, comparison of the finite sample power surface to the asymptotic power surface in the bottom right suggests that the asymptotic power results provide a good approximation to finite sample power.

1.6. Discussion

In this paper I demonstrated that score-type statistics based on the efficient score function can be used to perform uniformly valid inference in a wide class of models. A high level framework was provided in order to develop the theoretical results, based on the local asymptotic normality (LAN) framework of Le Cam.

The version of this framework considered permits many models and scenarios in which standard testing procedures fail to correctly control size, as demonstrated via specific examples. This class includes models which may suffer from identification problems, models where nuisance parameters may lie on the boundary of the parameter space and models which need a regularisation step for their estimation. I demonstrated that the efficient score test enjoys locally uniformly valid size control. Moreover, I showed that a number of standard testing optimality results continue to hold in this setup and demonstrated a minimax optimality result which applies in cases where, for example, the parameter of interest is underidentified.

A number of examples were studied in detail to demonstrate the applicability of the suggested framework and how the conditions it requires may be shown to hold. Simulation studies based on these examples suggest that the asymptotic results obtained provide a useful guide to finite sample performance. The simulations also show that – in the cases considered – the procedures based on the efficient score statistic perform better than alternative procedures.

The treatment in the current paper is restricted to cases where the observed data forms a random sample. This restriction was made to remove inessential complications in the derivation of the results. With these now established in the baseline i.i.d. case, an interesting potential extension would be to extend these results to other sampling schemes. An additional drawback of the current treatment is that the parameter of interest θ is required to be a bona fide parameter of the model as opposed to a function of the model parameters. Such extensions are left for future work.

[73] Which, of course, is exactly the nullspace of $\tilde{\mathcal{I}}_\gamma$ in this case.

Acknowledgements

I am grateful for the support, advice and encouragement I received from my advisor Geert Mesters during the writing of this chapter. I have also benefitted from discussions with and comments from Christian Brownlees, Bjarni G. Einarsson, Kirill Evdokimov, Lukas Hoesch, Vladislav Morozov, Katerina Petrova, Barbara Rossi, André B. M. Souza, Philipp Tiozzo as well as participants at the 2021 Barcelona GSE PhD Jamboree, the 2021 Symposium of the Spanish Economics Association, the 2021 SNDE Workshop for Young Researchers, the 2022 Spring Meeting of Young Economists, the "Advances in Econometrics" workshop of the 2022 BSE summer forum and seminar participants at UPF (Statistics), University of Surrey, BI Norwegian Business School (Data Science & Analytics), Toulouse School of Economics, Duke University, Erasmus University Rotterdam (Econometric Institute), University of Liverpool and Western University.

Appendices

A. Notation & conventions

$A \coloneqq B$ means that A is defined to be B. $A \subset B$ indicates that A is a subset of B. All vector spaces are over the real field $\mathbb{R}$. Given a positive integer K, $[K] \coloneqq \{1, \dots, K\}$. For any Euclidean parameter, say α, d_α denotes the dimension of the space in which it lives. Similarly for a vector of functions κ, d_κ is the number of component functions. For a sequence $(x_n)_{n\in\mathbb{N}}$, $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ denotes that each $x_n \in \mathcal{X}$. For any matrix M, $\|M\|_2$ is its spectral norm and $M^\dagger$ is its Moore-Penrose pseudo-inverse. "$\succeq$" is used to denote the Loewner partial order; that is, given two Hermitian matrices A, B, $A \succeq B$ iff $A - B$ is positive semi-definite and $A \succ B$ iff $A - B$ is positive definite. If A is a linear operator, $N(A)$ is its nullspace. Given a topological space S, $\mathcal{B}(S)$ is its Borel σ-algebra. Weak convergence is denoted by "$\rightsquigarrow$". Operator notation is often used for integrals: $Pf \coloneqq \int f \, \mathrm{d}P$. $\mathbb{P}_n$ denotes the empirical measure of a given sample and $\mathbb{G}_n$ the empirical process. Throughout this paper & unless otherwise noted the sample considered is denoted by $(W_i)_{i=1}^n \in \mathcal{W}^n$, hence $\mathbb{P}_n f = \int f \, \mathrm{d}\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n f(W_i)$. For a sequence of functions $(f_n)_{n\in\mathbb{N}}$ with each f_n having domain $\mathcal{W}^n$ and a sequence of probability measures $(P_n)_{n\in\mathbb{N}}$ on $\mathcal{W}$, convergence statements will often be written as $f_n \rightsquigarrow f$ under P_n. This is shorthand for weak convergence under the product measures P_n^n. If X has distribution G, I write $X \sim G$. If g is the density of G (with respect to some σ-finite measure), I also write $X \sim g$. $X \simeq Y$ indicates that X and Y have the same distribution. $L_p(P)$ denotes the space of functions f such that $P|f|^p < \infty$. In the case where $f = (f_1, \dots, f_K)$ is a vector of functions $f \in L_p(P)$ denotes that each $f_i \in L_p(P)$ for $i = 1, \dots, K$. $L_p^0(P)$ is the subspace of $L_p(P)$ whose members f satisfy $Pf = 0$. Given a (closed) subspace S of a Hilbert space H, the orthogonal projection of a function $f \in H$ onto S is denoted by $\Pi(f|S)$.

B. Additional details and proofs of results in the main text

B.1. Details and proofs for section 1.3

Construction of the sequence of experiments

In order to discuss power I use the limits of experiments framework of Le Cam (see e.g. chapter 9 of van der Vaart (1998) for an introduction). Under the additional assumption that $\mathscr{H}_\gamma$ is a linear space, I will obtain a Gaussian shift limit experiment on a particular inner-product space.[74]

To state the proposition, I need to define the inner-product space that will be used to parametrise the experiments. Let $N(A)$ denote the null space of a linear transformation A; in particular $N(\tilde{\mathcal{I}}_\gamma)$ denotes the null space of the matrix $\tilde{\mathcal{I}}_\gamma$. For the nuisance perturbations, h, it is more convenient to parametrise directly by the scores $g = B_\gamma h$. For each $g = B_\gamma h \in \mathscr{H}_\gamma$ let $\mathfrak{h}_{g,\eta} := \{h \in H_\eta : B_\gamma h = g\}$. Suppose that $\mathscr{H}_\gamma$ is a linear subspace of $L_2(P_\gamma)$ and note that it is therefore a dense subspace of a its completion (which is a Hilbert space). It therefore has an orthonormal basis, $(g_k)_{k\in\mathbb{N}}$.[75] For each element g_k in this basis select (arbitrarily) an element $h_k = h_{g_k}$ from each $\mathfrak{h}_{g_k,\eta}$. For any other element $g \in \mathscr{H}_\gamma$ choose $h_g = \sum_{k\in\mathbb{N}} a_k h_k$ where $g = \sum_{k\in\mathbb{N}} a_k g_k$. Denote the collection of such h_g as $\mathfrak{H}_\gamma := \{h_g : g \in \mathscr{H}_\gamma\} \subset H_\eta$.[76] I will consider sequences of experiments, where each consists of measures of the form $P_{\gamma_n,\tau,g} = P_{\gamma_n,\tau,h}$ for $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ and $g \in \mathscr{H}_\gamma$, $h = h_g \in \mathfrak{H}_\gamma$ (with $\gamma = \lim_{n\to\infty}\gamma_n$); that is to say, these experiments are parametrised by the (inner-product) space $\mathbb{H}_\gamma := N(\tilde{\mathcal{I}}_\gamma)^\perp \times \mathscr{H}_\gamma$ equipped with the inner-product given below in (52).

The choice of a particular "representative" $h = h_g$ for each score $g = B_\gamma h \in \mathscr{H}_\gamma$ as in the preceding construction is a technical point which will not impede statements being made about the behaviour of tests along sequences with $h_n \to h \in H_\eta \setminus \mathfrak{H}_\gamma$ due to the following lemma.

Lemma B.1. *Suppose that assumptions M, LAN, CM(i) hold and that $(\psi_n)_{n\in\mathbb{N}}$ is a sequence of tests on $\mathcal{W}^n$ (i.e. each $\psi_n : \mathcal{W}^n \to [0,1]$).*

1. *If $(\tau_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{d_\theta}$ and $(h_n)_{n\in\mathbb{N}} \subset H_\eta$ are convergent sequences with limits $\tau \in \mathbb{R}^{d_\theta}$ and $h \in H_\eta$ respectively, then*

$$\limsup_{n\to\infty} \left[P^n_{\gamma_n,\tau_n,h_n}\psi_n - P^n_{\gamma_n,\tau,h}\psi_n\right] = 0.$$

2. *If $h_1, h_2 \in H_\eta$ are such that $B_\gamma h_1 = B_\gamma h_2$ and $h_1 - h_2 \in H_\eta$, then for any convergent sequences $(\tau_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{d_\theta}$, $(h_{1,n})_{n\in\mathbb{N}} \subset H_\eta$, $(h_{2,n})_{n\in\mathbb{N}} \subset H_\eta$ with limits $\tau \in \mathbb{R}^{d_\theta}$ and $h_1, h_2 \in H_\eta$ respectively,*

$$\Lambda_n(\gamma_n(\tau_n, h_{1,n}), \gamma_n) - \Lambda_n(\gamma_n(\tau_n, h_{2,n}), \gamma_n) = o_{P_{\gamma_n}}(1),$$

[74] That is, the limit experiment is the restriction of a Gaussian shift experiment on a specific Hilbert space to the inner-product space of interest. See e.g. Le Cam (1986, Chapter 9, section 3) or Strasser (1985, Chapter 11) for an introduction to Gaussian shift experiments on Hilbert spaces.

[75] See footnote 85.

[76] I will suppose that the $h_g = h_0$ chosen to correspond to $g = 0$ is $h_g = h_0 = 0$. Note that if B_γ is injective there is only one such h_g for each $g \in \mathscr{H}_\gamma$.

and

$$\limsup_{n\to\infty}\left[P^n_{\gamma_n,\tau_n,h_{1,n}}\psi_n - P^n_{\gamma_n,\tau_n,h_{2,n}}\psi_n\right] = 0.$$

With the setup previously described the following result concerning convergence of experiments can be stated. This result is straightforward given the assumptions made, and is quite standard, aside from potentially one key aspect: the definition of the indexing set of the sequence of experiments — that $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$. This ensures that the inner-product in equation (52) *is* an inner-product. If $N(\tilde{\mathcal{I}}_\gamma)^\perp$ was replaced by $\mathbb{R}^{d_\theta}$ and $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) < d_\theta$, the map in (52) would only be a positive-semidefinite Hermitian form.[77]

Proposition B.2. *Suppose that assumptions M, LAN and CM(i) hold and that $\mathscr{H}_\gamma$ is a linear subspace of $L_2(P_\gamma)$. Suppose that $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0$ and let $\mathbb{H}_\gamma := N(\tilde{\mathcal{I}}_\gamma)^\perp \times \mathscr{H}_\gamma$. If the map $\langle\cdot,\cdot\rangle_{\mathbb{H}_\gamma} : \mathbb{H}_\gamma \times \mathbb{H}_\gamma \to \mathbb{R}$ is defined by*

$$\langle(\tau_1, g_1), (\tau_2, g_2)\rangle := \langle \tau_1'\dot{\ell}_\gamma + g_1, \tau_2'\dot{\ell}_\gamma + g_2\rangle_{P_\gamma}, \tag{52}$$

then $(\mathbb{H}_\gamma, \langle\cdot,\cdot\rangle)$ is an inner-product space. In addition, the sequence of experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$, where each

$$\mathscr{E}_n := \left(\mathcal{W}^n, \mathcal{B}(\mathcal{W}^n), \left\{P^n_{\gamma_n,\tau,g} : (\tau, g) \in \mathbb{H}_\gamma\right\}\right), \tag{53}$$

converges weakly to a Gaussian shift on $(\mathbb{H}_\gamma, \langle\cdot,\cdot\rangle)$.

Proofs

Proof of proposition 1.3.1. To simplify the notation, let $g_n := \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h$ and $g := \tau'\dot{\ell}_\gamma + B_\gamma h$. Let $\{W_{n,k} : k \le n, n \in \mathbb{N}\}$ be a triangular array, where each row $W_{n,1}, \ldots, W_{n,n}$ ($n \in \mathbb{N}$) is independently and identically distributed, with each random vector $W_{n,k}$ having law P_{γ_n}. Let $\{Z_{n,k} : k \le n, n \in \mathbb{N}\}$ be the array defined by $Z_{n,k} := \left(\tilde{\ell}_{\gamma_n}(W_{n,k})', g_n(W_{n,k})\right)'$. The rows of this array are i.i.d. with $\mathbb{E}Z_{n,k} = 0$ and $\mathbb{V}Z_{n,k} = \begin{bmatrix}\tilde{\mathcal{I}}_{\gamma_n} & \tilde{\mathcal{I}}_{\gamma_n}\tau \\ \tau'\tilde{\mathcal{I}}_{\gamma_n} & P_{\gamma_n}g_n^2\end{bmatrix}$ (for each k, n). [78] By assumption CM(ii)

$$\frac{1}{n}\sum_{k=1}^n \mathbb{V}Z_{n,k} = \mathbb{V}Z_{n,1} \to \begin{bmatrix}\tilde{\mathcal{I}}_\gamma & \tilde{\mathcal{I}}_\gamma\tau \\ \tau'\tilde{\mathcal{I}}_\gamma & \sigma^2_{\tau,h}\end{bmatrix}, \tag{54}$$

[77] That is, $\langle(\tau,g),(\tau,g)\rangle = 0$ whilst $(\tau,g) \neq 0$ would be possible. In particular, $\langle(\tau,0),(\tau,0)\rangle = 0$ would hold for all $\tau \in N(\tilde{\mathcal{I}}_\gamma)$, which has positive dimension whenever $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) < d_\theta$.

[78] We have that $P_{\gamma_n}\tilde{\ell}_{\gamma_n}\tilde{\ell}_{\gamma_n}' = \tilde{\mathcal{I}}_{\gamma_n}$ (e.g. Rudin, 1991, Theorem 12.14). $P_{\gamma_n}\tilde{\ell}_{\gamma_n}[B_{\gamma_n}h] = 0$ by the construction of the efficient score function.

where

$$\sigma^2_{\tau,h} := P_\gamma g^2 = P_\gamma[\tau'\dot{\ell}_\gamma + B_\gamma h]^2 = \lim_{n\to\infty} P_{\gamma_n}\left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right]^2 = \lim_{n\to\infty} P_{\gamma_n} g_n^2, \tag{55}$$

and hence (97) is satisfied. Moreover assumptions LAN and CM(ii) together yield that $(\|Z_{n,1}\|_2^2)_{n\in\mathbb{N}}$ is uniformly integrable and hence as the rows are identically distributed, (98) holds. It then follows by lemma C.1 that under P_{γ_n} we have

$$\sqrt{n}\mathbb{P}_n\left(\tilde{\ell}'_{\gamma_n}, \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n} h\right)' \rightsquigarrow \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tilde{\mathcal{I}}_\gamma & \tilde{\mathcal{I}}_\gamma\tau\\ \tau'\tilde{\mathcal{I}}_\gamma & \sigma^2_{\tau,h}\end{pmatrix}\right). \tag{56}$$

Combining equations (1.5), (54), (55) and (56) we have

$$\left(\sqrt{n}\mathbb{P}_n\tilde{\ell}'_{\gamma_n},\ \Lambda_n(\gamma_n(\tau_n,h_n),\gamma_n)\right)' \rightsquigarrow \mathcal{N}\left(\begin{pmatrix}0\\ -\frac{1}{2}\sigma^2_{\tau,h}\end{pmatrix}, \begin{pmatrix}\tilde{\mathcal{I}}_\gamma & \tilde{\mathcal{I}}_\gamma\tau\\ \tau'\tilde{\mathcal{I}}_\gamma & \sigma^2_{\tau,h}\end{pmatrix}\right). \tag{57}$$

The marginal convergence of the likelihood ratio yields that $(P^n_{\gamma_n})_{n\in\mathbb{N}}$ and $(P^n_{\gamma_n,\tau_n,h_n})_{n\in\mathbb{N}}$ are mutually contiguous (e.g. van der Vaart and Wellner, 1996, Example 3.10.6). We remark here that a completely analogous argument to the foregoing applied to the array $\{g_n(W_{n,k}) : k \leq n, n \in \mathbb{N}\}$ yields this same marginal convergence under assumption CM(i) rather than assumption CM(ii) and hence the mutual contiguity of these sequences of measures continues to hold under this weaker condition, as claimed in the statement of the proposition.

By Le Cam's third lemma (e.g. van der Vaart and Wellner, 1996, Example 3.10.8) it follows from (57) that under P_{γ_n,τ_n,h_n}

$$\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma).$$

Equation (1.6), the mutual contiguity and Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) allow us to conclude that

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] = o_{P_{\gamma_n,\tau_n,h_n}}(1)$$

It follows that under P_{γ_n,τ_n,h_n}

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_n} = \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} + \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma).$$

□

Proof of lemma B.1. For 1, use (1.5) to obtain that under P_{γ_n}

$$\Lambda_n(\gamma_n(\tau_n,h_n),\gamma_n(\tau,h)) = \Lambda_n(\gamma_n(\tau_n,h_n),\gamma_n) - \Lambda_n(\gamma_n(\tau,h),\gamma_n) = o_{P_{\gamma_n}}(1),$$

and so by the continuous mapping theorem, the mutual contiguity of $\left(P^n_{\gamma_n}\right)_{n\in\mathbb{N}}$ and $\left(P^n_{\gamma_n,\tau,h}\right)_{n\in\mathbb{N}}$ (Proposition 1.3.1) and Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4)

$$\exp(\Lambda_n(\gamma_n(\tau_n,h_n),\gamma_n(\tau,h))) \rightsquigarrow 1, \quad \text{under } P_{\gamma_n,\tau,h}.$$

Since ψ_n is bounded between 0 and 1, it is tight under $P_{\gamma_n,\tau,h}$ and hence by Prohorov's theorem (e.g. Billingsley, 1999, Theorem 5.1) for any subsequence $(n_j)_{j\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ there is a further subsequence $(n_k)_{k\in\mathbb{N}}$ such that $\psi_{n_k} \rightsquigarrow \psi$ for some $\psi \in [0,1]$ under $P_{\gamma_n,\tau,h}$. In conjunction with the preceding display, Slutsky's lemma yields

$$(\psi_n, \exp(\Lambda_n(\gamma_n(\tau_n,h_n),\gamma_n(\tau,h)))) \rightsquigarrow (\psi, 1) \quad \text{under } P_{\gamma_n,\tau,h}.$$

By Le Cam's third lemma (e.g. van der Vaart, 1998, Theorem 6.6) we have that under P_{γ_n,τ_n,h_n}, the law of ψ_{n_k} converges weakly to the law of ψ in the preceding display. Since each $\psi_n \in [0,1]$ it is both uniformly $P_{\gamma_n,\tau,h}$-integrable and uniformly P_{γ_n,τ_n,h_n}-integrable. These observations imply that

$$\lim_{k\to\infty}\left[P^{n_k}_{\gamma_{n_k},\tau_{n_k},g_{n_k}}\psi_n - P^{n_k}_{\gamma_{n_k},\tau,h}\psi_{n_k}\right] = 0.$$

Since the original subsequence $(n_j)_{j\in\mathbb{N}}$ was arbitrary, this holds also for the original sequence.

For 2, from (1.5), assumption CM(i) and the hypothesis that $B_\gamma h_1 = B_\gamma h_2$

$$\begin{aligned}\Lambda_n(\gamma_n(\tau_n,h_{1,n}),\gamma_n(\tau_n,h_{2,n})) &= \Lambda_n(\gamma_n(\tau_n,h_{1,n}),\gamma_n) - \Lambda_n(\gamma_n(\tau_n,h_{2,n}),\gamma_n)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n B_{\gamma_n}(h_1-h_2) + o_{P_{\gamma_n}}(1).\end{aligned}$$

$h := h_1 - h_2 \in H_\eta$ by assumption. Let $h_n := h$ for each $n \in \mathbb{N}$ and form g_n as in the proof of proposition 1.3.1 with $\tau = 0$. Argue analogously to the the proof of proposition 1.3.1 (noting that for this purpose assumption CM(i) rather than CM(ii) is sufficient) to obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n B_{\gamma_n}(h_1-h_2) = \frac{1}{\sqrt{n}}\sum_{i=1}^n B_{\gamma_n}h \rightsquigarrow \mathcal{N}(0,\sigma^2_{0,h}), \quad \text{under } P_{\gamma_n},$$

with $\sigma^2_{0,h} = P_\gamma\left[B_\gamma h\right]^2 = P_\gamma 0^2 = 0$. It follows from the two preceding displays that

$$\Lambda_n(\gamma_n(\tau_n,h_{1,n}),\gamma_n(\tau_n,h_{2,n})) = \Lambda_n(\gamma_n(\tau_n,h_{1,n}),\gamma_n) - \Lambda_n(\gamma_n(\tau_n,h_{2,n}),\gamma_n) = o_{P_{\gamma_n}}(1).$$

With this in hand, the second part of 2 can be established by an argument analogous to that used to establish 1. □

Proof of proposition B.2. That $\mathbb{H}_\gamma$ is a linear space is clear. Moreover, linearity, coordinate symmetry, and positive semi-definiteness of the map in (52) are clear from its definition.

It remains to prove that it is positive definite. Let Π denote the projection onto $\operatorname{cl} \mathscr{H}_\gamma \subset L_2(P_\gamma)$. Then, we can re-write

$$\langle (\tau, g), (\tau, g) \rangle = \tau' \tilde{\mathcal{I}}_\gamma \tau + \langle \tau' \Pi \dot{\ell}_\gamma + g, \tau' \Pi \dot{\ell}_\gamma + g \rangle_{P_\gamma}. \tag{58}$$

This is strictly positive whenever $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp \setminus \{0\}$.[79] If instead $\tau = 0$ but $g \neq 0$ it is positive since $\langle \cdot, \cdot \rangle_{P_\gamma}$ is an inner product. Thus $\langle \cdot, \cdot \rangle_{\mathbb{H}_\gamma}$ is an inner product and $(\mathbb{H}_\gamma, \langle \cdot, \cdot \rangle)$ is an inner-product space. Denote the completion of this space with respect to the norm induced by $\langle \cdot, \cdot \rangle$ as $(\overline{\mathbb{H}_\gamma}, \langle \cdot, \cdot \rangle)$.

A Gaussian shift on $(\mathbb{H}_\gamma, \langle \cdot, \cdot \rangle)$ is the restriction to $\mathbb{H}_\gamma$ of the standard Gaussian shift experiment of the Hilbert space $(\overline{\mathbb{H}_\gamma}, \langle \cdot, \cdot \rangle)$. Define

$$L_n(\tau, g) \coloneqq \Lambda(\gamma_n(\tau, h_g), \gamma_n) + \frac{1}{2} \| (\tau, g) \|^2, \tag{59}$$

and note that equation (55) and the marginal convergence of the log-likelihood (cf. equation (56)):

$$\sqrt{n} \mathbb{P}_n \tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h_g \rightsquigarrow \mathcal{N}\left(0, \sigma^2_{\tau,g}\right) \qquad \text{under } P_{\gamma_n}, \tag{60}$$

remain valid in this setting, where we write $\sigma^2_{\tau,g}$ for σ^2_{τ,h_g}.[80] By equation (55)

$$\| (\tau, g) \|^2 = \sigma^2_{\tau,g} = P_\gamma \left[\tau' \dot{\ell}_\gamma + g \right]^2 = \lim_{n \to \infty} P_{\gamma_n} \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h_g \right]^2. \tag{61}$$

Equations (1.5), (59) and (61) allow us to write

$$L_n(\tau, g) = \sqrt{n} \mathbb{P}_n \left[\tau' \dot{\ell}_{\gamma_n} + B_{\gamma_n} h_g \right] + o_{P_{\gamma_n}}(1),$$

and hence by (60),

$$L_n(\tau, g) \rightsquigarrow \mathcal{N}\left(0, \| (\tau, g) \|^2\right) \text{ under } P_{\gamma_n}, \qquad \text{for any } (\tau, g) \in \mathbb{H}_\gamma. \tag{62}$$

Moreover, for any $(\tau_1, g_1), (\tau_2, g_2) \in \mathbb{H}_\gamma$ and any $a_1, a_2 \in \mathbb{R}$ we have, where $R_{n,i} =$

[79] Suppose $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ and $\tau' \tilde{\mathcal{I}}_\gamma \tau = 0$. The latter implies that $\tilde{\mathcal{I}}_\gamma^{1/2} \tau = 0$, and hence $\tilde{\mathcal{I}}_\gamma \tau = \tilde{\mathcal{I}}_\gamma^{1/2} \tilde{\mathcal{I}}_\gamma^{1/2} \tau = 0$; i.e. $\tau \in N(\tilde{\mathcal{I}}_\gamma)$. Since τ is also in $N(\tilde{\mathcal{I}}_\gamma)^\perp$ we must have $\tau' \tau = 0$, i.e. $\tau = 0$.

[80] Proposition 1.3.1 requires assumption CM(ii) rather than the weaker CM(i). It is easy to see that an analogous argument as to that given in the proof of proposition 1.3.1 concerned only with marginal weak convergence of the log-likelihood in equation (56) holds under the weaker condition.

$o_{P_{\gamma_n}}(1)$ for $i = 1, 2, 3$,

$$\begin{aligned}
&a_1 L_n(\tau_1, g_1) + a_2 L_n(\tau_2, g_2) - L_n(a_1\tau_1 + a_2\tau_2, a_1 g_1 + a_2 g_2) \\
&\quad = a_1\sqrt{n}\mathbb{P}_n\left[\tau_1'\dot{\ell}_{\gamma_n} + B_{\gamma_n} h_{g_1}\right] + a_1 R_{n,1} - a_2\sqrt{n}\mathbb{P}_n\left[\tau_2'\dot{\ell}_{\gamma_n} + B_{\gamma_n} h_{g_2}\right] + a_2 R_{n,2} \\
&\qquad - \sqrt{n}\mathbb{P}_n\left[(a_1\tau_1 + a_2\tau_2)'\dot{\ell}_{\gamma_n} + B_{\gamma_n}[a_1 h_{g_1} + a_2 h_{g_2}]\right] + R_{n,3} \\
&\quad = a_1 R_{n,1} + a_2 R_{n,2} + R_{n,3} \\
&\quad = o_{P_{\gamma_n}}(1).
\end{aligned}$$

That is,

$$\begin{aligned}
a_1 L_n(\tau_1, g_1) + a_2 L_n(\tau_2, g_2) - L_n(a_1\tau_1 + a_2\tau_2, a_1 g_1 + a_2 g_2) = o_{P_{\gamma_n}}(1), \\
\text{whenever } a_1, a_2 \in \mathbb{R},\ (\tau_1, g_1), (\tau_2, g_2) \in \mathbb{H}_\gamma.
\end{aligned} \tag{63}$$

By imitating the proof of Theorem 69.4 in Strasser (1985), one obtains that the experiment

$$\mathscr{E} = (\Omega, \mathcal{F}, \{G_{\tau,g} : (\tau, g) \in \mathbb{H}_\gamma\}) \tag{64}$$

is the restriction to $\mathbb{H}_\gamma$ of a Gaussian shift experiment on $(\overline{\mathbb{H}_\gamma}, \langle\cdot,\cdot\rangle)$ if and only if the stochastic process $(L(\tau, g))_{(\tau,g)\in\mathbb{H}_\gamma}$, defined by

$$L(\tau, g) = \Lambda((\tau, h_g), (0, 0)) + \frac{1}{2}\|(\tau, g)\|^2, \tag{65}$$

with $\Lambda((\tau, h_g), (0, 0))$ the log-likelihood ratio of $G_{\tau,g}$ and $G_{(0,0)}$, is the restriction to $\mathbb{H}_\gamma$ of a standard Gaussian process defined on $\overline{\mathbb{H}_\gamma}$ under $G_{(0,0)}$.[81] Combining equations (62) and (63) we have that for any $K \in \mathbb{N}$, $a \in \mathbb{R}^K$ and $(\tau_k, g_k) \in \mathbb{H}_\gamma$ (for $k = 1, \ldots, K$) we have that under P_{γ_n}

$$\sum_{k=1}^K a_k L_n(\tau_k, g_k) \rightsquigarrow \sum_{k=1}^K a_k L^*(\tau_k, g_k) = L^*\left(\sum_{k=1}^K a_k(\tau_k, g_k)\right), \tag{66}$$

for a square integrable stochastic process L^* defined on $\mathbb{H}_\gamma$. Thus we have convergence of the finite dimensional marginal distributions of L_n to those of L^* by the Cramér-Wold theorem. Imitating the proof of Theorem 68.4 in Strasser (1985) yields that a square integrable stochastic process L defined on $\mathbb{H}_\gamma$ is the restriction to $\mathbb{H}_\gamma$ of a standard Gaussian process defined on $\overline{\mathbb{H}_\gamma}$ if and only if L is linear and has a $\mathcal{N}\left(0, \|(\tau, g)\|^2\right)$ marginal distribution for each $(\tau, g) \in \mathbb{H}_\gamma$. Since our process L^* satisfies these conditions, it follows that it is such a restriction of a standard Gaussian process. Therefore we have convergence of the finite dimensional distributions of $(L_n(\tau, g))_{(\tau,g)\in\mathbb{H}_\gamma}$ to those of (the restriction to $\mathbb{H}_\gamma$ of) a standard Gaussian process (on $(\overline{\mathbb{H}_\gamma}, \langle\cdot,\cdot\rangle)$). By (59) and (65) this implies the

[81] Such a standard Gaussian process is a square integrable stochastic process such that all its finite dimensional distributions are Gaussian with $\mathbb{E}L(\tau_1, g_1) = 0$ and $\mathbb{E}[L(\tau_1, g_1)L(\tau_2, g_2)] = \langle(\tau_1, g_1), (\tau_2, g_2)\rangle$ for all $(\tau_1, g_1), (\tau_2, g_2) \in \overline{\mathbb{H}_\gamma}$.

convergence of the finite dimensional distributions of $(\Lambda_n(\gamma_n(\tau, h_g), \gamma_n))_{(\tau,g)\in\mathbb{H}_\gamma}$ to those of $(\Lambda((\tau, h_g), (0, 0)))_{(\tau,g)\in\mathbb{H}_\gamma}$. With this in hand, the proof is completed by an appeal to Theorem 61.6 of Strasser (1985), upon noting that that the sequence of experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$ is contiguous (see e.g. Strasser, 1985, Definition 61.1) by an analogous argument as used to prove the contiguity claimed in proposition 1.3.1 and the transitivity of (mutual) contiguity. □

Lemma B.3. *Suppose that assumptions M, LAN, CM(ii), E and R hold for a sequence $(\gamma_n)_{n\in\mathbb{N}} \subset \Gamma$ with limit $\gamma \in \Gamma$. Then, for any $h_n \to h$ with each $h_n, h \in H_\eta$*

$$\lim_{n\to\infty} P^n_{\gamma_n,0,h_n}\phi_{n,\theta_n} = \begin{cases} \alpha & \textit{if } \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0 \\ 0 & \textit{if } \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) = 0 \end{cases}.$$

Proof of Lemma B.3. By proposition 1.3.1 we have that under $P_{\gamma_n,0,h_n}$

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_n} \rightsquigarrow \mathcal{N}(0, \tilde{\mathcal{I}}_\gamma).$$

Equations (1.7), (1.8) and Lemma C.6 imply that $\|\hat{\mathcal{I}}^\dagger_{n,\theta_n} - \tilde{\mathcal{I}}^\dagger_\gamma\|_2 = o_{P_{\gamma_n}}(1)$. The mutual contiguity established in proposition 1.3.1 along with Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) ensures that this result and equation (1.6) also hold under $P_{\gamma_n,0,h_n}$:

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] = o_{P_{\gamma_n,0,h_n}}(1) \quad \text{and} \quad \|\hat{\mathcal{I}}^\dagger_{n,\theta_n} - \tilde{\mathcal{I}}^\dagger_\gamma\|_2 = o_{P_{\gamma_n,0,h_n}}(1).$$

Write $\hat{Z}_n := \sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_n}$. We have

$$\hat{Z}_n = \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} + \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{\mathcal{I}}_\gamma)$$

under $P_{\gamma_n,0,h_n}$. We now cover the case of one-sided and two-sided tests separately. In the case of a two-sided test, the continuous mapping theorem implies that

$$\hat{S}_{n,\theta_n} = \hat{Z}_n'\hat{\mathcal{I}}^\dagger_{n,\theta_n}\hat{Z}_n \rightsquigarrow Z'\tilde{\mathcal{I}}^\dagger_\gamma Z =: S \sim \chi^2_r,$$

under $P_{\gamma_n,0,h_n}$ where $r = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$.[82]

Let c_n be the $1-\alpha$ quantile of the $\chi^2_{r_n}$ distribution and c the $1-\alpha$ quantile of the χ^2_r distribution. We have $P_{\gamma_n}\{c_n = c\} = P_{\gamma_n}\{r_n = r\} \to 1$ by assumption. This implies that $c_n - c \to 0$ in P_{γ_n}-probability and hence by the mutual contiguity and Le Cam's first lemma, also under $P_{\gamma_n,0,h_n}$. By continuous mapping once more we have $\hat{S}_{n,\theta_n} - c_n \rightsquigarrow S - c$ under $P_{\gamma_n,0,h_n}$.

Now, consider first the case where $r > 0$. In this case, since the χ^2_r distribution is continuous

[82] The distributional result is given by, for example, Theorem 9.2.2 in Rao and Mitra (1971).

the portmanteau theorem gives

$$P_{\gamma_n,0,h_n}\phi_{n,\theta_n} = P_{\gamma_n,0,h_n}\left(\hat{S}_{n,\theta_n} - c_n > 0\right) \to L\left(S - c > 0\right) = \alpha,$$

where L is the law of S. In the case where instead $r = 0$ we note that on the sets $\{r_n = r\} = \{r_n = 0\}$ we have that $\hat{\mathcal{I}}^{\dagger}_{n,\theta_n} = 0$ and $c_n = 0$ and hence do not reject since $\hat{S}_{n,\theta_n} = 0 \leq c_n = 0$. It follows that $P_{\gamma_n,0,h_n}\phi_{n,\theta_n} \leq 1 - P_{\gamma_n,0,h_n}\{r_n = r\} \to 0$.

Finally consider a one-sided test with $d_\theta = 1$ and $1 - \alpha \in [1/2, 1)$. By the continuous mapping theorem,

$$\hat{S}_{n,\theta_n} = \hat{Z}_n\sqrt{\hat{\mathcal{I}}^{\dagger}_{n,\theta_n}} \rightsquigarrow Z\sqrt{\tilde{\mathcal{I}}^{\dagger}_{\gamma}}.$$

If $r = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) = 1$, then $Z\sqrt{\tilde{\mathcal{I}}^{\dagger}_{\gamma}} = Z/\sqrt{\tilde{\mathcal{I}}_\gamma} \sim \mathcal{N}(0,1)$ and since this distribution is continuous, the portmanteau theorem yields

$$P_{\gamma_n,0,h_n}\phi_{n,\theta_n} \to 1 - \Phi(z_\alpha) = \alpha,$$

where Φ is the CDF of the standard normal distribution. If, instead $r = 0$, then again on the sets where $r_n = \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) = 0$ we have that $\hat{\mathcal{I}}_{n,\theta_n} = \hat{\mathcal{I}}^{\dagger}_{n,\theta_n} = 0$ and so $\hat{S}_{n,\theta_n} = 0 \leq z_\alpha$ and hence we do not reject. It follows that $P_{\gamma_n,0,h_n}\phi_{n,\theta_n} \leq 1 - P_{\gamma_n,0,h_n}\{r_n = r\} \to 0$. □

Lemma B.4. *Suppose that assumptions M, LAN, CM(ii), E and R hold for a convergent sequence $(\gamma_n)_{n\in\mathbb{N}} \subset \Gamma$ with limit $\gamma \in \Gamma$. Suppose we are given a convergent sequences $h_{n_k} \to h \in H_\eta$ with $(h_{n_k})_{k\in\mathbb{N}} \subset H_\eta$. If the limit*

$$\mathcal{S} := \lim_{k\to\infty} P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}} \tag{67}$$

exists, then $\mathcal{S} = \alpha \times \mathbf{1}\{\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0\}$.

Proof. The idea is to construct a new sequence to which proposition B.3 can be applied.[83] For all m with $m \in [n_k, n_{k+1}) \cap \mathbb{N}$ for some $k \in \mathbb{N}$ put $h^*_m = h_{n_k}$. For $m = 1, \ldots, n_1$, put $h^*_m = h_{n_1}$. For each m let $\gamma^*_m = \gamma_m$. By construction $h^*_m \to h$, and by our hypotheses and proposition B.3 we may conclude that

$$\lim_{m\to\infty} P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} = s_\gamma := \begin{cases} \alpha & \text{if } \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0 \\ 0 & \text{if } \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) = 0 \end{cases}.$$

Fix an arbitrary $\varepsilon > 0$. There is a $M \in \mathbb{N}$ such that for all $m \geq M$, $\left|P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} - s_\gamma\right| < \varepsilon/2$. By (67) there is a $K \in \mathbb{N}$ such that if $k \geq K$, $\left|\mathcal{S} - P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}}\right| < \varepsilon/2$. Hence for any k sufficiently large that $m = n_k \geq M$

[83] This construction is based on that used in the proofs of e.g. Lemma 6 in Andrews and Guggenberger (2010b), Lemma 2.1 in Andrews and Cheng (2012).

and $k \geq K$ we have

$$|\mathcal{S} - s_\gamma| \leq \left|\mathcal{S} - P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m}\right| + \left|P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} - s_\gamma\right| < \left|\mathcal{S} - P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}}\right| + \frac{\varepsilon}{2} < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the inequality $|\mathcal{S} - s_\gamma| < \varepsilon$ can be obtained for any $\varepsilon > 0$ and hence taking the limit as $\varepsilon \downarrow 0$ completes the proof. □

Proof of proposition 1.3.2. There is a sequence $(h_n)_{n\in\mathbb{N}} \subset H'_\eta$ and a subsequence $(n_j)_{j\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ such that

$$\mathcal{S} := \limsup_{n\to\infty} \sup_{h\in H'_\eta} P^n_{\gamma_n,0,h}\phi_{n,\theta_n} = \limsup_{n\to\infty} P^n_{\gamma_n,0,h_n}\phi_{n,\theta_n} = \lim_{j\to\infty} P^{n_j}_{\gamma_{n_j},0,h_{n_j}}\phi_{n_j,\theta_{n_j}}$$

There is a further subsequence $(n_k)_{k\in\mathbb{N}}$ such that $h_{n_k} \to h$ and $\mathcal{S} = \lim_{k\to\infty} P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}}$. Applying lemma B.4 yields that $\mathcal{S} = \alpha \times \mathbf{1}\{\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) > 0\}$. Since an analogous argument can be made to obtain the same conclusion but with “”lim inf” replacing “”lim sup” in the definition of $\mathcal{S}$, we obtain the desired result. □

Lemma B.5. *Fix a convergent sequence $(\eta_n)_{n\in\mathbb{N}}$ and denote its limit by η. Suppose that assumptions M, LAN, CM(ii), E and R hold for any sequence $(\gamma_n)_{n\in\mathbb{N}}$ where each $\gamma_n := (\theta_n, \eta_n)_{n\in\mathbb{N}} \subset \Theta' \times \mathcal{H} =: \Gamma'$ with $\theta_n \to \theta \in \Theta' \subset \Theta$. Suppose we are given convergent sequences $\gamma_{n_k} \to \gamma$ with $(\gamma_{n_k})_{k\in\mathbb{N}} \subset \Gamma'$ and $h_{n_k} \to h$ with $(h_{n_k})_{k\in\mathbb{N}} \subset H_\eta$. If the limit*

$$\mathcal{S} := \lim_{k\to\infty} P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}} \tag{68}$$

exists, then $\mathcal{S} \leq \alpha$.

Proof. The idea is to construct a new sequence to which proposition B.3 can be applied.[84] For all m with $m \in [n_k, n_{k+1}) \cap \mathbb{N}$ for some $k \in \mathbb{N}$ put $\theta^*_m = \theta_{n_k}$ and $h^*_m = h_{n_k}$. For $m = 1, \ldots, n_1$, put $\theta^*_m = \theta_{n_1}$ and $h^*_m = h_{n_1}$. For each m let $\gamma^*_m = (\theta^*_m, \eta_m)$. By construction $\gamma^*_m \to \gamma$ through Γ' and $h^*_m \to h$, and by our hypotheses and proposition B.3 we may conclude that

$$\lim_{m\to\infty} P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} \leq \alpha.$$

Fix an arbitrary $\varepsilon > 0$. There is a $M \in \mathbb{N}$ such that for all $m \geq M$, $P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} \leq \alpha + \varepsilon/2$. By (68) there is a $K \in \mathbb{N}$ such that if $k \geq K$, $\left|\mathcal{S} - P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}}\right| < \varepsilon/2$. Hence for any k sufficiently large that $m = n_k \geq M$ and $k \geq K$ we have

$$\mathcal{S} \leq \left|\mathcal{S} - P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m}\right| + P^m_{\gamma^*_m,0,h^*_m}\phi_{m,\theta^*_m} < \left|\mathcal{S} - P^{n_k}_{\gamma_{n_k},0,h_{n_k}}\phi_{n_k,\theta_{n_k}}\right|\alpha + \frac{\varepsilon}{2} \leq \alpha + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we can obtain the inequality $\mathcal{S} \leq \alpha + \varepsilon$ for any $\varepsilon > 0$ and hence taking the limit as $\varepsilon \downarrow 0$ completes the proof. □

[84] See footnote 83.

Proof of proposition 1.3.3. There are sequences $(\theta_n)_{n\in\mathbb{N}} \subset \Theta'$ and $(h_n)_{n\in\mathbb{N}} \subset H'_\eta$ and a subsequence $(n_j)_{j\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ such that

$$\mathcal{S} := \liminf_{n\to\infty} \inf_{\theta\in\Theta'} \inf_{h\in H'_\eta} P^n_{(\theta,\eta_n),0,h}(\theta \in \hat{C}_n) = \lim_{j\to\infty} P^{n_j}_{(\theta_{n_j},\eta_{n_j}),0,h_{n_j}}(\theta_{n_j} \in \hat{C}_{n_j}).$$

There is a further subsequence $(n_k)_{k\in\mathbb{N}}$ of $(n_j)_{j\in\mathbb{N}}$ such that $\theta_{n_j} \to \theta \in \Theta'$ and $h_{n_j} \to h \in H'_\eta$. We also clearly have

$$\mathcal{S} = \lim_{k\to\infty} P^{n_k}_{(\theta_{n_k},\eta_{n_k}),0,h_{n_k}}(\theta_{n_k} \in \hat{C}_{n_k}) = 1 - \lim_{k\to\infty} P^{n_k}_{(\theta_{n_k},\eta_{n_k}),0,h_{n_k}} \phi_{n_k,\theta_{n_k}}. \tag{69}$$

Apply lemma B.5 to conclude that $1-\mathcal{S} \le \alpha$, and rearrange to obtain the desired result. □

Proof of proposition 1.3.4. By (both parts of) lemma B.1, it suffices to show that

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau,h}\psi_n \le 1 - \Phi\left(z_\alpha - \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right) \qquad \text{for all } \tau > 0,\ h \in \mathfrak{H}_\gamma. \tag{70}$$

Since $d_\theta = 1$ and $\tilde{\mathcal{I}}_\gamma > 0$, $N(\tilde{\mathcal{I}}_\gamma)^\perp = \mathbb{R}$. Let $\tilde{g} = (g_k)_{k\in\mathbb{N}} \subset \mathscr{H}_\gamma$ be an orthonormal basis of $\operatorname{cl}\mathscr{H}_\gamma$.[85] Consider the subspace $\mathcal{G}^m := \operatorname{Span}\{g_1,\ldots,g_m\}$, and let Π^m denote the orthogonal projection onto $\mathcal{G}^m$. Fix $b = (\tau, g_b) \in (0,\infty) \times \mathscr{H}_\gamma =: K_1$ and any $\varepsilon > 0$.[86] By lemma C.2 we can take $m \in \mathbb{N}$ large enough that $\left\|(\Pi^m - \Pi)\dot{\ell}_\gamma\right\|_{P_\gamma,2} < \varepsilon$. Now consider the restriction of $\mathscr{E}$ to $\mathbb{R} \times \mathcal{G}^m$ for any $m \in \mathbb{N}$.[87] Choose $a = (0, g_a)$ with $g_a = \Pi^m\left(\tau\Pi\dot{\ell}_\gamma + g_b\right) = \tau\left(\Pi^m\Pi\dot{\ell}_\gamma\right) + g_b$ and note that by Lemma 28.1 of Strasser (1985) any test ψ of level$-\alpha$ of H_0 against H_1 satisfies

$$G_b\psi \le 1 - \Phi\left(z_\alpha - \|b-a\|\right)$$

Expand the square of the norm using the Pythagorean theorem to obtain

$$\|b-a\|^2 = \tau^2\tilde{\mathcal{I}}_\gamma + \tau^2\left\|(\Pi^m - \Pi)\dot{\ell}_\gamma\right\|^2_{P_\gamma,2} = \tau^2\tilde{\mathcal{I}}_\gamma + \tau^2\varepsilon^2.$$

Hence we have

$$G_b\psi \le 1 - \Phi\left(z_\alpha - \sqrt{\tau^2\tilde{\mathcal{I}}_\gamma + \tau^2\varepsilon^2}\right).$$

Since $\varepsilon > 0$ was arbitrary, we can take the limit as $\varepsilon \downarrow 0$ to obtain

$$G_b\psi \le 1 - \Phi\left(z_\alpha - \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right), \tag{71}$$

[85] Such a basis always exists: by assumption M, $\mathcal{W}$ is Polish. Take a metric d such that $(\mathcal{W}, d)$ is a complete (separable) metric space. By Theorem 1.3 in Billingsley (1999), P_γ is tight. By Proposition 7.14.12 in Bogachev (2007) this is a sufficient condition for separability of P_γ which is equivalent to separability of the $L_p(P_\gamma)$ spaces for $p \in (0,\infty)$ (e.g. Bogachev, 2007, Exercise 4.7.63). $\operatorname{cl}\mathscr{H}_\gamma$ is therefore separable as a subset of $L_2(P_\gamma)$. Choose a countable dense subset in $\mathscr{H}_\gamma$ and apply Gram-Schmidt to obtain an orthonormal basis which satisfies the the desired property.

[86] We can always change the choice of the orthonormal basis such that g_b lies in (each) $\mathcal{G}^m$.

[87] See equations (64), (65) and the surrounding text for the definitions of $\mathscr{E}$ and $G_{\tau,g}$.

which holds for all $b \in K_1$, since the choice of $b \in K_1$ was arbitrary. Moreover, since the test ψ was an arbitrary test of level-α, this power bound holds for all level-α tests in $\mathscr{E}$.

By proposition B.2 the the sequence of experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$ defined in (53) converge to the dominated experiment $\mathscr{E}$. (70) then follows on combining the power bound given by (71) with Theorem 7.2 in van der Vaart (1991). □

Proof of corollary 1.3.5. Since $\tilde{\mathcal{I}}_\gamma > 0$ and $d_\theta = 1$, assumption R is automatically satisfied given assumption E. By proposition 1.3.1 we have that

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_0}/\hat{\mathcal{I}}_{n,\theta_0}^{1/2} \rightsquigarrow N(\tilde{\mathcal{I}}_\gamma^{1/2}\tau, 1), \text{ under } P_{\gamma_n,\tau_n,h_n}.$$

Hence by the portmanteau theorem

$$\lim_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\phi_n = \lim_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}(\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_0}/\hat{\mathcal{I}}_{n,\theta_0}^{1/2} > z_\alpha) = 1 - \Phi(z_\alpha - \tilde{\mathcal{I}}_\gamma^{1/2}\tau).$$

For $\tau \leq 0$, $1 - \Phi(z_\alpha - \tilde{\mathcal{I}}_\gamma^{1/2}\tau) \leq \alpha$; hence this test is level-α as claimed. For any $\tau > 0$, it attains the power bound in equation (1.15). □

Proof of proposition 1.3.6. The proof is is very similar to that of proposition 1.3.4. By lemma B.1 it suffices to show that for all $\tau \neq 0$ and $h \in \mathfrak{H}_\gamma$

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau,h}\psi_n \leq 1 - \Phi\left(z_{\alpha/2} - \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right) + 1 - \Phi\left(z_{\alpha/2} + \tilde{\mathcal{I}}_\gamma^{1/2}\tau\right). \tag{72}$$

Since $d_\theta = 1$ and $\tilde{\mathcal{I}}_\gamma > 0$, $N(\tilde{\mathcal{I}}_\gamma)^\perp = \mathbb{R}$. Let $\tilde{g}$, $\mathcal{G}^m$ and Π^m be defined as in the proof of proposition 1.3.4 and consider the restriction of $\mathscr{E}$ to $L^m := \mathbb{R} \times \mathcal{G}^m$ for some $m \in \mathbb{N}$ which contains $(\tau, g) \in K_1 = \{(\tau, g) : \tau \neq 0,\ h \in \mathscr{H}_\gamma\}$.[88] This is a finite dimensional (hence closed) subspace of $\overline{\mathbb{H}_\gamma}$ (the completion of $\mathbb{H}_\gamma$) and so is a Hilbert space. Hence this restriction is a finite dimensional (standard) Gaussian shift. Take $f : \mathbb{R} \times \mathcal{G}^m \to \mathbb{R}$ as $f(\tau, g) = \tau$ and let $\Sigma^m := P_\gamma\left([I - \Pi^m]\dot{\ell}_\gamma\right)^2$, which can be ensured positive by taking $m \in \mathbb{N}$ sufficiently large.[89] Then, letting $g \in \mathcal{G}^m$ be such that $g = -\Pi^m\dot{\ell}_\gamma \in \mathcal{G}^m$, $e = (1, g)/\sqrt{\Sigma^m}$ is a unit vector in $\mathbb{R}\times\mathcal{G}^m \subset \mathbb{H}_\gamma$, orthogonal to $N(f) = \{(0, g) : g \in \mathcal{G}^m\}$ and has $f(e) = 1/\sqrt{\Sigma^m} > 0$. Thus, by Theorem 28.8 of Strasser (1985), any unbiased test ψ of level-α has power bounded by

$$G_{\tau,g}\psi \leq 1 - \Phi(z_{\alpha/2} - (\Sigma^m)^{1/2}\tau) + 1 - \Phi(z_{\alpha/2} + (\Sigma^m)^{1/2}\tau).$$

Since $\Sigma^m \to \tilde{\mathcal{I}}_\gamma$ as $m \to \infty$, by continuity we obtain that

$$G_{\tau,g}\psi \leq 1 - \Phi(z_{\alpha/2} - \tilde{\mathcal{I}}_\gamma^{1/2}\tau) + 1 - \Phi(z_{\alpha/2} + \tilde{\mathcal{I}}_\gamma^{1/2}\tau). \tag{73}$$

[88] See footnote 86.

[89] By lemma C.2 we have that $\Sigma^m \to \tilde{\mathcal{I}}_\gamma > 0$ as $m \to \infty$.

Since the point $(\tau, g) \in K_1$ was arbitrary, this bound holds for all K_1.

By proposition B.2 the sequence of experiments $(\mathscr{E}_n)_{n\in\mathbb{N}}$ converges to the dominated experiment $\mathscr{E}$. Let $\pi_n(\tau, g) := P^n_{\gamma_n,\tau,g}\psi_n \in [0, 1]$. Fix a $(\tau, g) \in K_1$ and let $(n_j)_{j\in\mathbb{N}}$ be a subsequence of $(n)_{n\in\mathbb{N}}$ along which $\limsup_{n\to\infty} P^n_{\gamma_n,\tau,g}\psi_n = \lim_{j\to\infty} P^{n_j}_{\gamma_{n_j},\tau,g}\psi_{n_j}$. Since $[0, 1]^{\mathbb{H}_\gamma}$ is compact in the product topology there is a subnet $(n_{j(\alpha)})_{\alpha\in A}$ of the subsequence $(n_j)_{j\in\mathbb{N}}$ and a function $\pi : \mathbb{H}_\gamma \to [0, 1]$ such that $\lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, g) = \pi(\tau, g)$ for every $(\tau, g) \in \mathbb{H}_\gamma$. By Theorem 7.1 in van der Vaart (1991) there is a test ψ in $\mathscr{E}$ with power function π. By our hypotheses and the pointwise convergence we have that for any $\tau \neq 0$ and any $g_1, g_2 \in \mathscr{H}_\gamma$

$$\pi(0, g_1) = \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(0, g_1) \leq \alpha \leq \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, g_2) = \pi(\tau, g_2).$$

It follows that ψ is unbiased and hence combining

$$\limsup_{n\to\infty} P^n_{\gamma_n,\tau,g}\psi_n = \limsup_{n\to\infty} \pi_n(\tau, g) = \lim_{j\to\infty} \pi_{n_j}(\tau, g) = \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, g) = \pi(\tau, g)$$

with the power bound given by (73) we obtain (72).[90] □

Proof of corollary 1.3.7. Since $\tilde{\mathcal{I}}_\gamma > 0$ and $d_\theta = 1$, assumption R is automatically satisfied given assumption E. By proposition 1.3.1 we have that

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_0}/\hat{\mathcal{I}}^{1/2}_{n,\theta_0} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}^{1/2}_\gamma\tau, 1), \text{ under } P_{\gamma_n,\tau_n,h_n}.$$

Let the $1 - \alpha$ quantile of the χ^2_1 distribution be denoted by c_α. By assumption R holds and the contiguity noted in proposition 1.3.1 we have that $P_{\gamma_n,\tau_n,h_n}(\hat{r}_n = 1) \to 1$ and hence $c_n \to c_\alpha$ in P_{γ_n,τ_n,h_n}-probability. Hence by the portmanteau theorem

$$\lim_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\phi_{n,\theta_0} = 1 - \Phi(z_{\alpha/2} - \tilde{\mathcal{I}}^{1/2}_\gamma\tau) + 1 - \Phi(z_{\alpha/2} + \tilde{\mathcal{I}}^{1/2}_\gamma\tau),$$

which is exactly the power bound given by equation (1.16). For $\tau = 0$, $1 - \Phi(z_{\alpha/2}) + 1 - \Phi(z_{\alpha/2}) = \alpha$; hence this test is level-α as claimed. It is unbiased since the last right hand side expression in the preceding display exceeds α for any $\tau \neq 0$. □

Lemma B.6. *If* $(\overline{\mathbb{H}_\gamma}, \langle\cdot,\cdot\rangle)$ *is the completion of* $(\mathbb{H}_\gamma, \langle\cdot,\cdot\rangle)$*, then*

1. *we can take* $\overline{\mathbb{H}_\gamma}$ *to be* $N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl}\mathscr{H}_\gamma$*;*
2. $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ *converges to* $(\tau, g) \in \overline{\mathbb{H}_\gamma}$ *if and only if* $\tau_n \to \tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ *and* $g_n \to g \in \operatorname{cl}\mathscr{H}_\gamma$*.*

Proof. We first note that $(x, y) \mapsto x'\tilde{\mathcal{I}}_\gamma y$ defines an inner-product on $N(\tilde{\mathcal{I}}_\gamma)^\perp$. Linearity and symmetry are obvious. Positive definiteness was established in footnote 79. On

[90] Where $g = B_\gamma h$ for the $h \in \mathfrak{H}_\gamma$ in the latter.

$\mathbb{R}^{d_\theta}$ it defines a positive-semidefinite Hermitian form and thus induces a semi-norm by $\|x\| := \sqrt{x'\tilde{\mathcal{I}}_\gamma x}$.

By the Pythagorean theorem we can decompose the square of the $\overline{\mathbb{H}_\gamma}$ norm as follows

$$\|(\tau_n, g_n) - (\tau, g)\|^2 = (\tau_n - \tau)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau) + \|(\tau_n - \tau)'\Pi\dot{\ell}_\gamma + g_n - g\|^2_{P_\gamma,2}. \tag{74}$$

We start with the first claim. Suppose that $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ is a Cauchy sequence. By (74) we must have that $(\tau_n - \tau_m)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau)m \to 0$ as $n, m \to \infty$. Let UDU' be an eigendecomposition of $\tilde{\mathcal{I}}_\gamma^{1/2}$ with eigenvalues $\lambda_1, \ldots, \lambda_{d_\theta}$ in decreasing order. Then the eigenvectors u_j for $j > r$ are in the null space of $\tilde{\mathcal{I}}_\gamma^{1/2}$ and so that of $\tilde{\mathcal{I}}_\gamma$. Letting U_1 be the $d_\theta \times r$ matrix of the first r columns of U and U_2 the remaining columns, we then have that $\|\tau_n - \tau_m\|_2 = \|U'(\tau_n - \tau_m)\|_2 = \|U_1'(\tau_n - \tau_m)\|_2$. Let $\tilde{\tau}_{n,m} := U_1'(\tau_n - \tau_m)$ and note that by hypothesis

$$(\tau_n - \tau_m)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau_m) = \sum_{i=1}^{r} \lambda_i \tilde{\tau}^2_{n,m,i} \to 0.$$

Since the λ_i are all positive this implies that $\|\tilde{\tau}_{n,m}\|_2 \to 0$, i.e. $\tau_n - \tau_m \to 0$. Since this is a Cauchy sequence in $N(\tilde{\mathcal{I}}_\gamma)^\perp$, which is a closed subspace of $\mathbb{R}^{d_\theta}$, it follows that τ_n has a limit, say $\tau^* \in N(\tilde{\mathcal{I}}_\gamma)^\perp$. From this and that $\left\|(\tau_n - \tau_m)'\Pi\dot{\ell}_\gamma + g_n - g_m\right\|_{P_\gamma,2} \to 0$ (as $m, n \to \infty$) we can also conclude that $(g_n)_{n\in\mathbb{N}}$ is Cauchy in $L_2(P_\gamma)$ and hence has a limit, say $g^* \in \operatorname{cl}\mathscr{H}_\gamma$.[91] Hence all such Cauchy sequences have limits in $N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl}\mathscr{H}_\gamma$ and so this is complete under the relevant norm.

To complete the proof we will now show that $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ converges to $(\tau, g) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl}\mathscr{H}_\gamma$ if and only if $\tau_n \to \tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$ and $g_n \to g \in \operatorname{cl}\mathscr{H}_\gamma$. Since this ensures that $N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl}\mathscr{H}_\gamma = \operatorname{cl}\mathbb{H}_\gamma$, this is the smallest closed set containing $\mathbb{H}_\gamma$, which completes the proof of the first part, and hence the second.

Suppose first that $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ converges to $(\tau, g) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl}\mathscr{H}_\gamma$. Then since each $\tau_n - \tau \in \mathbb{N}(\tilde{\mathcal{I}}_\gamma)^\perp$ we can argue as above via the same eigendecomposition (replacing τ_m with τ) to obtain that $\tau_n - \tau \to 0$. An argument analogous to that in footnote 91 (replace g_m with g) can be used to show the convergence $g_n \to g$ in the $L_2(P_\gamma)$ norm.

For the converse, suppose that $\tau_n \to \tau$ and $g_n \to g$. It follows immediately that $(\tau_n - \tau)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau) \to 0$ and $\|(\tau_n - \tau)'\Pi\dot{\ell}_\gamma\|_{P_\gamma,2} \to 0$. Using (74) we have

$$\|(\tau_n, g_n) - (\tau, g)\|^2 \lesssim (\tau_n - \tau)'\tilde{\mathcal{I}}_\gamma(\tau_n - \tau) + \|(\tau_n - \tau)'\Pi\dot{\ell}_\gamma\|^2_{P_\gamma,2} + \|g_n - g\|^2_{P_\gamma,2} = o(1).$$

□

[91] By the reverse triangle inequality we have

$$\lim_{n,m\to\infty} \|g_n - g_m\|_{P_\gamma,2} \leq \lim_{n,m\to\infty} \left\|(\tau_n - \tau_m)'\Pi\dot{\ell}_\gamma + g_n - g_m\right\|_{P_\gamma,2} = 0.$$

Proof of proposition 1.3.8. Let $\tilde{M}_a := \{(\tau, h) \in M_a : h \in \mathfrak{H}_\gamma\}$. We clearly have that

$$\limsup_{n\to\infty} \inf_{(\tau,h)\in M_a} P^n_{\gamma_n,\tau,h}\psi_n \leq \limsup_{n\to\infty} \inf_{(\tau,h)\in \tilde{M}_a} P^n_{\gamma_n,\tau,h}\psi_n,$$

so it will suffice to demonstrate the upper bound in

$$\limsup_{n\to\infty} \inf_{(\tau,h)\in \check{M}_a} P^n_{\gamma_n,\tau,g}\psi_n = \limsup_{n\to\infty} \inf_{(\tau,h)\in \tilde{M}_a} P^n_{\gamma_n,\tau,h}\psi_n \leq 1 - \mathrm{P}\left(\chi^2_r(a) \leq c_{r,\alpha}\right), \quad (75)$$

where $\check{M}_a := \{(\tau, g) \in \mathbb{H}_\gamma : \tau'\tilde{\mathcal{I}}_\gamma\tau \geq a\}$. We first observe that if $(\tau, g) \in \overline{\mathbb{H}_\gamma}$ then $\tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp$. Define $f : \overline{\mathbb{H}_\gamma} \to \mathbb{R}^{d_\theta}$ by $f(\tau, g) := \tau$ and let $L_0 := N(f)$. Let Π_0 denote the orthogonal projection onto L_0 in $\overline{\mathbb{H}_\gamma}$ and Π the orthogonal projection onto $\operatorname{cl}\mathscr{H}_\gamma$ in $L_2(P_\gamma)$. The (finite dimensional) subspace $L_0^\perp \subset \overline{\mathbb{H}_\gamma}$ consists of vectors

$$L_0^\perp = \left\{(\tau, -\tau'\Pi\dot{\ell}_\gamma) \in \overline{\mathbb{H}_\gamma}\right\}.$$

It follows from lemma B.6 that this has dimension r, since we can take $\overline{\mathbb{H}_\gamma} = N(\tilde{\mathcal{I}}_\gamma)^\perp \times \operatorname{cl}\mathscr{H}_\gamma$.

Consider the orthogonal projection onto L_0: we must have $\langle(\tau, g) - \Pi_0(\tau, g), (0, g')\rangle = 0$ for all $(0, g') \in L_0$. This implies that $\Pi_0(\tau, g) = (0, \tilde{g})$ must satisfy $\tilde{g} = \tau'\Pi\dot{\ell}_\gamma + g$. It follows that $\|(\tau, g) - \Pi_0(\tau, g)\|^2 = \tau'\tilde{\mathcal{I}}_\gamma\tau$. Define

$$\overline{M}_a = \left\{(\tau, h) \in \overline{\mathbb{H}_\gamma} : \tau'\tilde{\mathcal{I}}_\gamma\tau \geq a\right\},$$

and let $\overline{M}_a'$ be the set defined analogously to $\overline{M}_a$ where "$=$" replaces "$\geq$". We note here that $\overline{M}_a = \operatorname{cl}\check{M}_a$. For this, note firstly that any convergent $(t_n, g_n)_{n\in\mathbb{N}} \subset \overline{M}_a$ converges in $\overline{M}_a$ and hence this is a closed set.[92] It follows that $\operatorname{cl}\check{M}_a \subset \overline{M}_a$. Suppose that this inclusion were strict. Then there must be a point $(\tau, g) \in \overline{M}_a$ which is not the limit of a sequence $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \check{M}_a$. There must exist *a* sequence $(\tau_n, g_n)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ with $(\tau_n, g_n) \to (\tau, g)$. By the argument in footnote 92 we have that $\tau_n'\tilde{\mathcal{I}}_\gamma\tau_n \to \tau\tilde{\mathcal{I}}_\gamma\tau$. If the difference $e_n := \tau\tilde{\mathcal{I}}_\gamma\tau - \tau_n'\tilde{\mathcal{I}}_\gamma\tau_n \to 0$ is always negative there is nothing to do. Else take a sequence $(\tau_n', 0)_{n\in\mathbb{N}} \subset \mathbb{H}_\gamma$ which converges to $(0, 0)$ and satisfies $\tau_n'\tilde{\mathcal{I}}_\gamma\tau_n \geq \max\{e_n, 0\}$.[93] Then $(\tau_n + \tau_n', g_n)_{n\in\mathbb{N}} \subset \check{M}_a$ and converges to (τ, g). Hence no such point can exist and the two sets are equal.

Consider the testing problem of $K_0' = \{0\}$ against $K_1' = L_0^\perp \setminus \{0\}$ in the standard Gaussian

[92] That $(\tau, g) \in \overline{\mathbb{H}_\gamma}$ is clear since the latter is complete and hence closed. It remains to show that if $\tau_n\tilde{\mathcal{I}}_\gamma\tau_n \geq a$ for each $n \in \mathbb{N}$ then also $\tau\tilde{\mathcal{I}}_\gamma\tau \geq a$. For this, we note that if $(\tau_n, g_n) \to (\tau, g)$ then by lemma B.6 we have that $\tau_n \to \tau$. $(x, y) \mapsto x'\tilde{\mathcal{I}}_\gamma y$ defines a positive-semidefinite Hermitian form over $\mathbb{R}^{d_\theta}$ and thus induces a semi-norm $\|x\| := \sqrt{x'\tilde{\mathcal{I}}_\gamma x}$. Hence by the reverse triangle inequality

$$|\|\tau_n\| - \|\tau\|| \leq \|\tau_n - \tau\| \to 0.$$

That is $\|\tau_n\| \to \|\tau\|$ and hence by the continuity of $x \mapsto x^2$ we have $\tau_n\tilde{\mathcal{I}}_\gamma\tau_n = \|\tau_n\|^2 \to \|\tau\|^2 = \tau'\tilde{\mathcal{I}}_\gamma\tau$.

[93] An explicit construction of such a sequence can be given based on the eigendecomposition of $\tilde{\mathcal{I}}_\gamma$.

shift experiment on $L_0^\perp$. For any $a' \geq a$ and any level$-\alpha$ test ψ we have by Theorem 30.2 of Strasser (1985) that (Cf. Strasser, 1985, Theorem 71.10)

$$\inf_{t \in \overline{M}'_{a'}} G_t\psi \leq \inf_{t \in \overline{M}'_{a'} \cap L_0^\perp} G_t\psi \leq \mathrm{P}\left(\chi_r^2(a') > c_{r,\alpha}\right).$$

Since $\overline{M}_a = \operatorname{cl} \check{M}_a$ and $t \mapsto G_t\psi$ is continuous, taking the infimum over $a' \geq a$ yields[94]

$$\inf_{t \in \check{M}_a} G_t\psi = \inf_{t \in \overline{M}_a} G_t\psi \leq \mathrm{P}\left(\chi_r^2(a) > c_{r,\alpha}\right) =: \mathcal{R}. \tag{76}$$

By proposition B.2 $(\mathscr{E}_n)_{n\in\mathbb{N}}$ converges to $\mathscr{E}$. Suppose that (1.17) does not hold for all sequences of asymptotically level-α tests for $H_0 : \tau = 0$ against $H_1 : \tau \in N(\tilde{\mathcal{I}}_\gamma)^\perp \setminus \{0\}$ in $\mathscr{E}_n$. Then there is such a sequence of tests $(\psi_n)_{n\in\mathbb{N}}$ and a subsequence $(n_j)_{j\in\mathbb{N}}$ such that for some $\varepsilon > 0$

$$\lim_{j\to\infty} \inf\{\pi_{n_j}(\tau, h) : (\tau, h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta,\ \tau'\tilde{\mathcal{I}}_\gamma\tau \geq a\} \geq \mathcal{R} + \varepsilon,$$

where $\pi_n(\tau, h) := P^n_{\gamma_n,\tau,h}\psi_n$. Since $[0, 1]^{N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta}$ is compact in the product topology there is a subnet $(n_{j(\alpha)})_{\alpha\in A}$ of the subsequence $(n_j)_{j\in\mathbb{N}}$ and a function $\pi : \mathbb{N}(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta \to [0, 1]$ such that $\lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, h) = \pi(\tau, h)$ for every $(\tau, h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta$. Combine this with the preceding display to conclude that for any $(\tau, h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta$ with $\tau'\tilde{\mathcal{I}}_\gamma\tau \geq a$ we have

$$\pi(\tau, h) = \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, h) \geq \lim_{\alpha\in A} \inf\{\pi_{n_{j(\alpha)}}(\tau, h) : (\tau, h) \in N(\tilde{\mathcal{I}}_\gamma)^\perp \times H_\eta,\ \tau'\tilde{\mathcal{I}}_\gamma\tau \geq a\} \geq \mathcal{R} + \varepsilon.$$

However, by Theorem 7.1 in van der Vaart (1991) there is a test ψ in $\mathscr{E}$ with power function π and it follows from our hypothesis that this test is of level-α, since for any $g \in \mathscr{H}_\gamma$ there is a $h \in \mathfrak{H}_\gamma$ with $B_\gamma h = g$ and so

$$G_{0,g}\psi = \pi(0, h) = \lim_{\alpha\in A} \pi_{n_{j(\alpha)}}(\tau, h) \leq \limsup_n \pi_n(\tau, h) \leq \alpha.$$

Then by the preceding two displays we have $G_{0,g}\psi \leq \alpha$ for any $(0, g) \in \mathbb{H}_\gamma$ and for any $(\tau, g) \in \check{M}_a$

$$G_{\tau,g}\psi = \pi(\tau, h_g) \geq \mathcal{R} + \varepsilon,$$

which contradicts (76).

□

Proof of corollary 1.3.9. By proposition 1.3.1 we have that for $\tau_n \to \tau$ and $h_n \to h$,

$$\sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_0} \rightsquigarrow \mathcal{N}(\tilde{\mathcal{I}}_\gamma\tau, \tilde{\mathcal{I}}_\gamma), \text{ under } P_{\gamma_n,\tau_n,h_n}.$$

[94] The continuity of the indicated map follows directly from the fact that a Gaussian shift experiment is continuous in the total variation norm.

As in the proof of proposition B.3, equations (1.7), (1.8) and Lemma C.6 imply that $\|\hat{\mathcal{I}}^{\dagger}_{n,\theta_n} - \tilde{\mathcal{I}}^{\dagger}_{\gamma}\|_2 = o_{P_{\gamma_n}}(1)$. The mutual contiguity established in proposition 1.3.1 along with Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) ensures that this result and equation (1.6) also hold under P_{γ_n,τ_n,h_n}:

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_0} - \tilde{\ell}_{\gamma_n}\right] = o_{P_{\gamma_n,\tau_n,h_n}}(1) \quad \text{and} \quad \|\hat{\mathcal{I}}^{\dagger}_{n,\theta_0} - \tilde{\mathcal{I}}^{\dagger}_{\gamma}\|_2 = o_{P_{\gamma_n,\tau_n,h_n}}(1).$$

Write $\hat{Z}_n := \sqrt{n}\mathbb{P}_n\hat{\ell}_{n,\theta_0}$. We have

$$\hat{Z}_n = \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\gamma_n} + \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_0} - \tilde{\ell}_{\gamma_n}\right] \rightsquigarrow Z \sim \mathcal{N}(\tilde{\mathcal{I}}_{\gamma}\tau, \tilde{\mathcal{I}}_{\gamma})$$

under P_{γ_n,τ_n,h_n}. The continuous mapping theorem and Theorem 9.2.3 of Rao and Mitra (1971) imply that

$$\hat{S}_{n,\theta_0} = \hat{Z}_n'\hat{\mathcal{I}}^{\dagger}_{n,\theta_0}\hat{Z}_n \rightsquigarrow Z'\tilde{\mathcal{I}}^{\dagger}_{\gamma}Z =: S \sim \chi^2_r(\tau'\tilde{\mathcal{I}}_{\gamma}\tau),$$

under P_{γ_n,τ_n,h_n} where $r = \operatorname{rank}(\tilde{\mathcal{I}}_{\gamma})$.

Let c_n be the $1-\alpha$ quantile of the $\chi^2_{r_n}$ distribution and c the $1-\alpha$ quantile of the χ^2_r distribution. We have $P_{\gamma_n}\{c_n = c\} = P_{\gamma_n}\{r_n = r\} \to 1$ by assumption. This implies that $c_n - c \to 0$ in P_{γ_n}-probability and hence by the mutual contiguity and Le Cam's first lemma, also under P_{γ_n,τ_n,h_n}. By continuous mapping once more we have $\hat{S}_{n,\theta_0} - c_n \rightsquigarrow S - c$ under P_{γ_n,τ_n,h_n}. Hence by the portmanteau theorem

$$\lim_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\phi_{n,\theta_0} = 1 - \mathrm{P}\left(\chi^2_r\left(\tau\tilde{\mathcal{I}}_{\gamma}\tau\right) \le c\right). \tag{77}$$

For $\tau = 0$, $1 - \mathrm{P}\left(\chi^2_r(0) \le c\right) = \alpha$; hence this test is level-α as claimed.

Let $K_a \subset M_a$ be compact and suppose $(\tau_n, h_n)_{n\in\mathbb{N}} \subset K_a$ is such that $\tau_n \to \tau$ and $h_n \to h$. Then, by equation (77) we have that

$$\lim_{n\to\infty} P^n_{\gamma_n,\tau_n,h_n}\phi_{n,\theta_0} = \mathrm{P}(\chi^2_r\left(\tau'\tilde{\mathcal{I}}_{\gamma}\tau\right) > c) \ge \mathrm{P}(\chi^2_r(a) > c) =: \mathcal{R}. \tag{78}$$

Taking a constant sequence in K_a with $\tau'\tilde{\mathcal{I}}_{\gamma}\tau = a$ we obtain from the preceding display that $\limsup_{n\to\infty}\inf_{(\tau,h)\in K_a} P^n_{\gamma_n,\tau,h}\phi_{n,\theta_0} \le \lim_{n\to\infty} P^n_{\gamma_n,\tau,h}\phi_{n,\theta_0} = \mathcal{R}$. It follows that if equation (1.18) does not hold then there is a sequence $(\tau_n, h_n)_{n\in\mathbb{N}} \subset K_a$ and a subsequence $(n_j)_{j\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ such that

$$\mathcal{S} = \lim_{j\to\infty} P^{n_j}_{\gamma_{n_j},\tau_{n_j},h_{n_j}}\phi_{n_j,\theta_0} < \mathcal{R}. \tag{79}$$

Take a further subsequence $(n_k)_{k\in\mathbb{N}}$ along which $\tau_n \to \tau$ and $h_n \to h$ with $(\tau, h) \in K_a$. Construct new sequences $(h^*_m)_{m\in\mathbb{N}}$ and $(\tau^*_m)_{m\in\mathbb{N}}$ as follows. For all $m \in [n_k, n_{k+1}) \cap \mathbb{N}$ for some $k \in \mathbb{N}$ put $\tau^*_m = \tau_{n_k}$ and $h^*_m = h_{n_k}$. For $m = 1, \ldots, n_1$ put $\tau^*_m = \tau_{n_1}$ and

$h_m^* = h_{n_1}$. By construction we have that $\tau_m^* \to \tau$ and $h_m^* \to h$. By (78) we have that

$$\lim_{m\to\infty} P^m_{\gamma_m,\tau_m^*,h_m^*}\phi_{m,\theta_0} \geq \mathcal{R}.$$

Fix an arbitrary $\varepsilon > 0$. There is an $M \in \mathbb{N}$ such that for all $m \geq M$ we have $P^m_{\gamma_m,\tau_m^*,h_m^*}\phi_{m,\theta_0} \geq \mathcal{R} - \varepsilon/2$. Hence for any k sufficiently large that $m = n_k \geq M$ we have

$$\mathcal{S} = \mathcal{S} - P^{n_k}_{\gamma_{n_k},\tau_{n_k},h_{n_k}}\phi_{n_k,\theta_0} + P^m_{\gamma_m,\tau_m^*,h_m^*}\phi_{m,\theta_0} \geq \mathcal{S} - P^{n_k}_{\gamma_{n_k},\tau_{n_k},h_{n_k}}\phi_{n_k,\theta_0} + \mathcal{R} - \varepsilon/2.$$

This holds for all large enough k and so taking the limit first as $k \to \infty$ and then as $\varepsilon \downarrow 0$ yields that $\mathcal{S} \geq \mathcal{R}$. But this contradicts equation (79). □

Proof of proposition 1.3.10. By proposition A.8 in van der Vaart (1988b) and our assumptions

$$\begin{aligned}\Lambda_n(\gamma_n(\tau_n, h_n), \gamma_n) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n g_n - \frac{1}{2}P_{\gamma_n}g_n^2 + o_{P_{\gamma_n}}(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h\right] - \frac{1}{2}P_{\gamma_n}\left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h\right]^2 + o_{P_{\gamma_n}}(1).\end{aligned}$$

since $\frac{1}{2}P_{\gamma_n}\left[\tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h\right]^2 - \frac{1}{2}P_{\gamma_n}g_n^2 = \frac{1}{2}P_{\gamma_n}\left(f_n^2 - g_n^2\right) \to 0$, where $f_n := \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h$, as

$$\left|P_{\gamma_n}\left(f_n^2 - g_n^2\right)\right| = \left|\|f_n\|^2_{P_{\gamma_n},2} - \|g_n\|^2_{P_{\gamma_n},2}\right| \leq \|f_n - g_n\|^2_{P_{\gamma_n},2} + 2\|f_n - g_n\|_{P_{\gamma_n},2}\|g\|_{P_{\gamma_n},2} \to 0,$$

as $(g_n)_{n\in\mathbb{N}}$ is uniformly square P_{γ_n}-integrable and hence $P_{\gamma_n}g_n^2 \leq M$ for some $M \in (0, \infty)$.

It remains to show that $(f_n)_{n\in\mathbb{N}}$ is uniformly square P_{γ_n}-integrable. The preceding display yields that $P_{\gamma_n}f_n^2 = P_{\gamma_n}g_n^2 - P_{\gamma_n}(g_n^2 - f_n^2) = P_{\gamma_n}g_n^2 + o(1)$. Hence there is an $N \in \mathbb{N}$ such that $n > N$ has $P_{\gamma_n}f_n^2 \leq M + 1$. It follows that $P_{\gamma_n}f_n^2 \leq K < \infty$ with $K := \max\{M + 1, P_{\gamma_1}f_1^2, \ldots, P_{\gamma_N}f_N^2\}$. Let $\varepsilon > 0$ be given and note that there is a $\delta > 0$ such that if $P_{\gamma_n}(A) < \delta$ we have $P_{\gamma_n}(g_n^2\mathbf{1}_A) < \varepsilon/4$.[95] Hence

$$P_{\gamma_n}\left(f_n^2\mathbf{1}_A\right) \leq 2P_{\gamma_n}\left((f_n - g_n)^2\mathbf{1}_A\right) + 2P_{\gamma_n}\left(g_n^2\mathbf{1}_A\right) = o(1) + \frac{\varepsilon}{2}.$$

Hence there is an $N' \in \mathbb{N}$ such that for all $n \geq N'$ we have $P_{\gamma_n}(f_n^2\mathbf{1}_A) < \varepsilon$ if $P_{\gamma_n}(A) < \delta$. By Markov's inequality we have that for $K' > K/\delta$, $P_{\gamma_n}(f_n^2 > K') \leq P_{\gamma_n}f_n^2/K' \leq \delta$ and

[95] Given $\varepsilon > 0$, take $M < \infty$ large enough that $P_n(g_n^2\mathbf{1}\{g_n^2 > M\}) < \varepsilon/8$ for all $n \in \mathbb{N}$ and let $\delta < \varepsilon/(8M)$. Then if $P_{\gamma_n}(A) < \delta$ we have

$$P_\gamma(g_n^2\mathbf{1}_A) \leq P_{\gamma_n}(g_n^2\mathbf{1}_A\mathbf{1}\{g_n^2 \leq M\}) + P_{\gamma_n}(g_n^2\mathbf{1}_A\mathbf{1}\{g_n^2 > M\}) \leq MP_{\gamma_n}(A) + P_n(g_n^2\mathbf{1}\{g_n^2 > M\}) < \varepsilon/4.$$

hence for all $n \geq N'$, $P_{\gamma_n}(f_n^2 \mathbf{1}\{f_n^2 > K'\}) < \varepsilon$. That is, $(f_n)_{n\in\mathbb{N}}$ is asymptotically uniformly square P_{γ_n}-integrable, which implies that $(f_n)_{n\in\mathbb{N}}$ is uniformly square P_{γ_n}-integrable.[96] □

Proof of lemma 1.3.11. This is implied by Corollary 2.9 of Feinberg et al. (2016). □

Proof of lemma 1.3.12. Define $Q_n,\ Q$ respectively as the pushforward measures of P_n under f_n and P under f. By the extended continuous mapping theorem of van der Vaart and Wellner (1996, Theorem 1.11.1), $Q_n \rightsquigarrow Q$ and by hypothesis,

$$\lim_{M\to\infty} \sup_{n\in\mathbb{N}} \int_{|x|>M} |x| \,\mathrm{d}Q_n(x) = \lim_{M\to\infty} \sup_{n\in\mathbb{N}} \int_{|f_n(s)|>M} |f(s)| \,\mathrm{d}P_n(s) = 0.$$

The result now follows from the equivalence of (ii) and (iii) in Proposition A.6.1 of Bickel et al. (1998). □

Proof of proposition 1.3.13. Throughout let $\hat{r}_n := \operatorname{rank}(\hat{M}_n)$, $r := \operatorname{rank}(M)$, $R_n := \{\hat{r}_n = r\}$ and $\lambda_l, \lambda_{n,l}, \check{\lambda}_{n,l}$ and $\hat{\lambda}_{n,l}$ respectively the l-th largest eigenvalue of M, M_n, $\check{M}_n$ and $\hat{M}_n$.

Start with the case $r = 0$. By Weyl's perturbation theorem and the fact that $M_n = 0$ for all n larger than some $N \in \mathbb{N}$,

$$P_n(R_n) = P_n\left(\max_{l=1,\dots,L} |\check{\lambda}_{n,l}| < \nu_n\right) \geq P_n(\|\check{M}_n - M_n\|_2 < \nu_n) \to 1.$$

On the sets R_n we have that $\hat{M}_n = 0 = M$ and so $\hat{M}_n \xrightarrow{P_n} M$ as $P(R_n) \to 1$.

Now suppose that $r > 0$. let $\underline{\nu} := \lambda_r/2 > 0$ and note that (1.20) implies that $\|\check{M}_n - M_n\|_2 = o_{P_n}(1)$ and so, by Weyl's perturbation theorem (e.g. Bhatia, 1997, Corollary III.2.6), $\max_{l=1,\dots,L} |\check{\lambda}_{n,l} - \lambda_{n,l}| \leq \|\check{M}_n - M_n\|_2 = o_{P_n}(1)$. Hence, defining $E_n := \{\check{\lambda}_{n,r} \geq \nu_n\}$, for n large enough such that $\nu_n < \underline{\nu}$ and $\|M_n - M\|_2 < \underline{\nu}/2$ we have

$$P_n(E_n) = P_n\left(\check{\lambda}_{n,r} \geq \nu_n\right) \geq P_n\left(\check{\lambda}_{n,r} \geq \underline{\nu}\right) \geq P_n\left(|\check{\lambda}_{n,r} - \lambda_{n,r}| < \underline{\nu}/2\right) \to 1.$$

If $r = L$ we have that $R_n \supset E_n$ and therefore $P_n(R_n) \to 1$. Additionally, if $\check{\lambda}_{n,L} \geq \nu_n$ then $\hat{\lambda}_{n,l} = \check{\lambda}_{n,l}$ for each $l \in [L]$ and hence $\hat{M}_n = \check{M}_n$, implying $\|\hat{M}_n - M\|_2 \leq \|\check{M}_n - M_n\|_2 + \|M_n - M\|_2 = o_{P_n}(1)$.

Now suppose instead that $r < L$ and define $F_n := \{\check{\lambda}_{n,r+1} < \nu_n\}$. It follows by Weyl's perturbation theorem and the fact that $\lambda_{n,l} = 0$ for $l > r$ and $n \geq N$ that as $n \to \infty$

$$P_n(F_n) = P_n(\check{\lambda}_{n,r+1} < \nu_n) \geq P_n(\|\check{M}_n - M_n\|_2 < \nu_n) \to 1.$$

[96] Increase K' to K'' as necessary to ensure that also $P_{\gamma_n}(f_n^2 \mathbf{1}\{f_n^2 > K''\}) < \varepsilon$ for all $1 \leq n < N'$.

Since $R_n \supset E_n \cap F_n$, this implies that $P_n(R_n) \to 1$ as $n \to \infty$. Additionally, if $\check{\lambda}_{n,r} \geq \nu_n$, $\check{\lambda}_{n,r+1} < \nu_n$ and $\|\check{M}_n - M\|_2 \leq \upsilon$, we have that $\hat{\lambda}_{n,k} = \check{\lambda}_{n,k}$ for $k \leq r$ and $\hat{\lambda}_{n,l} = 0 = \lambda_l$ for $l > r$ and so

$$\|\Lambda_n(\nu_n) - \Lambda\|_2 = \max_{l=1,\ldots,r} |\hat{\lambda}_{n,l} - \lambda_l| = \max_{l=1,\ldots,r} |\check{\lambda}_{n,l} - \lambda_l| \leq \|\check{\Lambda}_n - \Lambda\|_2 \leq \|\check{M}_n - M\|_2 \leq \upsilon,$$

and hence $\{\|\check{M}_n - M\|_2 \leq \upsilon\} \cap E_n \cap F_n \subset \{\|\Lambda_n(\nu_n) - \Lambda\|_2 \leq \upsilon\}$, from which it follows that $\Lambda_n(\nu_n) \xrightarrow{P_n} \Lambda$ as $\|\check{M}_n - M\|_2 \leq \|\check{M}_n - M_n\|_2 + \|M_n - M\|_2 \xrightarrow{P_n} 0$. Suppose that $(\lambda_1, \ldots, \lambda_r)$ consists of s distinct eigenvalues with values $\lambda^1 > \lambda^2 > \cdots > \lambda^s$ and multiplicities $\mathfrak{m}_1, \ldots, \mathfrak{m}_s$ (each at least one).[97] $\lambda^{s+1} = 0$ is an eigenvalue with multiplicity $\mathfrak{m}_{s+1} = L - r$. Let l_i^k for $k = 1, \ldots, s+1$ and $i = 1, \ldots, \mathfrak{m}_k$ denote the column indices of the eigenvectors in U corresponding to each λ^k. For each λ^k, the total eigenprojection is $\Pi_k := \sum_{i=1}^{\mathfrak{m}_k} u_{l_i^k} u'_{l_i^k}$.[98] Total eigenprojections are continuous.[99] Therefore, if we construct $\Pi_{n,k}$ in in an analogous fashion to Π_k but replace columns of U with columns of $\check{U}_n$, we have $\Pi_{n,k} \xrightarrow{P_n} \Pi_k$ for each $k \in [s+1]$ since $\check{M}_n \xrightarrow{P_n} M$. Spectrally decompose M as $M = \sum_{k=1}^s \lambda^k \Pi_k$, where the sum runs to s rather than $s+1$ since $\lambda^{s+1} = 0$. Then,

$$\hat{M}_n = \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} \hat{\lambda}_{n,l_i^k} u_{n,l_i^k} u'_{n,l_i^k} = \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} (\hat{\lambda}_{n,l_i^k} - \lambda^k) u_{n,l_i^k} u'_{n,l_i^k} + \sum_{k=1}^{s} \lambda^k \Pi_{n,k},$$

whence

$$\|\hat{M}_n - M\|_2 \leq \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} |\hat{\lambda}_{n,l_i^k} - \lambda^k| \|u_{n,l_i^k} u'_{n,l_i^k}\|_2 + \sum_{k=1}^{s} |\lambda^k| \|\Pi_{n,k} - \Pi_k\|_2 \xrightarrow{P_n} 0,$$

by $\hat{\Pi}_{n,k} \xrightarrow{P_n} \Pi_k$, $\hat{\Lambda}_n(\nu_n) \xrightarrow{P_n} \Lambda$ and since we have $\|u_{n,l_i^k} u'_{n,l_i^k}\|_2 = 1$ for any i, k, n. □

Proof of corollary 1.3.14. Apply proposition 1.3.13 with $\check{\mathcal{I}}_{n,\theta_n} = \check{M}_n$, $\hat{\mathcal{I}}_{n,\theta_n} = \hat{M}_n$, $\tilde{\mathcal{I}}_n = M_n$, $\tilde{\mathcal{I}}_\gamma = M$ and $P_{\gamma_n} = P_n$. □

B.2. Additional miscellaneous results

Lemma B.7. *Suppose that assumption M holds and assumptions LAN and CM(i) hold along a convergent sequence $(\gamma_n)_{n \in \mathbb{N}}$ with $\gamma_n := (\theta_n, \eta) \to \gamma \in \Gamma$,that $\eta = (\eta_1, \eta_2)$ with $\eta_1 \in \mathcal{H}_1 \subset \mathbb{R}^{d_{\eta_1}}$ and that the efficient score function takes the form*

$$\tilde{\ell}_{\gamma_n} = \check{\ell}_{\gamma_n,1} - \check{I}_{\gamma_n,12} \check{I}_{\gamma_n,22}^{-1} \check{\ell}_{\gamma_n,2}, \quad \check{I}_{\gamma_n} := P_{\gamma_n} \check{\ell}_{\gamma_n} \check{\ell}'_{\gamma_n},$$

[97] The superscripts on the λs are indices, not exponents.

[98] See e.g Chapter 8.8 of Magnus and Neudecker (2019).

[99] E.g. Theorem 8.7 of Magnus and Neudecker (2019).

for a L-dimensional vector of functions $\breve{\ell}_{\gamma_n} := \left(\breve{\ell}'_{\gamma_n,1}, \breve{\ell}'_{\gamma_n,2}\right)'$. *Suppose that* $\tilde{\mathcal{I}}_{\gamma_n} \to \tilde{\mathcal{I}}_\gamma$ *and* $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ *for all sufficiently large* $n \in \mathbb{N}$. *Moreover, suppose that along any sequence* $(\gamma'_n)_{n\in\mathbb{N}}$ *with* $\gamma'_n := (\theta_n, (\eta_{n,1}, \eta_2)) \to \gamma$ *where* $\sqrt{n}\|\eta_{n,1} - \eta_1\| = O(1)$*,*

1. $P_{\gamma'_n}\breve{\ell}_{\gamma'_n} = o(n^{-1/2})$*,*
2. $(\|\breve{\ell}_{\gamma_n}\|_2^2)_{n\in\mathbb{N}}$ *is uniformly* $P_{\gamma'_n}$*-integrable,*
3. $\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\xi_n} - \breve{\ell}_{\gamma'_n}\right] = o_{P_{\gamma'_n}}(1)$*,*
4. $\nu_n^{-1}\|\hat{I}_{n,\xi_n} - \breve{I}_{\gamma_n}\|_2 = o_{P_{\gamma'_n}}(1)$*,*
5. $\int\left[\breve{\ell}_{\gamma'_n,l}\sqrt{p_{\gamma'_n}} - \breve{\ell}_{\gamma_n,l}\sqrt{p_{\gamma_n}}\right]^2 \mathrm{d}\nu \to 0$ *for each* $l \in [L]$*,*

with $\xi_n := (\theta_n, \eta_{n,1})$. *Finally suppose that* $\hat{\eta}_{n,1}$ *satisfies* $\sqrt{n}\|\hat{\eta}_{n,1} - \eta_1\| = O_{P_{\gamma_n}}(1)$. *Then if* $\bar{\xi}_n := (\theta_n, \bar{\eta}_{n,1})$ *where* $\bar{\eta}_{n,1}$ *is the version of* $\hat{\eta}_{n,1}$ *discretised on* $n^{-1/2}C\mathbb{Z}^{d_{\eta_1}} \cap \mathcal{H}_1$*,*

$$\hat{\ell}_{n,\theta_n} := \hat{\ell}_{n,\bar{\xi}_n,1} - \hat{I}_{n,\bar{\xi}_n,12}\hat{I}^{-1}_{n,\bar{\xi}_n,22}\hat{\ell}_{n,\bar{\xi}_n,2}, \quad \breve{\mathcal{I}}_{n,\theta_n} := \hat{I}_{n,\bar{\xi}_n,11} - \hat{I}_{n,\bar{\xi}_n,12}\hat{I}^{-1}_{n,\bar{\xi}_n,22}\hat{I}_{n,\bar{\xi}_n,21}, \tag{80}$$

and $\hat{\mathcal{I}}_{n,\theta_n}$ *is the eigendecomposition-truncated version of* $\breve{\mathcal{I}}_{n,\theta_n}$ *at* ν_n *analogously to* (1.21)*, then assumptions E and R hold.*

Proof. Define $b_n := \sqrt{n}(\eta_{n,1} - \eta_1)$. Take an arbitrary subsequence $(n_m)_{m\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ and a further subsequence $(n_k)_{k\in\mathbb{N}}$ along which $b_{n_k} \to b \in \mathbb{R}^{d_{\eta_1}}$. Construct a "full" sequence $(b^\star_n)_{n\in\mathbb{N}}$ according to $b^\star_{n_k} := b_{n_k}$ for all $k \in \mathbb{N}$ and for all $m \in \mathbb{N}$ such that $m \notin \{n_k : k \in \mathbb{N}\}$ set $b^\star_m := b^\star_{m-1}$ (arbitrarily put $b_0 = 0$). Constructed in this manner $b^\star_n \to b$ as $n \to \infty$ and hence $\beta^\star_{n,1} := \eta + \sqrt{n}b^\star_n$ is a deterministic sequence satisfying $\sqrt{n}(\eta^\star_{n,1} - \eta) = O(1)$. Note that we can write $\gamma^\star_n := (\theta_n, (\eta^\star_{n,1}, \eta_2))$ as $\gamma^\star_n = \gamma_n(0, h^\star_n)$ for $h^\star_n := (b^\star_n, 0)$. Since conditions 1 - 5 are valid along $(\gamma'_n)_{n\in\mathbb{N}}$ formed with an arbitrary deterministic $\sqrt{n}$-consistent sequence $(\eta_{n,1})_{n\in\mathbb{N}}$, they apply along $(\gamma^\star_n)_{n\in\mathbb{N}}$ in particular. Since LAN holds, these observations, in conjunction with Proposition A.10 in van der Vaart (1988b) yield that

$$\sqrt{n}\mathbb{P}_n\left[\breve{\ell}_{\gamma^\star_n} - \breve{\ell}_{\gamma_n}\right] + \breve{I}_{\gamma_n}(0', (b^\star_n)')' = o_{P_{\gamma_n}}(1).$$

This clearly implies also that

$$\sqrt{n_k}\mathbb{P}_{n_k}\left[\breve{\ell}_{\gamma'_{n_k}} - \breve{\ell}_{\gamma_{n_k}}\right] + \breve{I}_{\gamma_{n_k}}(0', b'_{n_k})' = \sqrt{n_k}\mathbb{P}_{n_k}\left[\breve{\ell}_{\gamma^\star_{n_k}} - \breve{\ell}_{\gamma_{n_k}}\right] + \breve{I}_{\gamma_{n_k}}(0', (b^\star_{n_k})')' = o_{P_{\gamma_{n_k}}}(1),$$

and therefore, as the original subsequence $(n_m)_{m\in\mathbb{N}}$ was arbitrary,

$$\sqrt{n}\mathbb{P}_n\left[\breve{\ell}_{\gamma'_n} - \breve{\ell}_{\gamma_n}\right] + \sqrt{n}\breve{I}_{\gamma_n}(0', (\eta_{n,1} - \eta)')' = o_{P_{\gamma_n}}(1). \tag{81}$$

Moreover we have by Proposition 1.3.1 that $(P^n_{\gamma_n})_{n\in\mathbb{N}}$ and $(P^n_{\gamma^\star_n})_{n\in\mathbb{N}}$ are mutually contiguous. Hence the same is true of $(P^{n_k}_{\gamma_{n_k}})_{k\in\mathbb{N}}$ and $(P^{n_k}_{\gamma^\star_{n_k}})_{k\in\mathbb{N}} = (P^{n_k}_{\gamma'_{n_k}})_{k\in\mathbb{N}}$. This observation in conjunction with 3, 4 and the fact that our initial subsequence $(n_m)_{m\in\mathbb{N}}$ was arbitrary yields the conclusion that

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\xi_n} - \breve{\ell}_{\gamma'_n}\right] = o_{P_{\gamma_n}}(1), \quad \text{and} \quad \left\|\hat{I}_{n,\xi_n} - \breve{I}_{\gamma_n}\right\|_2 = o_{P_{\gamma_n}}(\nu_n). \tag{82}$$

Now, for $\eta_1^\sharp \in \mathcal{H}_1$ let

$$R_{1,n}(\eta_1^\sharp) := \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\xi_n^\sharp} - \breve{\ell}_{\gamma_n}\right] + \sqrt{n}\breve{I}_{\gamma_n}(0', (\eta_1^\sharp - \eta)'), \quad R_{2,n}(\eta_1^\sharp) := \nu_n^{-1}\left[\hat{I}_{n,\xi_n^\sharp} - \breve{I}_{\gamma_n}\right]$$

where $\xi_n^\sharp := (\theta_n, \eta_1^\sharp)$, $\eta^\sharp := (\eta_1^\sharp, \eta_2)$ and $\gamma_n^\sharp := (\theta_n, \eta^\sharp)$. Let As $\bar{\beta}_n$ is discretised on $n^{-1/2}C\mathbb{Z}^{d_{\eta_1}} \cap \mathcal{H}_1$ from $\hat{\eta}_{1,n}$ it remains $\sqrt{n}$-consistent under P_{γ_n} and hence for any $\varepsilon > 0$ there is an $M \in (0,\infty)$ and N such that for all $n \geq N$, $P_{\gamma_n}\left(\sqrt{n}\|\bar{\eta}_{n,1} - \eta_1\|_2 > M\right) < \varepsilon$. If $\sqrt{n}\|\bar{\eta}_{n,1} - \eta_1\|_2 \leq M$ then $\bar{\eta}_{n,1} \in \mathfrak{S}_n := \{\eta_1^\flat \in n^{-1/2}C\mathbb{Z}^{d_{\eta_1}} \cap \mathcal{H}_1 : \|\eta_1^\flat - \eta_1\|_2 \leq M/\sqrt{n}\}$. For any fixed M, $\mathfrak{S}_n$ has a finite number of elements bounded independently of n, call this number $\overline{\mathfrak{S}}$. For $R_n \in \{R_{1,n}, R_{2,n}\}$, any $\upsilon > 0$ and $n \geq N$

$$\begin{aligned} P_{\gamma_n}\left(\|R_n(\bar{\eta}_{n,1})\| > \upsilon\right) &\leq \varepsilon + \sum_{\eta_{n,1}\in\mathfrak{S}_n} P_{\gamma_n}\left(\{\|R_n(\eta_{n,1})\| > \upsilon\} \cap \{\bar{\eta}_{n,1} = \eta_{n,1}\}\right) \\ &\leq \varepsilon + \overline{\mathfrak{S}}P_{\gamma_n}\left(\|R_n(\eta^*_{n,1})\| > \upsilon\right), \end{aligned}$$

where $\eta^*_{n,1} \in \mathfrak{S}_n$ maximises $\eta_1 \mapsto P_{\gamma_n}\left(\|R_n(\eta_1)\| > \upsilon\right)$. Since $(\eta^*_{n,1})_{n\in\mathbb{N}}$ is deterministic and $\sqrt{n}$-consistent for η_1, $P_{\gamma_n}\left(\|R_n(\eta^*_{n,1})\| > \upsilon\right) \to 0$ by equations (81) & (82). It follows that $\|R_{i,n}(\bar{\eta}_{n,1})\| = o_{P_{\gamma_n}}(1)$ for $i \in \{1,2\}$. It follows that $\|\hat{\mathcal{K}}_{\bar{\xi}_n} - \tilde{\mathcal{K}}_{\gamma_n}\|_2 \xrightarrow{P_{\gamma_n}} 0$ where

$$\tilde{\mathcal{K}}_{\gamma_n} := \left[I \ - \breve{I}_{\gamma_n,12}\breve{I}^{-1}_{\gamma_n,22}\right], \quad \hat{\mathcal{K}}_{\bar{\xi}_n} := \left[I \ - \hat{I}_{n,\bar{\xi}_n,12}\hat{I}^{-1}_{n,\bar{\xi}_n,22}\right],$$

with the partitions of the matrices $\hat{I}_{\bar{\xi}_n}$, $\breve{I}_{\gamma_n}$ corresponds to the partition of the vectors $\hat{\ell}_{n,\bar{\xi}_n} = (\hat{\ell}'_{n,\bar{\xi}_n,1}, \hat{\ell}'_{n,\bar{\xi}_n,2})'$, $\breve{\ell}_{\gamma_n} = (\tilde{\ell}'_{\gamma_n,1}, \tilde{\ell}'_{\gamma_n,2})'$, $\bar{\xi}_n := (\theta_n, \bar{\eta}_{n,1})$ and $\breve{I}^{-1}_{\gamma_n,22}$ exists by assumption. Using these results, (80) and the uniform P_{γ_n}-integrability of $\|\breve{\ell}_{\gamma_n}\|_2^2$,

$$\begin{aligned} &\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\right] \\ &\quad = \left(\hat{\mathcal{K}}_{\bar{\xi}_n} - \tilde{\mathcal{K}}_{\gamma_n}\right)\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\bar{\xi}_n} - \breve{\ell}_{\gamma_n}\right] + \tilde{\mathcal{K}}_{\gamma_n}\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{n,\bar{\xi}_n} - \breve{\ell}_{\gamma_n}\right] + \left(\hat{\mathcal{K}}_{\bar{\xi}_n} - \tilde{\mathcal{K}}_{\gamma_n}\right)\sqrt{n}\mathbb{P}_n\breve{\ell}_{\gamma_n} \\ &\quad = -\left[I \ - \breve{I}_{\gamma_n,12}\breve{I}^{-1}_{\gamma_n,22}\right]\begin{bmatrix}\breve{I}_{\gamma_n,11} & \breve{I}_{\gamma_n,12} \\ \breve{I}_{\gamma_n,21} & \breve{I}_{\gamma_n,22}\end{bmatrix}\begin{bmatrix}0 \\ \sqrt{n}(\bar{\eta}_{n,1} - \eta_1)\end{bmatrix} + o_{P_{\gamma_n}}(1) \\ &\quad = o_{P_{\gamma_n}}(1), \end{aligned}$$

which gives (1.6). To show that equation (1.7) and assumption R hold, Corollary 1.3.14 indicates that it suffices to show that the requirements of assumption T are satisifed. For this note that by assumption $\tilde{\mathcal{I}}_{\gamma_n} \to \tilde{\mathcal{I}}_\gamma$ with $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ for all sufficiently large $n \in \mathbb{N}$ and (1.23) follows from $\|R_{2,n}(\bar{\eta}_{n,1})\| = o_{P_{\gamma_n}}(1)$. □

B.3. Proofs for section 1.4

Throughout this section I use the notation $\iota(\theta, X) := X_1 + X_2'\theta$.

Proof of Proposition 1.4.1. Fix arbitrary $\tau_n \to \tau \in \mathbb{R}^{d_\theta}$ and $h_n \to h \in H_\eta$. The perturbed law is P_{γ_n,τ_n,h_n} with density

$$p_{\gamma_n,\tau_n,h_n}(W) := \zeta(e_n, X)(1 + h_{n,2}(e_n, X)/\sqrt{n}),$$

where $e_n := Y - f(\iota(\theta_n + n^{-1/2}\tau_n, X)) - n^{-1/2}h_{n,1}(\iota(\theta_n + n^{-1/2}\tau_n, X))$. Since Θ is are open and $\theta_n \to \theta$, $\theta_n + n^{-1/2}\tau_n \in \Theta$ for all large enough $n \in \mathbb{N}$. The restrictions on $\dot{\mathscr{F}}$ ensure that $f + n^{-1/2}h_{n,1} \in \mathscr{F}$. The restrictions on $\dot{\mathscr{Z}}_\eta$ along with the norm on H suffice to ensure that $\zeta(1 + h_{n,2}/\sqrt{n}) \in \mathscr{Z}$. Specifically, for all large enough n, $\zeta(1+h_{n,2}/\sqrt{n}) \geq 0$ (λ-a.e.) since $h_{n,2}$ is bounded (λ-a.e.) and the conditions on $\dot{\mathscr{Z}}$ ensure that $\int \zeta(1 + h_{n,2}/\sqrt{n})\,\mathrm{d}\lambda = \int \zeta\,\mathrm{d}\lambda + \frac{1}{\sqrt{n}}\int h_{n,2}\zeta\,\mathrm{d}\lambda = 1$. Continuous differentiability (λ-a.e.) of $e \mapsto \sqrt{\zeta(1 + h_{n,2}/\sqrt{n})(e, X)}$ follows from the same requirement on $\sqrt{\zeta}$ and $h_{n,2}$, the boundedness of $h_{n,2}$ (which ensures that eventually $1 + h_{n,2}/\sqrt{n}$ is bounded away from zero λ-a.e.) and the chain rule. Finally it remains to check the conditions in (1.26). For any $A \in \sigma(Z)$, letting G denote the measure corresponding to ζ

$$\begin{aligned}\int_A \epsilon\zeta(\epsilon, X)(1 + h_{n,2}(\epsilon, X)/\sqrt{n})\,\mathrm{d}\lambda &= \int_A \epsilon\,\mathrm{d}G + \frac{1}{\sqrt{n}}\int_A \epsilon h_{n,2}(\epsilon, X)\,\mathrm{d}G\\ &= \int_A \mathbb{E}[\epsilon|X]\,\mathrm{d}G + \frac{1}{\sqrt{n}}\int_A \mathbb{E}[\epsilon h_{n,2}(\epsilon, X)|X]\,\mathrm{d}G\\ &= 0,\end{aligned}$$

and hence $\mathbb{E}[\epsilon|X] = 0$ (a.s. under $\zeta(1 + h_{n,2}/\sqrt{n})$). For the rest, firstly let $m(\epsilon, X)$ be non-negative and integrable under G. By the (λ-a.e.) boundedness of $h_{n,2}$ (by $\bar{h}_2$, say)

$$\int m(\epsilon, X)\zeta(\epsilon, X)(1 + h_{n,2}(\epsilon, X)/\sqrt{n})\,\mathrm{d}\lambda \leq \left(1 + \frac{\bar{h}_2}{\sqrt{n}}\right)\int m(\epsilon, X)\,\mathrm{d}G < \infty.$$

Secondly, note that by Jensen's inequality

$$\left\|\int XX'\zeta(1 + h_{n,2}/\sqrt{n})\,\mathrm{d}\lambda - \int XX'\,\mathrm{d}G\right\|_2 \leq \frac{\bar{h}_2}{\sqrt{n}}\left\|\int XX'\,\mathrm{d}G\right\| \leq \frac{\bar{h}_2}{\sqrt{n}}\int \|X\|_2^2\,\mathrm{d}G \to 0,$$

which implies that for all large enough n, $\int XX'\zeta(1 + h_{n,2}/\sqrt{n})\,\mathrm{d}\lambda \succ 0$.

To establish (1.19), first let $\gamma \in \Gamma$, $u = (\tau, h) \in \mathbb{R}^{d_\theta} \times H_\eta$, $t \in (0, \infty)$ and $\varphi := \varphi(u) := (\tau, h_1, \zeta h_2)$ and let $\Delta_\gamma(\varphi) := \frac{1}{2}[\tau\dot{\ell}_\gamma + B_\gamma h]\sqrt{p_\gamma}$. By arguing analogously to the preceding paragraph it is seen that for all t in a sufficiently small neighbourhood $\mathscr{U}$ of 0 in $[0, \infty)$, $p_{\gamma+t\varphi}$ is a probability density. $t \mapsto \sqrt{p_{\gamma+t\varphi}}$ is continuously differentiable λ-a.e. by the corresponding conditions imposed on $e \mapsto \sqrt{\zeta(e, X)}$ and $e \mapsto h_3(e, X)$. For $t \in \mathscr{U}$, define $e(t) = Y - f(\iota(\theta(t), X)) - th_1(\iota(\theta(t), X))$ with $\theta(t) := \theta + t\tau$. Define

$g(t) := \frac{\partial}{\partial s}_{|s=t} \log p_{\gamma + s\varphi}$ and note

$$g(t) = -\phi(e(t), X)\left[f'(\iota(\theta(t), X))X_2'\tau + h_1(\iota(\theta(t), X)) + th_1'(\iota(\theta(t), X))X_2'\tau\right] \\ + \frac{h_2(e(t), X) + th_2'(e(t), X)\left[f'(\iota(\theta(t), X))X_2'\tau + h_1(\iota(\theta(t), X)) + th_1'(\iota(\theta(t), X))X_2'\tau\right]}{1 + th_2(e(t), X)}.$$

By taking $\mathscr{U}$ smaller if necessary suppose that $1 + th_2 > c > 0$, and $|f'|$, $|h_1|$, $|h_1'|$, $|h_2|$ and $|h_2'|$ are bounded by $C \in (0, \infty)$ λ-a.e.. Let $t_n \to t$ through $\mathscr{U}$ and note that $g(t_n) \to g(t)$ λ-a.e. by the continuity and continuous differentiability assumptions. For any $t \in \mathscr{U}$

$$\int |g(t)|^{2+\rho} \,\mathrm{d}P_{\gamma + t\varphi} \lesssim \int (\phi(\epsilon, X)^{2+\rho} + 1)\|X\|_2^{2+\rho}\zeta(\epsilon, X)\,\mathrm{d}\lambda < \infty,$$

which can be used in conjunction with Markov's inequality to obtain the uniform $P_{\gamma + t_n\varphi}$-integrability of $(g(t_n)^2)_{n\in\mathbb{N}}$. Since also $p_{\gamma + t_n\varphi} \to p_{\gamma + t\varphi}$ λ-a.e. as is easily verified by inspection, Lemma 1.3.11 implies that $\int g(t_n)^2 \,\mathrm{d}P_{\gamma + t_n\varphi} \to \int g(t)^2 \,\mathrm{d}P_{\gamma + t\varphi}$. By Lemma 1.8 in van der Vaart (2002)

$$\lim_{t\downarrow 0} \left\| \frac{\sqrt{p_{\gamma + t\varphi}} - \sqrt{p_\gamma}}{t} - \Delta_\gamma(\varphi) \right\|_{\lambda,2} = 0. \tag{83}$$

Next let $(\delta_n)_{n\in\mathbb{N}} \subset [0, 1]$ be an arbitrary sequence, $t_n \downarrow 0$ and define $\upgamma_n := \gamma_n + \delta_n t_n \varphi_n$ for $\varphi_n := \varphi(u_n)$ with $u_n \to u \in \mathbb{R}^{d_\theta} \times H_\eta$. Define $\tilde{e}_n := Y - f(\iota(\tilde{\theta}_n, X)) - \delta_n t_n h_{n,1}(\iota(\tilde{\theta}_n, X))$ with $\tilde{\theta}_n := \theta_n + \delta_n t_n \tau_n$,

$$\phi_n := \phi(\tilde{e}_n, X) + \frac{\delta_n t_n h_{n,2}'(\tilde{e}_n, X)}{1 + \delta_n t_n h_{n,2}(\tilde{e}_n, X)}.$$

Then, $\Delta_{\upgamma_n}(\varphi_n) := \frac{1}{2}[\tau_n' \dot{\ell}_{\upgamma_n} + B_{\upgamma_n} h_n]\sqrt{p_{\upgamma_n}}$, with

$$\begin{aligned} p_{\upgamma_n}(W) &= \zeta(\tilde{e}_n, X)(1 + \delta_n t_n h_{n,2}(\tilde{e}_n, X)) \\ \dot{\ell}_{\upgamma_n}(W) &= -\phi_n f'(\iota(\tilde{\theta}_n, X))X_2 \\ [B_{\upgamma_n} h](W) &= -\phi_n h_{n,1}(\iota(\tilde{\theta}_n, X)) + h_{n,2}(\tilde{e}_n, X). \end{aligned}$$

It may be verified by inspection that $\Delta_{\upgamma_n}(\varphi_n) \to \Delta_\gamma(\varphi)$ λ-a.e. under our assumptions. Argue analogously to the demonstration that $\int g(t_n)^2 \,\mathrm{d}P_{\gamma + t_n\varphi} \to \int g(t)^2 \,\mathrm{d}P_{\gamma + t\varphi}$ above to conclude $\|\Delta_{\upgamma_n}(\varphi_n)\|_{\lambda,2}^2 \to \|\Delta_\gamma(\varphi)\|_{\lambda,2}^2$ and hence by Proposition 2.29 in van der Vaart (1998),

$$\|\Delta_{\upgamma_n}(\varphi_n) - \Delta_\gamma(\varphi)\|_{\lambda,2} \to 0. \tag{84}$$

Now we establish (1.19). First suppose that $\theta_n = \theta$ for all $n \in \mathbb{N}$, let $u_n \to u$ be arbitrary, put $\varphi_n := \varphi(u_n)$, $\varphi := \varphi(u)$ and $t_n \downarrow 0$. For all large enough n, $\gamma + t_n\varphi_n \in \Gamma$ and so using (83) and the mean-value theorem (e.g. Drabek and Milota, 2007, Theorem 3.2.7), for such

n

$$\begin{aligned}\left\|\frac{\sqrt{p_{\gamma+t_n\varphi_n}}-\sqrt{p_\gamma}}{t_n}-\Delta_\gamma(\varphi)\right\|_{\lambda,2} &\leq \left\|\frac{\sqrt{p_{\gamma+t_n\varphi_n}}-\sqrt{p_{\gamma+t_n\varphi}}}{t_n}\right\|_{\lambda,2}+\left\|\frac{\sqrt{p_{\gamma+t_n\varphi}}-\sqrt{p_\gamma}}{t_n}-\Delta_\gamma(\varphi)\right\|_{\lambda,2}\\ &\leq \sup_{\delta\in[0,1]}\left\|\Delta_{\gamma+\delta t_n(\varphi_n-\varphi)(\varphi_n-\varphi)}\right\|_{\lambda,2}+o(1)\\ &= o(1),\end{aligned} \tag{85}$$

where the last step uses that for any sequence $(\delta_n)_{n\in\mathbb{N}}\subset[0,1]$, $\|\Delta_{\gamma+\delta_n t_n(\varphi_n-\varphi)}(\varphi_n-\varphi)-\Delta_\gamma(0)\|_{\lambda,2}\to 0$ by (84) and $\Delta_\gamma(0)=0$. Now consider an arbitrary sequence $\theta_n\to\theta$ and $\gamma_n=(\theta_n,\eta)$. Using (85) and applying the mean-value theorem at each $n\in\mathbb{N}$ gives

$$\begin{aligned}\left\|\frac{\sqrt{p_{\gamma_n+t_n\varphi_n}}-\sqrt{p_{\gamma_n}}}{t_n}-\Delta_{\gamma_n}(\varphi)\right\|_{\lambda,2} &\leq |t_n^{-1}|\sup_{\delta\in[0,1]}\|\Delta_{\gamma_n+\delta_n t_n\varphi_n}(t_n\varphi_n)-t_n\Delta_{\gamma_n}(\varphi)\|_{\lambda,2}\\ &= \sup_{\delta\in[0,1]}\|\Delta_{\gamma_n+\delta_n t_n\varphi_n}(\varphi_n)-\Delta_{\gamma_n}(\varphi)\|_{\lambda,2}.\end{aligned}$$

By (84), for some sequence $(\delta_n)_{n\in\mathbb{N}}\subset[0,1]$[100]

$$\begin{aligned}&\limsup_{n\to\infty}\sup_{\delta\in[0,1]}\|\Delta_{\gamma_n+\delta t_n\varphi_n}(\varphi_n)-\Delta_{\gamma_n}(\varphi)\|_{\lambda,2}\\ &\quad\leq\limsup_{n\to\infty}\|\Delta_{\gamma_n+\delta_n t_n\varphi_n}(\varphi_n)-\Delta_\gamma(\varphi)\|_{\lambda,2}+\limsup_{n\to\infty}\|\Delta_{\gamma_n}(\varphi)-\Delta_\gamma(\varphi)\|_{\lambda,2}\\ &\quad= o(1).\end{aligned}$$

Combine the two preceding displays and take $t_n=n^{-1/2}$ to yield (1.19):

$$\left\|\sqrt{n}\left(\sqrt{p_{\gamma_n,\tau_n,h_n}}-\sqrt{p_{\gamma_n}}\right)-\frac{1}{2}g_n\sqrt{p_{\gamma_n}}\right\|_{\lambda,2}=\left\|\frac{\sqrt{p_{\gamma_n+t_n\varphi_n}}-\sqrt{p_\gamma}}{t_n}-\Delta_{\gamma_n}(\varphi)\right\|_{\lambda,2}=o(1).$$

To conclude we note that Lemma 1.8 in van der Vaart (2002) along with (83) applied for each γ_n separately yields that $P_{\gamma_n}g_n=0$. The uniform square P_{γ_n}-integrability of g_n follows by Lemma C.8 on noting that by (84) (applied with $\delta_n=t_n=0$ and $u_n=0$) $P_{\gamma_n}g_n^2\to P_\gamma g^2$ (where $g:=\tau\dot{\ell}_\gamma+B_\gamma h$), and $p_{\gamma_n}\to p_\gamma$ λ-a.e.. Linearity of each B_{γ_n} is clear. □

Lemma B.8. *In the setting of Proposition 1.4.2, let G be the measure on $\mathbb{R}^{1+K}$ corresponding to ζ and $U=(\epsilon,X)\sim\zeta$. Let $\mathscr{N}:=\left\{-\phi(\epsilon,X)h_1(\iota(\theta,X))+h_2(\epsilon,X):h_1\in\dot{\mathscr{F}},h_2\in\dot{\mathscr{Z}}_\eta\right\}$. The closed linear span of $\mathscr{N}$ in $L_2(G)$ is*

$$\overline{\text{lin}}\,\mathscr{N}=\{q\in L_2(G):\mathbb{E}[q(U)]=0,\ \mathbb{E}[\epsilon q(U)|X]=\mathbb{E}[\epsilon q(U)|\iota(\theta,X)]\}.$$

[100] On the right hand side take $\varphi_n=\varphi$ and $\delta_n=0$.

Proof. [101] Let $h_1 \in \dot{\mathscr{F}}$ and $h_2 \in \dot{\mathscr{Z}}_\eta$. The definition of the sets $\dot{\mathscr{F}}$, $\dot{\mathscr{Z}}_\eta$ and (1.26) ensure that $\mathscr{N} \subset L_2(G)$. Taking $h_1 = 0$ and $h_2 = 0$, we have that $\mathbb{E}[-\phi(\epsilon, X)h_1(\iota(\theta, X))] = 0$ by Proposition 1.4.1. $\mathbb{E}[h_2(\epsilon, X)] = 0$ by definition. Additionally, we have by (1.28)

$$\mathbb{E}[-\epsilon\phi(U)h_1(\iota(\theta, X)) + \epsilon h_2(U)|X] = h_1(\iota(\theta, X)),$$

and since $\sigma(\iota(\theta, X)) \subset \sigma(X)$, by (1.28) and the law of iterated expectations

$$\mathbb{E}[-\epsilon\phi(U)h_1(\iota(\theta, X)) + \epsilon h_2(U)|\iota(\theta, X)] = h_1(\iota(\theta, X)).$$

Hence $\mathscr{N} \subset \{q \in L_2(G) : \mathbb{E}[q(U)] = 0,\ \mathbb{E}[\epsilon q(U)|X] = \mathbb{E}[\epsilon q(U)|\iota(\theta, X)]\}$. Both sets are clearly linear spaces, hence it suffices to show that the latter is the closure of the former. Suppose that $q \in \{q \in L_2(G) : \mathbb{E}[q(U)] = 0,\ \mathbb{E}[\epsilon q(U)|X] = \mathbb{E}[\epsilon q(U)|\iota(\theta, X)]\}$.

It follows from the defintion of $\tilde{m}$ that $\bar{m}(U) := \tilde{m}(\epsilon) - \mathbb{E}[\tilde{m}(\epsilon)|X]$ is bounded and $e \mapsto \bar{m}((e, X))$ is continuously differentiable with bounded derivative. For any bounded function $U \mapsto \tilde{\mathfrak{q}}(U)$ such that $e \mapsto \tilde{\mathfrak{q}}((e, X))$ is continuously differentiable with bounded deriviatives, define $\bar{\mathfrak{q}}(U) := \tilde{\mathfrak{q}}(U) - \mathbb{E}[\tilde{\mathfrak{q}}(U)|X]$ and put for a bounded function $a : \mathbb{R} \to \mathbb{R}$ where a is continuously differentiable with bounded derivative,

$$\mathfrak{q}(U) := \bar{\mathfrak{q}}(U) - \bar{m}(\epsilon)\left[\mathbb{E}[\bar{m}(\epsilon)\epsilon|X]\right]^{-1}\left[\mathbb{E}[\bar{\mathfrak{q}}(U)\epsilon|X] - a(\iota(\theta, X))\right].$$

By construction, $\mathfrak{q}$ is bounded, $e \mapsto \mathfrak{q}((e, X))$ is continuously differentiable with bounded derivative, $\mathbb{E}\left[\mathfrak{q}(U)|X\right] = 0$ and $\mathbb{E}\left[\epsilon\mathfrak{q}(U)|X\right] = 0$. Hence $\mathfrak{q} \in \dot{\mathscr{Z}}_\eta$. For any $\varepsilon > 0$, by Lemma C.7 of Newey (1991), there are $\tilde{\mathfrak{q}}$, a and ψ such that $\tilde{\mathfrak{q}}$ and a satisfy the conditions required for the construction of $\mathfrak{q}$ above and $\|q - \tilde{\mathfrak{q}}\|_{G,2}^2 < \varepsilon$, $\|\mathbb{E}[\epsilon q|\iota(\theta, X)] - a(\iota(\theta, X))\|_{G,2}^2 < \varepsilon$ and $\|\mathbb{E}[q|X] - \psi(X)\|_{G,2}^2 < \varepsilon$.[102] The proof is completed by arguing as in display (A.11) of Newey and Stoker (1993, p. 1220). □

Proof of Proposition 1.4.2. Lemma B.8 establishes the closed linear span of the nuisance tangent set. The orthogonal projection (in $L_2(G)$) of a function onto the orthocomplement of this set is given by Lemma A.2 in Newey and Stoker (1993). In particular, for $U = (\epsilon, X) \sim G$ and $V_n := \iota(\theta_n, X)$, the projection $\Pi\left(-\phi(U)f'(V_n)X_2)\middle|\mathscr{N}^\perp\right)$ has the form

$$\begin{aligned}
&\omega(X)\epsilon\left[\mathbb{E}[-\epsilon\phi(U)f'(V_n)X_2)|X] - \frac{\mathbb{E}\left[-\omega(X)\epsilon\phi(U)f'(V_n)X_2)|V_n\right]}{\mathbb{E}\left[\omega(X)|V_n\right]}\right] \\
&\quad = \omega(X)\epsilon f'(V_n)\left[\mathbb{E}\left[X_2\mathbb{E}\left[-\epsilon\phi(U)|X\right]|V_n\right] - \frac{\mathbb{E}\left[\omega(X)X_2\mathbb{E}\left[-\epsilon\phi(U)|X\right]|V_n\right]}{\mathbb{E}\left[\omega(X)|V_n\right]}\right] \\
&\quad = \omega(X)\epsilon f'(V_n)\left[\mathbb{E}\left[X_2|V_n\right] - \frac{\mathbb{E}\left[\omega(X)X_2|V_n\right]}{\mathbb{E}\left[\omega(X)|V_n\right]}\right],
\end{aligned}$$

[101] Cf. the proof of Lemma A.1 in Newey and Stoker (1993, pp. 1219 – 1220).

[102] I.e. a is bounded, continuously differentiable with bounded derivative and $\tilde{\mathfrak{q}}$ is bounded and $e \mapsto \tilde{\mathfrak{q}}((e, X))$ is continuously differentiable with bounded deriviatives.

where the last equality is by (1.28). As $(Y - f(V_n), X) \sim G$ under P_{γ_n}, the claimed form of the efficient score function follows. □

Proof of Lemma 1.4.3. We first show that $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = O_{P_{\gamma_n}}(1)$. For $\tilde{\epsilon}_i^2 := \epsilon_i^2 - \sigma^2$ we have

$$\begin{aligned}\sqrt{n}|\hat{\sigma}_n^2 - \sigma^2| &\lesssim \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\epsilon}_i^2 + \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(f(V_{n,i}) - \hat{f}_{n,i}\right)^2 \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\epsilon}_i^2 + \frac{1}{\sqrt{n}}\left[\sum_{i\in N^{(1)}} \left(f(V_{n,i}) - \hat{f}_{n,i}\right)^2 + \sum_{i\in N^{(2)}} \left(f(V_{n,i}) - \hat{f}_{n,i}\right)^2\right],\end{aligned}$$

The first right hand side term is $O_{P_{\gamma_n}}(1)$ by the CLT. Next define $\tilde{f}_{n,i} := (f(V_{n,i}) - \hat{f}_{n,i})$ and $\mathcal{C}_n := (W_j)_{j\in N_{-i}}$. On a set E_n with $P_{\gamma_n}(E_n) \to 1$ we have $\mathbb{E}[\tilde{f}_{n,i}^2|\mathcal{C}_n] \leq \mathcal{R}_{1,n,i} \leq r_n^2 = o(n^{-1/2})$ and hence by Markov's inequality, the second and third terms are $o_{P_{\gamma_n}}(1)$. Finally note that

$$\sqrt{n}|\hat{\sigma}_n^{-2} - \sigma^{-2}| = \frac{\sqrt{n}|\hat{\sigma}^2 - \sigma^2|}{|\hat{\sigma}_n^2\sigma^2|} = o_{P_{\gamma_n}}(1),$$

by $\sqrt{n}|\hat{\sigma}_n^2 - \sigma^2| = O_{P_{\gamma_n}}(1)$ and since for some $c > 0$, $\sigma^2 > c$ and with P_{γ_n}-probability approaching 1, $\hat{\sigma}_n^2 > c$ and so $1/|\hat{\sigma}_n^2\sigma^2| = O_{P_{\gamma_n}}(1)$. □

Proof of Proposition 1.4.4. That assumptions M, LAN and CM(ii) hold follows from Propositions 1.3.10, 1.4.1 and 1.4.2. We next show (1.6) holds. Let $\mathcal{C}_n$ be some collection of random vectors. Let $\delta_n \to 0$, $\delta_n' \to 0$. For a triangular array of random vectors $(R_{n,i})_{n\in\mathbb{N},i\leq n}$ if with P_{γ_n}-probability approaching one either (a) $\mathbb{E}[\|R_{n,i}\|_2|\mathcal{C}_n] \leq \delta_n n^{-1/2}$ or (b) for each element $R_{n,i,s}$ of $R_{n,i}$ and each $j \leq n'$, $\mathbb{E}[R_{n,i,s}R_{n,j,s}|\mathcal{C}_n] = 0$ (P_{γ_n}-a.s.) and $\mathbb{E}[R_{n,i,s}^2|\mathcal{C}_n] \leq \delta_n'$ then by Markov's inequality, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n'} R_{n,i} = o_{P_{\gamma_n}}(1)$ for $n' \leq n$. We establish that (a) or (b) holds for terms which sum to $\hat{\ell}_{n,\theta_n}(W_i) - \tilde{\ell}_{\gamma_n}(W_i)$. Abbreviate $Z_{n,i} := Z(V_{n,i})$ and let

$$\begin{aligned}R_{1,n,i} &:= (\hat{f}_{n,i} - f(V_{n,i}))f'(V_{n,i})(X_{2,i} - Z_{n,i}) \\ R_{2,n,i} &:= (Y_i - f(V_{n,i}))\left(f'(V_{n,i}) - \widehat{f'}_{n,i}\right)(X_{2,i} - Z_{n,i}) \\ R_{3,n,i} &:= (Y_i - f(V_{n,i}))\widehat{f'}_{n,i}\left(\hat{Z}_{n,i} - Z_{n,i}\right) \\ R_{4,n,i} &:= (\hat{f}_{n,i} - f(V_{n,i}))\left(f'(V_{n,i}) - \widehat{f'}_{n,i}\right)(X_{2,i} - Z_{n,i}) \\ R_{5,n,i} &:= (\hat{f}_{n,i} - f(V_{n,i}))\widehat{f'}_{n,i}\left(\hat{Z}_{n,i} - Z_{n,i}\right)\end{aligned}$$

For some $a_j \in \{-1, 1\}$, we have that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\ell}_{n,\theta_n}(W_i) - \tilde{\ell}_{\gamma_n}(W_i) = \sqrt{n}(\hat{\sigma}_n^{-2} - \sigma^{-2})\sigma^2\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}_{\gamma_n}(W_i) + \hat{\sigma}_n^{-2}\sum_{j=1}^{5}a_j\frac{1}{\sqrt{n}}\left[\sum_{i\in N^{(1)}}R_{j,n,i} + \sum_{i\in N^{(2)}}R_{j,n,i}\right].$$

The first term on the right hand side is $o_{P_{\gamma_n}}(1)$ by Lemma 1.4.3 and Proposition 1.3.1. For the second right hand side term first note that Lemma 1.4.3 also implies that $\hat{\sigma}_n^{-2} = O_{P_{\gamma_n}}(1)$. Let E_n be sets on which conditions (i) and (ii) in assumption SIM-NP(i) hold with $P_{\gamma_n}(E_n) \to 1$. For $j \in [3]$ we will show that (b) holds on E_n (for $i \in N^{(1)}$ or $i \in N^{(2)}$). That these terms are conditionally mean zero follows from the construction of the estimates. Specifically, using the fact that each $\hat{f}_{n,i}$, $\widehat{f'}_{n,i}$, $\hat{Z}_{n,i}$ is $\sigma(V_{n,i}, \{W_j\}_{j\in N_{-i}})$ measurable, independence, the LIE, Lemma C.5, $\mathbb{E}[\epsilon_i|X_i] = 0$ and $\mathbb{E}[(X_{2,i} - Z_{n,i})|V_{n,i}] = 0$, it follows that each $\mathbb{E}[R_{j,n,i,s}R_{j,n,k,s}|\mathcal{C}_n] = 0$ for $j \in [3]$ and $k \notin N_{-i}$ with $\mathcal{C}_n = (W_j)_{j\in N^{(1)}}$ for $i \in N^{(2)}$ and $\mathcal{C}_n = (W_j)_{j\in N^{(2)}}$ for $i \in N^{(1)}$. Similar arguments along with the (P_{γ_n}-a.s.) boundedness of X_2 and assumption SIM-NP(i) show that on E_n each component $\mathbb{E}[R^2_{j,n,i,s}|\mathcal{C}_n] \leq r_n^2$. For $j \in \{4, 5\}$ (a) holds on E_n as by SIM-NP(i), on E_n, each $\mathbb{E}[\|R_{j,n,i}\|_2|\mathcal{C}_n] \lesssim \mathcal{R}_{l,n,i}\mathcal{R}_{k,n,i} \leq r_n^2 = o(n^{-1/2})$ for $l, k \in [3]$.

For the second part we will verify assumption T, which suffices to establish (1.7) and assumption R by Corollary 1.3.14. Note first that by (1.28) and assumption SIM-NP(i) the elements of $\tilde{\ell}_{\gamma_n}$ satisfy $\mathbb{E}[\tilde{\ell}^4_{\gamma_n,l}] = \mathbb{E}[(\epsilon_i f'(V_{n,i})\omega(X_i)(X_{2,i} - Z_{n,i}))^4] \lesssim \mathbb{E}[\epsilon_i^4] < \infty$ and so by Cauchy-Schwarz and e.g. Theorem 2.5.11 in Durrett (2019), $\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}_{\gamma_n,l}\tilde{\ell}_{\gamma_n,k} - \mathbb{E}\tilde{\ell}_{\gamma_n,l}\tilde{\ell}_{\gamma_n,k} = O_{P_{\gamma_n}}(n^{-1/2}\log(n)^{1/2+\varkappa})$ for any $\varkappa > 0$. The distributional observation that under P_{γ_n}, $(Y - f(V_n), X) \sim G$ and the form of $\tilde{\ell}_{\gamma_n}$ then implies that $\tilde{\mathcal{I}}_{\gamma_n} = \tilde{\mathcal{I}}_{\gamma}$ and hence

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}_{\gamma_n}\tilde{\ell}'_{\gamma_n} - \tilde{\mathcal{I}}_{\gamma}\right\|_2 \leq \left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}_{\gamma_n}\tilde{\ell}'_{\gamma_n} - \tilde{\mathcal{I}}_{\gamma}\right\|_F = O_{P_{\gamma_n}}(n^{-1/2}\log(n)^{1/2+\varkappa}). \tag{86}$$

Secondly, write

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{n,\theta_n,l} - \tilde{\ell}_{\gamma_n,l}\right)^2 \lesssim \hat{\sigma}_n^{-4}\sum_{j=1}^{5}\frac{1}{n}\left(\sum_{i\in N^{(1)}}R^2_{j,n,i,l} + \sum_{i\in N^{(2)}}R^2_{j,n,i,l}\right) + (\hat{\sigma}_n^{-2} - \sigma^2)^2\sigma^4\mathbb{P}_n\tilde{\ell}^2_{\gamma_n,l}.$$

By Lemma 1.4.3, $\hat{\sigma}_n^{-4} = O_{P_{\gamma_n}}(1)$. Under assumptions SIM and SIM-NP(i), on E_n, each $\mathbb{E}[R^2_{j,n,i,l}|\mathcal{C}_n] \lesssim r_n^2$ as noted above. Since $r_n = o(\nu_n)$, Markov's inequality then implies that $\frac{1}{n}\sum_{i\in N^{(s)}}R^2_{j,n,i,l} = o_{P_{\gamma_n}}(\nu_n^2)$ for $s = 1, 2$. By Lemma 1.4.3 and equation (54), the second RHS term is $O_{P_{\gamma_n}}(n^{-1})$. Adding and subtracting and using Cauchy-Schwarz yields

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_{n,\theta_n}\hat{\ell}'_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\tilde{\ell}'_{\gamma_n}\right\|_2 \leq \left\|\frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_{n,\theta_n}\hat{\ell}'_{n,\theta_n} - \tilde{\ell}_{\gamma_n}\tilde{\ell}'_{\gamma_n}\right\|_F = o_{P_{\gamma_n}}(\nu_n). \tag{87}$$

Combine (86) and (87) to see that assumption T is satisfied with any sequence $(\nu_n)_{n\in\mathbb{N}}$ as in the statement of the proposition. $\square$

Proof of Proposition 1.4.7. Let $V_n \coloneqq \iota(\theta_n, X)$. We first note that (i) $\breve{\ell}_{\gamma_n} \in L_2^0(P_{\gamma_n})$ and (ii) $P_{\gamma_n}\left[\breve{\ell}_{\gamma_n} B_{\gamma_n} h\right] = 0$ for all $h \in H_\eta$. For (i) use the LIE to obtain that if $W \sim P_{\gamma_n}$

$$\mathbb{E}\breve{\ell}_{\gamma_n}(W) = \mathbb{E}\left[\mathbb{E}[\epsilon|X] f'(V_n)\breve{\omega}(X)\left(X_2 - \frac{\mathbb{E}\left[\breve{\omega}(X)X_2|V_n\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_n\right]}\right)\right] = 0,$$

and note that by boundedness of $\breve{\omega}$ (above and below), f', compactness of $\mathscr{X}$ we have $\mathbb{E}\breve{\ell}_{\gamma_n,k}(W)^4 < \infty$ for each $k = 1, \ldots, K-1$ which implies (i) and moreover that $\|\breve{\ell}_{\gamma_n}\|_2^2$ is uniformly P_{γ_n}-integrable. For (ii), if $W \sim P_{\gamma_n}$ then by the LIE, definition of $\dot{\mathscr{Z}}_\eta$ and (1.28)

$$\begin{aligned}
\mathbb{E}\left[\breve{\ell}_{\gamma_n}(W)[B_{\gamma_n}h](W)\right] &= \mathbb{E}\left[\mathbb{E}[\epsilon h_2(\epsilon, X)|X] f'(V_n)\breve{\omega}(X)\left(X_2 - \frac{\mathbb{E}\left[\breve{\omega}(X)X_2|V_n\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_n\right]}\right)\right] \\
&\quad + \mathbb{E}\left[-\mathbb{E}[\epsilon\phi(\epsilon, X)|X] f'(V_n)\breve{\omega}(X)\left(X_2 - \frac{\mathbb{E}\left[\breve{\omega}(X)X_2|V_n\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_n\right]}\right)\right] \\
&= \mathbb{E}\left[f'(V_n)\mathbb{E}\left[\breve{\omega}(X)X_2 - \frac{\breve{\omega}(X)\mathbb{E}\left[\breve{\omega}(X)X_2|V_n\right]}{\mathbb{E}\left[\breve{\omega}(X)|V_n\right]}\middle| V_n\right]\right] \\
&= 0.
\end{aligned}$$

The distributional observation that under P_{γ_n}, $(Y - f(V_n), X) \sim G$ and the form of $\breve{\ell}_{\gamma_n}$ then implies that $\Upsilon_{\gamma_n} = \Upsilon_\gamma$. Using this, along with (a) and (b) above, we can argue analogously to as in the proof of Proposition 1.3.1 (with $\tilde{\ell}_{\gamma_n}$ replaced by $\breve{\ell}_{\gamma_n}$ and $\tilde{\mathcal{I}}_{\gamma_n}$ replaced by Υ_{γ_n}) to conclude that under P_{γ_n,τ_n,h_n}, $\sqrt{n}\mathbb{P}_n\breve{\ell}_{\gamma_n} \rightsquigarrow \mathcal{N}(\Upsilon_\gamma\tau, \Upsilon_\gamma)$. Arguing as in the proofs of Propositions 1.3.2, 1.3.3 and Lemmas B.3, B.4, B.5 reveals that this suffices for the result provided we show that equations (1.6), (1.7) and (1.8) hold with $\breve{\ell}_{n,\theta_n}$ replacing $\hat{\ell}_{n,\theta_n}$, $\breve{\ell}_{\gamma_n}$ replacing $\tilde{\ell}_{\gamma_n}$, $\breve{\Upsilon}_{n,\theta_n}$ replacing $\hat{\mathcal{I}}_{n,\theta_n}$ and Υ_γ replacing $\tilde{\mathcal{I}}_\gamma$.

To this end we argue as in the proof of Proposition 1.4.4. Let $\mathcal{C}_n$ be some collection of random vectors, $\delta_n \to 0$ and $\delta_n' \to 0$. For any triangular array of random vectors $(R_{n,i})_{n\in\mathbb{N}, i\leq n}$ if with P_{γ_n}-probability approaching one either (a) $\mathbb{E}[\|R_{n,i}\|_2|\mathcal{C}_n] \leq \delta_n n^{-1/2}$ or (b) for each element $R_{n,i,s}$ of $R_{n,i}$ and any $j \leq n'$, $\mathbb{E}[R_{n,i,s}R_{n,j,s}|\mathcal{C}_n] = 0$ (P_{γ_n}-a.s.) and $\mathbb{E}[R_{n,i,s}^2|\mathcal{C}_n] \leq \delta_n'$ then by Markov's inequality, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n'} R_{n,i} = o_{P_{\gamma_n}}(1)$ for $n' \leq n$. We establish that (a) or (b) holds for terms which sum to $\breve{\ell}_{n,\theta_n}(W_i) - \breve{\ell}_{\gamma_n}(W_i)$. Abbreviate

$Z_{l,n,i} \coloneqq Z_l(V_{n,i})$ for $l \in [2]$ and let

$$\begin{aligned}
R_{1,n,i} &\coloneqq (\hat{f}_{n,i} - f(V_{n,i}))f'(V_{n,i})\breve{\omega}(X_i)(X_{2,i} - Z_{n,i}) \\
R_{2,n,i} &\coloneqq (Y_i - f(V_{n,i}))\left(f'(V_{n,i}) - \widehat{f'}_{n,i}\right)\breve{\omega}(X_i)(X_{2,i} - Z_{n,i}) \\
R_{3,n,i} &\coloneqq (Y_i - f(V_{n,i}))\widehat{f'}_{n,i}\breve{\omega}(X_i)\left(\hat{Z}_{n,i} - Z_{n,i}\right) \\
R_{4,n,i} &\coloneqq (\hat{f}_{n,i} - f(V_{n,i}))\left(f'(V_{n,i}) - \widehat{f'}_{n,i}\right)\breve{\omega}(X_i)(X_{2,i} - Z_{n,i}) \\
R_{5,n,i} &\coloneqq (\hat{f}_{n,i} - f(V_{n,i}))\widehat{f'}_{n,i}\breve{\omega}(X_i)\left(\hat{Z}_{n,i} - Z_{n,i}\right),
\end{aligned}$$

with $Z_{n,i} \coloneqq Z_{1,n,i}/Z_{2,n,i}$ and $\hat{Z}_{n,i} \coloneqq \hat{Z}_{1,n,i}/\hat{Z}_{2,n,i}$. For some $a_j \in \{-1, 1\}$, we have that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \breve{\ell}_{n,\theta_n}(W_i) - \breve{\ell}_{\gamma_n}(W_i) = \sum_{j=1}^{5} a_j \frac{1}{\sqrt{n}}\left[\sum_{i \in N^{(1)}} R_{j,n,i} + \sum_{i \in N^{(2)}} R_{j,n,i}\right].$$

Note also that

$$\hat{Z}_{n,i} - Z_{n,i} = \frac{(\hat{Z}_{1,n,i} - Z_{1,n,i})Z_{2,n,i} + (Z_{2,n,i} - \hat{Z}_{2,n,i})Z_{1,n,i}}{\hat{Z}_{2,n,i}Z_{2,n,i}},$$

and by assumption SIM-NP(ii) there is a sequence of sets E_n with $P_{\gamma_n}(E_n) \to 1$ such that each $\breve{\mathcal{R}}_{l,n,i} \le r_n$ and each $\hat{f}_{n,i}$, $\widehat{f'}_{n,i}$, $\hat{Z}_{1,n,i,k}$ are bounded uniformly in i and for all large enough $n \in \mathbb{N}$ and $\hat{Z}_{2,n,i}$ is bounded below and above, uniformly in i and for all large enough $n \in \mathbb{N}$. From this it follows that $\mathbb{E}\left[\|\hat{Z}_{n,i} - Z_{n,i}\|_2^2 | \mathcal{C}_n\right] \lesssim r_n^2 = o(n^{-1/2})$ on E_n where $\mathcal{C}_n = (W_j)_{j \in N^{(1)}}$ for $i \in N^{(2)}$ and $\mathcal{C}_n = (W_j)_{j \in N^{(2)}}$ for $i \in N^{(1)}$. Combining these observations we obtain that for $j \in \{4, 5\}$, on E_n, $\mathbb{E}[\|R_{j,n,i}\|_2 | \mathcal{C}_n] \lesssim r_n^2 = o(n^{-1/2})$, which establishes (a). For $j \in [3]$ we establish (b). Specifically, using the fact that each $\hat{f}_{n,i}$, $\widehat{f'}_{n,i}$, $\hat{Z}_{n,i}$ is $\sigma(V_{n,i}, \{W_j\}_{j \in N_{-i}})$ measurable, independence, the LIE, Lemma C.5, $\mathbb{E}[\epsilon_i | X_i] = 0$ and $\mathbb{E}[\breve{\omega}(X_i)(X_{2,i} - Z_{n,i}) | V_{n,i}] = 0$, it follows that $\mathbb{E}[R_{j,n,i,s}R_{j,n,k,s} | \mathcal{C}_n] = 0$ for $j \in [3]$ and $k \notin N_{-i}$ with $\mathcal{C}_n$ as above. Similar arguments along with the (P_{γ_n}-a.s.) boundedness of X_2 and the probabilistic rate and boundedness observations above show that on E_n each component $\mathbb{E}[R_{j,n,i,s}^2 | \mathcal{C}_n] \lesssim r_n^2$. For the second part we will verify assumption T, which suffices to establish the required modifications of (1.7) and (1.8) by Corollary 1.3.14. Note first that as noted above the components of $\breve{\ell}_{\gamma_n}$ satisfy $\mathbb{E}[\breve{\ell}_{\gamma_n,l}^4] < \infty$ and so by Cauchy-Schwarz and e.g. Theorem 2.5.11 in Durrett (2019), $\frac{1}{n}\sum_{i=1}^{n} \breve{\ell}_{\gamma_n,l}\breve{\ell}_{\gamma_n,k} - \mathbb{E}\breve{\ell}_{\gamma_n,l}\breve{\ell}_{\gamma_n,k} = O_{P_{\gamma_n}}(n^{-1/2}\log(n)^{1/2+\kappa})$ for any $\kappa > 0$. As noted above $\Upsilon_{\gamma_n} = \Upsilon_\gamma$ and hence

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \breve{\ell}_{\gamma_n}\breve{\ell}_{\gamma_n}' - \Upsilon_\gamma\right\|_2 \le \left\|\frac{1}{n}\sum_{i=1}^{n} \breve{\ell}_{\gamma_n}\breve{\ell}_{\gamma_n}' - \Upsilon_\gamma\right\|_F = O_{P_{\gamma_n}}(n^{-1/2}\log(n)^{1/2+\kappa}). \quad (88)$$

Secondly, write

$$\frac{1}{n}\sum_{i=1}^{n}\left(\check{\ell}_{n,\theta_n,l}-\check{\ell}_{\gamma_n,l}\right)^2 \lesssim \sum_{j=1}^{5}\frac{1}{n}\left(\sum_{i\in N^{(1)}} R^2_{j,n,i,l}+\sum_{i\in N^{(2)}} R^2_{j,n,i,l}\right).$$

As noted above, on E_n, each $\mathbb{E}[R^2_{j,n,i,l}|\mathcal{C}_n] \lesssim r_n^2$. Since $r_n = o(\nu_n)$, Markov's inequality then implies that $\frac{1}{n}\sum_{i\in N^{(s)}} R^2_{j,n,i,l} = o_{P_{\gamma_n}}(\nu_n^2)$ for $s = 1, 2$. Adding and subtracting and using Cauchy-Schwarz yields

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\check{\ell}_{n,\theta_n}\check{\ell}'_{n,\theta_n}-\check{\ell}_{\gamma_n}\check{\ell}'_{\gamma_n}\right\|_2 \leq \left\|\frac{1}{n}\sum_{i=1}^{n}\check{\ell}_{n,\theta_n}\check{\ell}'_{n,\theta_n}-\check{\ell}_{\gamma_n}\check{\ell}'_{\gamma_n}\right\|_F = o_{P_{\gamma_n}}(\nu_n). \tag{89}$$

Combine (88) and (89) to see that assumption T is satisfied with any sequence $(\nu_n)_{n\in\mathbb{N}}$ as in the statement of the proposition. □

B.4. Proofs for section 1.5

Proof of proposition 1.5.1. Fix arbitrary $\tau_n \to \tau \in \mathbb{R}^{d_\theta}$ and $h_n \to h \in H_\eta$. Since Θ is open and $\theta_n+\tau_n/\sqrt{n} \to \theta \in \Theta$ for sufficiently large n, $\theta_n+\tau_n/\sqrt{n} \in \Theta$. The construction of H_η ensures that also $\beta+b_n/\sqrt{n} \in \mathscr{B}$ for large enough n. The restrictions on $\dot{\mathscr{Z}}_\eta$ and $\dot{\mathscr{G}}_{\eta,k}$ along with the norm on H suffice to ensure that $\eta_0(1+t_nh_{n,0}) \in \mathscr{Z}$ and each $\eta_k(1+t_nh_{n,k}) \in \mathscr{G}$. Specifically, for $k = 0, 1, \ldots, K$, the convergence in ensures the exists of an $M \in (0, \infty)$ such that, for all large enough n, $|h_{n,k}| \leq M$ and $M/\sqrt{n} < 1$, λ-a.e.. This ensures that each $(1 + t_nh_{n,k}) > 0$ and hence $\eta_k(1 + t_nh_{n,k}) \geq 0$ (λ-a.e.). Moreover, the positivity of $1 + t_nh_{n,k}$ in combination with the continuous differentiability of $e \mapsto \sqrt{\eta_k(e)}$ and the fact that the square-root function is continuously differentiable away from 0, yields (via the chain rule) that $\sqrt{\eta_k(1 + t_nh_{n,k})}$ is continuously differentiable λ-a.e. (for $k \in [K]$). Moreover, for $k \in [K] \cup \{0\}$,

$$\int \eta_k(1+t_nh_{n,k})\,\mathrm{d}\lambda = \int \eta_k\,\mathrm{d}\lambda + t_n\int h_{n,k}\eta_k\,\mathrm{d}\lambda = 1 + t_n\int h_k\,\mathrm{d}G_k = 1.$$

Additionally by Jensen's inequality

$$\left\|\int \tilde{X}\tilde{X}'\zeta(1+h_{n,0}/\sqrt{n})\,\mathrm{d}\lambda - \int \tilde{X}\tilde{X}'\,\mathrm{d}G\right\|_2 \leq \frac{M}{\sqrt{n}}\left\|\int \tilde{X}\tilde{X}'\,\mathrm{d}G\right\| \leq \frac{M}{\sqrt{n}}\int \|\tilde{X}\|_2^2\,\mathrm{d}G \to 0,$$

which implies that for all large enough n, $\int \tilde{X}\tilde{X}'\zeta(1 + h_{n,1}/\sqrt{n})\,\mathrm{d}\lambda \succ 0$. By the boundedness of each $h_{n,k}$ for large enough n, for such n and any non-negative function f with $G_kf < \infty$,

$$\int \eta_k(1+t_nh_{n,k})f\,\mathrm{d}\lambda \leq (1+t_nM)G_kf < \infty.$$

Applying this with $k = 0$ and $f(\tilde{x}) = \|\tilde{x}\|_2^{4+\delta}$ completes the demonstration that $\eta_0(1 + t_n h_{n,0}) \in \mathscr{X}$ for all large enough n. Similarly applying it with $k \in [K]$ and $f(e) = |e|^{4+\delta}$ & $f(e) = |\phi_k(e)|^{4+\delta}$ ensures that the finite moment requirements in (1.35) are satisfied under $\eta_k(1 + t_n h_{n,k})$ for large enough n. By the definitions of $\mathscr{G}$ and $\dot{\mathscr{G}}_{\eta,k}$,

$$\int \iota\, \eta_k(1 + t_n h_{n,k})\, \mathrm{d}\lambda = \int \kappa\, \eta_k(1 + t_n h_{n,k})\, \mathrm{d}\lambda = 0,$$

verifying that the first two conditions of (1.35) hold under $\eta_k(1 + t_n h_{n,k})$. Lastly, since $G_k|\epsilon_k|^{4+\delta} < \infty$, the boundedness of $|h_{n,k}|$ ensures that

$$\int e^4 t_n h_{n,k}(e)\, \mathrm{d}G_k(e) \to 0, \quad \left[\int e^3 t_n h_{n,k}(e)\, \mathrm{d}G_k(e)\right]^2 \to 0$$

which, combined with $\mathbb{E}\epsilon_k^4 - 1 > (\mathbb{E}\epsilon_k^3)^2$ implies that for large enough n,

$$\int e^4(1 + t_n h_{n,k}(e))\, \mathrm{d}G_k(e) - 1 > \left[\int e^3(1 + t_n h_{n,k}(e))\, \mathrm{d}G_k(e)\right]^2,$$

completing the verification that $\eta_k(1 + t_n h_{n,k}) \in \mathscr{G}$ for all large enough n.

The next step is to establish (1.19). Firstly, for any given $u := (\tau, h) \in \mathbb{R}^{d_\theta} \times H_\eta$ let $\varphi := \varphi(u) := (\tau, b_1, b_2, \eta_0 h_0, \ldots, \eta_K h_K)$. Then, for any $\gamma \in \Gamma$, $t \in [0, \infty)$ and $u \in \mathbb{R}^{d_\theta} \times H_\eta$, define $q_{\gamma,t,u} := p_{\gamma + t\varphi}$ and $q_\gamma := q_{\gamma,0,0} = p_\gamma$. Finally, let $\Delta_\gamma(\varphi) := \frac{1}{2}[\tau'\dot{\ell}_\gamma + B_\gamma h]\sqrt{p_\gamma}$. For any $\gamma \in \Gamma$ and any $u \in \mathbb{R}^{d_\theta} \times H_\eta$, by Lemma S4 in Lee and Mesters (2022b),

$$\lim_{t\downarrow 0} \left\| \frac{\sqrt{p_{\gamma+t\varphi}} - \sqrt{p_\gamma}}{t} - \Delta_\gamma(\varphi) \right\|_{\lambda,2} = \lim_{t\downarrow 0} \left\| \frac{\sqrt{q_{\gamma,t,u}} - \sqrt{q_\gamma}}{t} - \frac{1}{2}[\tau'\dot{\ell}_\gamma + B_\gamma h]\sqrt{q_\gamma} \right\|_{\lambda,2} = 0. \tag{90}$$

In order to strengthen this directional differentiability into the result required by (1.19), we first establish an intermediate result. Let $(\delta_n)_{n\in\mathbb{N}} \subset [0, 1]$ be an arbitrary sequence, $t_n \downarrow 0$ and define $\gamma_n := \gamma_n + \delta_n t_n \varphi_n$ for $\varphi_n := \varphi(u_n)$ with $u_n \to u \in \mathbb{R}^{d_\theta} \times H_\eta$. Define also $A_n := A(\theta_n + \delta_n t_n \tau_n, \beta_1 + \delta_n t_n b_{n,1})$, $D_{1,l,n} := D_{1,l}(\theta_n + \delta_n t_n \tau_n, \beta_1 + \delta t_n b_{n,1})$, $\zeta_{l,k,j,n} := [D_{1,l,n}]_k [A_n^{-1}]_j'$, R_n is such that $\mathrm{vec}(R_n) = \beta_2 + \delta_n t_n b_{n,2}$, $V_n := Y - R_n X$ and finally

$$\phi_{k,n} := \phi_k + \frac{\delta_n t_n h'_{n,k}}{1 + \delta_n t_n h_{n,k}}.$$

We will show that $\|\Delta_{\gamma_n}(\varphi_n) - \Delta_\gamma(\varphi)\|_{\lambda,2} \to 0$ $(*)$. By Proposition 2.29 in van der Vaart (1998) it suffices to show that (i) $\Delta_{\gamma_n}(\varphi_n) \to \Delta_\gamma(\varphi)$ λ-a.e. and (ii) $\limsup_{n\to\infty} \|\Delta_{\gamma_n}(\varphi_n)\|_{\lambda,2}^2 \leq \|\Delta_\gamma(\varphi)\|_{\lambda,2}^2 < \infty$. We have that $\Delta_{\gamma_n}(\varphi_n) := \frac{1}{2}[\tau_n'\dot{\ell}_{\gamma_n} +$

$B_{\gamma_n} h_n]\sqrt{p_{\gamma_n}}$, with

$$p_{\gamma_n}(W) = |\det(A_n)| \prod_{k=1}^{K} [\eta_k(1 + \delta_n t_n h_{n,k})] (A_{n,k} V_n) \times [\eta_0(1 + \delta_n t_n h_{n,0})](\tilde{X})$$

$$\dot{\ell}_{\gamma_n,l}(W) = \sum_{k=1}^{K} \zeta_{l,k,k,n} [\phi_{k,n} (A_{n,k} V_n) A_{n,k} V_n + 1] + \sum_{k=1}^{K} \sum_{j=1,\ j \neq k}^{K} \zeta_{l,k,j,n} \phi_{k,n} (A_{n,k} V_n) A_{n,j} V_n$$

$$\begin{aligned}[B_{\gamma_n} h_n](W) = {} & h_{n,0}(\tilde{X}) + \sum_{k=1}^{K} h_{n,k}(A_{n,k} V_n) - \sum_{l=1}^{d_{\beta_2}} b_{n,2,l} \sum_{k=1}^{K} \phi_{k,n} (A_{n,k} V_n) A_{n,k} D_{2,l} X \\ & + \sum_{m=d_\theta+1}^{d_\theta+d_{\beta_1}} b_{n,1,m} \left[\sum_{k=1}^{K} \zeta_{m,k,k,n} [\phi_{k,n} (A_{n,k} V_n) A_{n,k} V_n + 1] \right] \\ & + \sum_{m=d_\theta+1}^{d_\theta+d_{\beta_1}} b_{n,1,m} \left[\sum_{k=1}^{K} \sum_{j=1,\ j \neq k}^{K} \zeta_{m,k,j,n} \phi_{k,n} (A_{n,k} V_n) A_{n,j} V_n \right].\end{aligned}$$

Note first that there is a $N \in \mathbb{N}$ such that for $n \geq N$ each $|h_{n,k}|$ and $|h'_{n,k}|$ is bounded above λ-a.e. by some $\bar{h} \in (0, \infty)$. This implies that $\phi_{k,n} \to \phi_k$ λ-a.e. The assumed continuity of $D_{1,l}$ and A imply that $A_n \to A$ and each $\zeta_{l,j,k,n} \to \zeta_{l,j,k}$ and it is clear from its definition that $V_n \to V := Y - RX$. Inspection of the preceding display in light of these observations reveals that (i) holds. For (ii), the finiteness of $\|\Delta_\gamma(\varphi)\|_{\lambda,2}^2 = {}^1\!/_4 P_\gamma[\tau' \dot{\ell}_\gamma + B_\gamma h]^2$ follows from Lemma 1.7 of van der Vaart (2002) and (90). For the remaining inequality it suffices to show that $P_{\gamma_n} \left[\tau_n \dot{\ell}_{\gamma_n} + B_{\gamma_n} h_n \right]^2 \to P_\gamma \left[\tau \dot{\ell}_\gamma + B_\gamma h \right]^2$. This will follow by Lemma 1.3.11 if we show that (a) P_{γ_n} converges to P_γ in total variation, (b) $g'_n := \tau_n \dot{\ell}_{\gamma_n} + B_{\gamma_n} h_n \in L_2(P_{\gamma_n})$ and $g := \tau \dot{\ell}_\gamma + B_\gamma h \in L_2(P_\gamma)$, (c) $g'_n \to g$ in P_γ-probability and (d) $(g'_n)_{n \in \mathbb{N}}$ is uniformly square P_{γ_n}-integrable.[103] For (a), note that inspection of the preceding display reveals that $p_{\gamma_n} \to p_\gamma$ λ-a.e.. Hence, $P_{\gamma_n} \to P_\gamma$ in total variation by Scheffé's theorem. (b) follows from the fact that (90) holds for each $\gamma \in \Gamma$, $\tau \in \mathbb{R}^{d_\theta}$, $h \in H_\eta$ and Lemma 1.7 in van der Vaart (2002). For (c) note that inspection of the preceding display once more gives that $g'_n \to g$ λ-a.e. and hence P_γ-a.s. as $P_\gamma \ll \lambda$. Finally, for (d), let $\rho = 2 + \delta/2$ where $\delta > 0$ is as in (1.35) & (1.36). Let N be large enough that for $n \geq N$, $t_n \in [0, 1)$, each $|h_{n,k}|$, $|h'_{n,k}| \leq \bar{h} \in (0, \infty)$, each $|\tau_{n,l}| \leq 2|\tau_l|$, $|\varsigma_{n,l}| \leq 2|\varsigma_l|$ $\|A_n\|_2 \leq 2\|A\|_2$, each $|\zeta_{l,k,j,n}| \leq 2|\zeta_{l,k,j}|$, $|\phi_{n,k}| \leq |\phi_k| + \bar{h}$ and $P_{\gamma_n} \in \mathcal{P}$.[104] It suffices to show that $\sup_{n \geq N} P_{\gamma_n} |g'_n|^\rho < \infty$. In particular, by Hölder's inequality (and given the bounds just discussed holding for $n \geq N$), it is enough to show that each of $P_{\gamma_n} |\phi_{n,k}(A_{n,k} V_n) A_{n,j} V_n|^\rho$ for all $(k, j) \in [K]^2$ and $P_{\gamma_n} |\phi_{n,k}(A_{n,k} V_n) A_{n,k} D_{2,l} X|^\rho$ for all $k \in [K]$ and $l \in [d_{\beta_2}]$ are bounded independently of n (for $n \geq N$). Note that under P_{γ_n}, $A_{n,k} V_n \sim \eta_k(1 + \delta_n t_n h_{n,k})$

[103] Since we are interested only in the limiting behaviour, we can replace any P_{γ_n} which are not probability measures with $P_{\gamma_{n'}}$ where n' indicates the first index for which all subsequent elements of the sequence are probability measures. That each P_{γ_n} is a probability measure for n sufficiently large can be established analogously to the same for $P_{\gamma_n, \tau_n, h_n}$, which was established at the start of this proof, upon replacing the t_n used in the argument there with $\delta_n t_n$.

[104] See footnote 103.

and $\tilde{X} \sim \eta_0(1 + \delta_n t_n h_{n,0})$. By Cauchy-Schwarz we have

$$P_{\gamma_n}\left[|\phi_{n,k}(A_{n,k}V_n)|^{\rho}|A_{n,j}V_n|^{\rho}\right] \leq P_{\gamma_n}|\phi_{n,k}(A_{n,k}V_n)|^{4+\delta}P_{\gamma_n}|A_{n,j}V_n|^{4+\delta},$$
$$P_{\gamma_n}\left[|\phi_{n,k}(A_{n,k}V_n)|^{\rho}|A_{n,k}D_{2,l}X|^{\rho}\right] \leq P_{\gamma_n}|\phi_{n,k}(A_{n,k}V_n)|^{4+\delta}P_{\gamma_n}|A_{n,k}D_{2,l}X|^{4+\delta}.$$

For $n \geq N$, $\eta_k(1 + \delta_n t_n h_{n,k}) \leq \eta_k(1 + \bar{h})$ and so by (1.35) & (1.36), for a constant C which does not depend on n,

$$P_{\gamma_n}|A_{n,j}V_n|^{4+\delta} \leq (1 + \bar{h})\int e^{4+\delta}\eta_j(e)\,\mathrm{d}\lambda < \infty,$$
$$P_{\gamma_n}|\phi_{n,k}(A_{n,k}V_n)|^{4+\delta} \leq C(1 + \bar{h})\int \left[|\phi_k(e)|^{4+\delta} + \bar{h}^{d+\delta}\right]\eta_k(e)\,\mathrm{d}\lambda < \infty,$$
$$P_{\gamma_n}|A_{n,j}D_{2,l}X|^{4+\delta} \leq (1 + \bar{h})[2\|A\|_2\|D_{2,l}\|_2]^{4+\delta}\int \|(1, \tilde{x}')\|_2^{4+\delta}\eta_0(\tilde{x})\,\mathrm{d}\lambda < \infty.$$

As each right hand side term in the preceding display does not depend on n, this completes the demonstration of (d) and hence of $(*)$.

We now establish (1.19). Suppose first that $\theta_n = \theta$ and let $u_n \to u$ be arbitrary and put $\varphi_n := \varphi(u_n)$, $\varphi := \varphi(u)$ and $t_n \downarrow 0$. Also let $g_\gamma := \tau'\dot{\ell}_\gamma + B_\gamma h$. For large enough n, $\gamma + \varphi_n \in \Gamma$ and so applying (90) and the mean value theorem (e.g. Drabek and Milota, 2007, Theorem 3.2.7) for all such n,

$$\begin{aligned}
&\left\|t_n^{-1}\left(\sqrt{q_{\gamma,t_n,u_n}} - \sqrt{q_\gamma}\right) - \frac{1}{2}g_\gamma\sqrt{q_\gamma}\right\|_{\lambda,2} \\
&\quad\leq \left\|t_n^{-1}\left(\sqrt{q_{\gamma,t_n,u_n}} - \sqrt{q_{\gamma,t_n,u}}\right)\right\|_{\lambda,2} + \left\|t_n^{-1}\left(\sqrt{q_{\gamma,t_n,u}} - \sqrt{q_\gamma}\right) - \frac{1}{2}g_\gamma\sqrt{q_\gamma}\right\|_{\lambda,2} \qquad (91)\\
&\quad\leq \sup_{\delta\in[0,1]}\left\|\Delta_{\gamma+\delta t_n(\varphi_n-\varphi)}(\varphi_n - \varphi)\right\|_{\lambda,2} + o(1).
\end{aligned}$$

For any sequence $(\delta_n)_{n\in\mathbb{N}} \subset [0,1]$ we have that $\|\Delta_{\gamma+\delta_n t_n(\varphi_n-\varphi)}(\varphi_n - \varphi) - \Delta_\gamma(0)\|_{\lambda,2} \to 0$ by $(*)$ and $\|\Delta_\gamma(0)\|_{\lambda,2} = 0$.[105] It follows that $\limsup_{n\to\infty}\sup_{\delta\in[0,1]}\left\|\Delta_{\gamma+\delta t_n(\varphi_n-\varphi)}(\varphi_n - \varphi)\right\|_{\lambda,2} = 0$ and hence

$$\left\|\frac{\sqrt{p_{\gamma+t_n\varphi_n}} - \sqrt{p_\gamma}}{t_n} - \Delta_\gamma(\varphi)\right\|_{\lambda,2} = \left\|\frac{\sqrt{q_{\gamma,t_n,u_n}} - \sqrt{q_\gamma}}{t_n} - \frac{1}{2}g_\gamma\sqrt{q_\gamma}\right\|_{\lambda,2} = o(1), \qquad (92)$$

which we note holds for any $\gamma \in \Gamma$, since such γ was arbitrary. Now, consider an arbitrary sequence $\theta_n \to \theta$ and $\gamma_n = (\theta_n, \eta)$. Using (92) and applying the mean value theorem at

[105] The latter observation follows directly from the definition of Δ_γ

each $n \in \mathbb{N}$ gives (e.g. Drabek and Milota, 2007, Theorem 3.2.7)

$$\begin{aligned}\left\| t_n^{-1}\left(\sqrt{q_{\gamma_n,t_n,u_n}} - \sqrt{q_{\gamma_n}}\right) - \frac{1}{2} g_{\gamma_n}\sqrt{q_{\gamma_n}} \right\|_{\lambda,2} &\leq |t_n^{-1}| \sup_{\delta\in[0,1]} \left\| \Delta_{\gamma_n+\delta t_n\varphi_n}(t_n\varphi_n) - t_n\Delta_{\gamma_n}(\varphi) \right\|_{\lambda,2} \\ &= \sup_{\delta\in[0,1]} \left\| \Delta_{\gamma_n+\delta t_n\varphi_n}(\varphi_n) - \Delta_{\gamma_n}(\varphi) \right\|_{\lambda,2}.\end{aligned} \tag{93}$$

By $(*)$ we have for some sequence $(\delta_n)_{n\in\mathbb{N}} \subset [0,1]$,[106]

$$\begin{aligned}&\limsup_{n\to\infty} \sup_{\delta\in[0,1]} \left\| \Delta_{\gamma_n+\delta t_n\varphi_n}(\varphi_n) - \Delta_{\gamma_n}(\varphi) \right\|_{\lambda,2} \\ &\quad \leq \limsup_{n\to\infty} \left\| \Delta_{\gamma_n+\delta_n t_n\varphi_n}(\varphi_n) - \Delta_{\gamma}(\varphi) \right\|_{\lambda,2} + \limsup_{n\to\infty} \left\| \Delta_{\gamma_n}(\varphi) - \Delta_{\gamma}(\varphi) \right\|_{\lambda,2} \\ &\quad = o(1).\end{aligned}$$

Combining this with (93) and taking $t_n = n^{-1/2}$ yields

$$\left\| \sqrt{n}\left(\sqrt{p_{\gamma_n,\tau_n,h_n}} - \sqrt{p_{\gamma_n}}\right) - \frac{1}{2} g_{\gamma_n}\sqrt{p_{\gamma_n}} \right\|_{\lambda,2} = \left\| t_n^{-1}\left(\sqrt{q_{\gamma_n,t_n,u_n}} - \sqrt{q_{\gamma_n}}\right) - \frac{1}{2} g_{\gamma_n}\sqrt{q_{\gamma_n}} \right\|_{\lambda,2} = o(1),$$

which implies (1.19).

Finally we demonstrate that $P_{\gamma_n} g_n = 0$ and the uniform square P_{γ_n}-integrability of the score functions g_n. That $P_{\gamma_n} g_n = 0$ and $g_n \in L_2(P_{\gamma_n})$ follows from (90) applied separately for each $n \in \mathbb{N}$ (with $\gamma = \gamma_n$) and Lemma 1.7 in van der Vaart (2002). The uniform square P_{γ_n}-integrability of $(g_n)_{n\in\mathbb{N}}$ follows from the uniform square P_{γ_n}-integrability of $(g_n')_{n\in\mathbb{N}}$ established in (d) above applied with $\delta_n = 0$, any $t_n \downarrow 0$ and $u_n = 0$. □

Proof of proposition 1.5.2. The claim regarding the form of the efficient score function follows from proposition 1.5.1, Lemma 3 of Lee and Mesters (2022a) and Lemma C.4.

For assumption CM(ii), fix $\tau \in \mathbb{R}^{d_\theta}$ and $h \in H_\eta$ and let $g_n = \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n} h$ and $g := \tau'\dot{\ell}_\gamma + B_\gamma h$ where $\dot{\ell}_\gamma$ and B_γ are defined analogously to in Proposition 1.5.1 but with $A = A(\theta, \beta_1)$ in place of $A_n = A(\theta_n, \beta_1)$. During the demonstration of $(*)$ in the proof of Proposition 1.5.1 it was shown that $\lim_{n\to\infty} P_{\gamma_n}(g_n')^2 = P_\gamma g^2$. Applying this result with $\delta_n = 0$, any $t_n \downarrow 0$ and $u_n = 0$ yields $\lim_{n\to\infty} P_{\gamma_n} g_n^2 = P_\gamma g^2$.

A similar argument can be used for the efficient score function. Let $\breve{\ell}_\gamma := (\tilde{\ell}_{\gamma,1}', \tilde{\ell}_{\gamma,2}')'$. Applied with $\delta_n = 0$, any $t_n \downarrow 0$ and $u_n = 0$, (a) in the proof of Proposition 1.5.1 yields that $P_{\gamma_n} \to P_\gamma$ in total variation. Since the components of $\breve{\ell}_\gamma$ and $\breve{\ell}_\gamma$ are defined as orthogonal projections onto subspaces of $L_2(P_{\gamma_n})$ and $\in L_2(P_\gamma)$ respectively, they lie in these spaces. Inspection of the form of each element of $\breve{\ell}_{\gamma_n}$ and $\breve{\ell}_\gamma$ reveals that $\breve{\ell}_{\gamma_n} \to \breve{\ell}_\gamma$ λ-a.e. and hence P_γ-a.s. as $P_\gamma \ll \lambda$. Let $\rho = 2 + \delta/2$ where δ is as in (1.35) & (1.36). Let $N \in \mathbb{N}$ be large enough that for $n \geq N$, each $|\tau_{n,l}| \leq 2|\tau_l|$, $|\varsigma_{n,l}| \leq 2|\varsigma_l|$, $\|A_n\|_2 \leq 2\|A\|_2$,

[106] On the right hand side take the trivial sequences $\varphi_n = \varphi$ and $\delta_n = 0$.

each $|\zeta_{l,k,j,n}| \leq 2|\zeta_{l,k,j}|$ and $P_{\gamma_n} \in \mathcal{P}$. To show that $\breve{\ell}^2_{\gamma_n,l}$ is uniformly P_{γ_n}-integrable for each $l \in [d_\theta + d_\beta]$ it suffices to show that $\sup_{n \geq N} P_{\gamma_n}|\breve{\ell}_{\gamma_n,l}|^\rho < \infty$ for each such l. In particular, by Hölder's inequality (and given the bounds just discussed holding for $n \geq N$) it is sufficient to show that each of (for all $(k,j) \in [K]^2$ with $k \neq j$ and $s \in [d_{\beta_2}]$)

$$P_{\gamma_n}|A_{n,k}V_n|^\rho,\ P_{\gamma_n}|\kappa(A_{n,k}V_n)|^\rho,\ P_{\gamma_n}|\phi_k(A_{n,k}V_n)A_{j,n}V_n|^\rho,\ P_{\gamma_n}|A_{n,k}D_{2,s}(X-\mu)\phi_k(A_{k,n}V_n)|^\rho,$$

are bounded independently of n (for $n \geq N$). Under P_{γ_n} $A_{n,k}V_n \sim \eta_k$ and $X \sim \eta_0$. Using independence, Hölder's inequality and (1.35) & (1.36) for constants $C_1, C_2 \in (0,\infty)$ independent of n

$$\begin{aligned}
P_{\gamma_n}|A_{n,k}V_n|^\rho &= \int e^\rho \,\mathrm{d}G_k(e) < \infty \\
P_{\gamma_n}|\kappa(A_{n,k}V_n)|^\rho &\leq C_1 \int (e^{4+\delta} + 1)\,\mathrm{d}G_k(e) < \infty \\
P_{\gamma_n}|\phi_k(A_{n,k}V_n)A_{j,n}V_n|^\rho &= \int |\phi_k(e_k)|^\rho \,\mathrm{d}G_k(e_k) \int |e_j|^\rho \,\mathrm{d}G_j(e_j) < \infty \\
P_{\gamma_n}|A_{n,k}D_{2,s}(X-\mu)\phi_k(A_{k,n}V_n)|^\rho &\leq C_2 \int (\|(1,\tilde{x})\|_2^\rho + \|\mu\|_2^\rho)\,\mathrm{d}G_0(\tilde{x}) \int |\phi_k(e_k)|^\rho \,\mathrm{d}G_k(e_k) < \infty.
\end{aligned}$$

Since each right hand side term in the preceding display does not depend on n, this establishes the uniform P_{γ_n}-integrability of each $\breve{\ell}^2_{\gamma_n,l}$. By Cauchy-Schwarz, the continuous mapping theorem and Lemma 1.3.11 it then follows that $P_{\gamma_n}\left[\breve{\ell}_{\gamma_n}\breve{\ell}'_{\gamma_n}\right] \to P_\gamma\left[\breve{\ell}_\gamma\breve{\ell}'_\gamma\right]$. To complete the argument, note that the convergence just established along with the uniform P_{γ_n}-integrability of each $\breve{\ell}^2_{\gamma_n,l}$ implies that also each component $\tilde{\ell}^2_{\gamma_n,l}$ (for $l \in [d_\theta]$) is uniformly P_{γ_n}-integrable and so the same holds for $\|\tilde{\ell}_{\gamma_n}\|_2^2$. Again by definition each component $\tilde{\ell}_{\gamma_n,l} \in L_2(P_{\gamma_n})$ and $\tilde{\ell}_{\gamma,l} \in L_2(P_\gamma)$ and so using the uniform P_{γ_n}-integrability just established, (1.46), $P_{\gamma_n}\left[\breve{\ell}_{\gamma_n}\breve{\ell}'_{\gamma_n}\right] \to P_\gamma\left[\breve{\ell}_\gamma\breve{\ell}'_\gamma\right]$, Cauchy-Schwarz, the continuous mapping theorem and Lemma 1.3.11 once more we may conclude that $\lim_{n\to\infty}\tilde{\mathcal{I}}_{\gamma_n} = \tilde{\mathcal{I}}_\gamma$.

It remains to check the boundedness of B_γ, which follows directly as

$$\|B_\gamma h\|_{P_\gamma,2} \lesssim \|b_1\|_2 + \|b_2\|_2 + \sum_{k=1}^K \|h_k\|_{G_k,2} \lesssim \|b\|_2 + \sum_{k=1}^K \|h_k\| = \|h\|.$$

□

Proof of proposition 1.5.3. That assumption M holds is a consequence of the model setup in assumption LSEM & the sampling assumption. Assumption CM(ii) follows by proposition 1.5.2. Assumption DQM holds by proposition 1.5.1, the proof of which also shows that the scores $\dot{\ell}_{\gamma_n} \in L_2^0(P_{\gamma_n})$ & $B_{\gamma_n} : H_\eta \to L_2^0(P_{\gamma_n})$. Then proposition 1.3.10 applied with $g_n = \tau'\dot{\ell}_{\gamma_n} + B_{\gamma_n}h$ yields that assumption LAN holds.

It remains to show that assumptions E and R hold.[107] Suppose that $(\beta_n)_{n\in\mathbb{N}} \subset \mathscr{B}$ is a deterministic $\sqrt{n}$-consistent sequence for β (as in assumption DSE) and let $\hat{\ell}_{\xi_n,1}$ & $\hat{\ell}_{\xi_n,2}$ be formed as in equation (1.49). Let $\gamma_n' := (\theta_n, \eta_n)$ with $\eta_n := (\beta_n, \eta_0, \dots, \eta_K)$. Let $\breve{\ell}_\gamma := (\tilde{\ell}_{\gamma,1}', \tilde{\ell}_{\gamma,2}')'$ and $\breve{\ell}_{\xi_n} := (\hat{\ell}_{\xi_n,1}', \hat{\ell}_{\xi_n,2}')'$. Components of $\breve{\ell}_{\xi_n}$ have one of two forms:

$$\hat{\ell}_{\xi_n,m,l}(W_i) = \sum_{k=1}^{K} \left[\zeta_{l,k,k,n} \left(\hat{\tau}_{n,k,1} e_{n,k,i} + \hat{\tau}_{n,k,2} \kappa(e_{n,k,i}) \right) + \sum_{j=1, j\neq k}^{K} \zeta_{l,k,j,n} \hat{\phi}_{n,k}(e_{n,k,i}) e_{n,j,i} \right],$$

$$\hat{\ell}_{\xi_n,2,d_{b_1}+s}(W_i) = \sum_{k=1}^{K} [-A_{n,k} D_{2,s}] \left[(X_i - \bar{X}_n) \hat{\phi}_{n,k}(e_{n,k,i}) - \bar{X}_n \left(\hat{\varsigma}_{n,k,1} e_{n,k,i} + \hat{\varsigma}_{n,k,2} \kappa(e_{n,k,i}) \right) \right]$$

(with $m = 1$ and $l \in [d_\theta]$ or $m = 2$ and $l \in [d_{\beta_1}]$ and $s \in [d_{\beta_2}]$). Under $P_{\gamma_n'}$, $e_{n,k,i} \simeq \epsilon_k$ and $e_{n,j,i} \simeq \epsilon_k$. Therefore, by assumptions LSEM and DSE, $\frac{1}{n}\sum_{i=1}^n \left[\hat{\phi}_{n,k}(e_{n,k,i}) - \phi_k(e_{n,k,i}) \right] e_{n,j,i} = o_{P_{\gamma_n'}}(n^{-1/2})$ and $\frac{1}{n}\sum_{i=1}^n \left[\hat{\phi}_{n,k}(e_{n,k,i}) - \phi_k(e_{n,k,i}) \right] (X_i - \mu) = o_{P_{\gamma_n'}}(n^{-1/2})$. Additionally, since $(e_{n,k,i})_{i=1}^n$ and $(\kappa(e_{n,k,i}))_{i=1}^n$ and $(\phi_k(e_{n,k,i}))_{n\in\mathbb{N}}$ are i.i.d. samples from mean zero distributions with finite variance under $P_{\gamma_n'}$ given assumption LSEM and equation (1.45), it follows that $\frac{1}{\sqrt{n}}\sum_{i=1}^n a_{n,k,i} = O_{P_{\gamma_n'}}(1)$, for $a_{n,k,i} \in \{e_{n,k,i}, \kappa(e_{n,k,i}), \phi_k(e_{n,k,i})\}$. The argument of Lemma 7 in Lee and Mesters (2022a) implies that $\|\hat{\varkappa}_{n,k} - \varkappa_k\|_2 = o_{P_{\gamma_n'}}(\nu_n) = o_{P_{\gamma_n}'}(1)$ for $\varkappa \in \{\tau, \varsigma\}$ where ν_n is defined as in assumption DSE.[108] Since $\tilde{X} \sim \eta_0$ under $P_{\gamma_n'}$, $\frac{1}{n}\sum_{i=1}^n X_i - \mu = o_{P_{\gamma_n'}}(1)$ by the LLN. The continuity of A and $D_{1,l}$ yields that each $\zeta_{l,k,j,n} \to \zeta_{l,k,j}$ and hence are bounded. Combining these observations yields that

$$\sqrt{n}\mathbb{P}_n \left[\breve{\ell}_{\xi_n} - \breve{\ell}_{\gamma_n'} \right] = o_{P_{\gamma_n'}}(1). \tag{94}$$

Let $\hat{I}_{\xi_n} := \mathbb{P}_n \breve{\ell}_{\xi_n} \breve{\ell}_{\xi_n}'$, $\breve{I}_{\gamma_n'} := \mathbb{P}_n \breve{\ell}_{\gamma_n'} \breve{\ell}_{\gamma_n'}'$ and $\breve{I}_{\gamma_n} := P_{\gamma_n} \breve{\ell}_{\gamma_n} \breve{\ell}_{\gamma_n}'$. Firstly, let $m, r \in \{1, 2\}$ and l, s be indices such that $\hat{\ell}_{\xi_n,m,l}$ and $\hat{\ell}_{\xi_n,r,s}$ are components of $\breve{\ell}_{\xi_n}$. Let $\hat{U}_{n,i,m,l} := \hat{\ell}_{\xi_n,m,l}(W_i)$, $\tilde{U}_{n,i,m,l} := \tilde{\ell}_{\xi_n,m,l}(W_i)$ and $D_{n,i,m,l} := \hat{U}_{n,i,m,l} - \tilde{U}_{n,i,m,l}$. By Cauchy-Schwarz, assumptions LSEM, DSE, (1.45) and arguing analogously to Lemma 8 of Lee and Mesters (2022a)

$$\left| \frac{1}{n}\sum_{i=1}^n D_{n,i,l,m} \tilde{U}_{n,i,r,s} \right| \leq \left(\frac{1}{n}\sum_{i=1}^n \tilde{U}_{n,i,r,s}^2 \right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n D_{n,i,l,m}^2 \right)^{1/2} = o_{P_{\gamma_n}'}(\nu_n)$$

$$\left| \frac{1}{n}\sum_{i=1}^n \tilde{U}_{n,i,l,m} D_{n,i,r,s} \right| \leq \left(\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,l,m}^2 \right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n D_{n,i,r,r}^2 \right)^{1/2} = o_{P_{\gamma_n}'}(\nu_n),$$

[107] The argument in this section proceeds similarly to the relevant parts of the proofs of Theorem 2 & Proposition 2 of Lee and Mesters (2022a).

[108] The Lemma as stated does not apply directly since it is for the case where $\theta_n = \theta$. Regardless, since $e_{n,k,i} \sim \eta_k$ and $\tilde{X} \sim \eta_0$ under $P_{\gamma_n'}$ the argument also holds in our case.

and hence $\mathscr{R}_{1,n} := \|\hat{I}_{\xi_n} - \check{I}_{\xi_n}\|_2 \leq \|\hat{I}_{\xi_n} - \check{I}_{\xi_n}\|_F = o_{P_{\gamma'_n}}(\nu_n)$.[109] Next let

$$Q_{n,i,l,m,r,s} := \tilde{\ell}_{\gamma'_n,l,m}(W_i)\tilde{\ell}_{\gamma'_n,r,s}(W_i) - \tilde{\ell}_{\gamma_n,l,m}(W_i)\tilde{\ell}_{\gamma_n,r,s}(W_i),$$

and let $\breve{Q}_{n,i,l,m,r,s}$ be defined analogously except with each $e_{n,k,i}$ replaced by $\epsilon_{i,k}$. Note that the distribution of $Q_{n,i,l,m,r,s}$ under $P_{\gamma'_n}$ is the same as that of $\breve{Q}_{n,i,l,m,r,s}$ under the product measure $G = \prod_{k=0}^{K} G_k$. Therefore, arguing analogously to the corresponding part of the proof of proposition 2 in Lee and Mesters (2022a), using their Lemma 6 and Theorems 2.5.11 & 2.5.12 in Durrett (2019) gives that $\mathscr{R}_{2,n} := \|\check{I}_{\xi_n} - \check{I}_{\gamma_n}\|_2 = o_{P'_{\gamma_n}}(\nu_n)$. Combining this with the result for $\mathscr{R}_{1,n}$ we have that

$$\|\hat{I}_{\xi_n} - \check{I}_{\gamma_n}\|_2 = o_{P_{\gamma'_n}}(\nu_n). \tag{95}$$

Next we demonstrate that for each pair m, l indexing and element of $\check{\ell}_{\gamma_n}$ we have

$$\int [\tilde{\ell}_{\gamma'_n,m,l}\sqrt{p_{\gamma'_n}} - \tilde{\ell}_{\gamma_n,m,l}\sqrt{p_{\gamma_n}}]^2 \, \mathrm{d}\lambda \to 0. \tag{96}$$

Note that λ-a.e. each $\tilde{\ell}_{\gamma'_n,m,l}\sqrt{p_{\gamma'_n}} \to \tilde{\ell}_{\gamma,m,l}\sqrt{p_\gamma}$ and $\tilde{\ell}_{\gamma_n,m,l}\sqrt{p_{\gamma_n}} \to \tilde{\ell}_{\gamma,m,l}\sqrt{p_\gamma}$ by the assumed continuity of A, each $D_{1,l}$, each η_k and each ϕ_k and the form of these functions. Hence by Proposition 2.29 in van der Vaart (1998) it suffices to show that $\int \tilde{\ell}^2_{\gamma'_n,m,l} \, \mathrm{d}P_{\gamma'_n} \to \int \tilde{\ell}^2_{\gamma,m,l} \, \mathrm{d}P_\gamma$ and $\int \tilde{\ell}^2_{\gamma_n,m,l} \, \mathrm{d}P_{\gamma_n} \to \int \tilde{\ell}^2_{\gamma,m,l} \, \mathrm{d}P_\gamma$, since $\tilde{\ell}_{\gamma,m,l} \in L_2(P_\gamma)$ by its definition. Define $Q_{n,i,l,m} := \tilde{\ell}^2_{\gamma_n,m,l}$, $Q'_{n,l,m} := \tilde{\ell}^2_{\gamma'_n,m,l}$ and $\breve{Q}_{n,l,m}$, $\breve{Q}'_{n,l,m}$ which are defined analogously except with each $e_{n,k,i}$ replaced by $\epsilon_{i,k}$. Under P_{γ_n}, $Q_{n,l,m}$ has the same distribution as $\breve{Q}_{n,l,m}$ has under G; similarly under $P_{\gamma'_n}$, $Q'_{n,l,m}$ has the same distribution as $\breve{Q}'_{n,l,m}$ has under G. Hence, $\int \tilde{\ell}^2_{\gamma'_n,m,l} \, \mathrm{d}P_{\gamma'_n} = \int Q'_{n,m,l} \, \mathrm{d}G$ and $\int \tilde{\ell}^2_{\gamma_n,m,l} \, \mathrm{d}P_{\gamma_n} = \int Q_{n,m,l} \, \mathrm{d}G$. This observation and the the continuity of A and each $D_{1,l}$ is sufficient for the required integral convergence to hold.[110] We note that the same argument which yielded the uniform P_{γ_n}-integrability of $\|\tilde{\ell}_{\gamma_n}\|_2^2$ in the proof of Proposition 1.5.2 can be used to show that that $\|\tilde{\ell}_{\gamma'_n}\|_2^2$ is uniform $P_{\gamma'_n}$-integrable. Since $\theta \mapsto \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ is locally constant, for all sufficiently large $n \in \mathbb{N}$ we have $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$. $\tilde{\mathcal{I}}_{\gamma_n} \to \tilde{\mathcal{I}}_\gamma$ (which holds as we have shown that assumption CM(ii) does). The proof is completed by applying Lemma B.7. □

Proof of corollary 1.5.4. This follows from propositions 1.5.3, 1.3.2 and 1.3.3. □

Proof of corollary 1.5.5. This follows from proposition 1.5.3 and corollaries 1.3.7 & 1.3.9, on noting that H_η – as defined in equation (1.43) – is a linear subspace of H whenever $\beta \in \operatorname{int} \mathscr{B}$. □

[109] Similarly to footnote 108, whilst Lemma 8 in Lee and Mesters (2022a) cannot be directly applied since it assumes $\theta_n = \theta$, the underlying argument continues to apply here as it is based on the fact that under the relevant measure (here $P_{\gamma'_n}$) $e_{n,k,i} \sim \eta_k$ and $\tilde{X} \sim \eta_0$. Moreover their assumptions 5 & 6 hold under assumptions LSEM, DSE and (1.45).

[110] See the corresponding part of the proof of proposition 2 in Lee and Mesters (2022a) for additional details.

C. Supporting results

Lemma C.1. *Let $\{Z_{n,k} : k \leq n, n \in \mathbb{N}\}$ be a triangular array of $L-$dimensional random vectors, such that each row is independent with $\mathbb{E}[Z_{n,k}] = 0$ and $\Sigma_{n,k} := \mathbb{E}\left[Z_{n,k}Z_{n,k}'\right]$ exists. Suppose that*

$$\frac{1}{n}\sum_{k=1}^{n} \Sigma_{n,k} \to \Sigma_{\star}, \tag{97}$$

with $\Sigma_{\star}$ positive semi-definite (and finite) and that for each $\varepsilon > 0$

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\|Z_{n,k}\|^2 \mathbf{1}\{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}\}\right] \to 0. \tag{98}$$

Then

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_{n,k} \rightsquigarrow \mathcal{N}(0, \Sigma_{\star}).$$

Proof. Put $\xi_{n,k} := Z_{n,k}/\sqrt{n}$ for $k \leq n$ and $\xi_{n,k} := 0$ otherwise. Fix $a \in \mathbb{R}^L$. For each $n \in \mathbb{N}$, let $\mathcal{F}_{n,k} = \sigma(\xi_{n,t} : t \leq k)$ for $k \leq n$ and $\mathcal{F}_{n,k} = \mathcal{F}_{n,n}$ otherwise. The adapted sequence $(a'\xi_{n,k}, \mathcal{F}_{n,k})_{k\in\mathbb{N}}$ is clearly a martingale difference sequence by the independence, mean zero and (square) integrability of each $Z_{n,k}$. Moreover, the sums $\sum_{k=1}^{\infty} a'\xi_{n,k} = \sum_{k=1}^{n} a'\xi_{n,k}$ and $\sum_{k=1}^{\infty} \mathbb{E}[(a'\xi_{n,k})^2] = \sum_{k=1}^{n} \mathbb{E}[(a'\xi_{n,k})^2]$ trivially converge with probability 1 for each $n \in \mathbb{N}$. By linearity and continuity we have that

$$\sum_{k=1}^{\infty} \mathbb{E}[(a'\xi_{n,k})^2] = \sum_{k=1}^{n} \mathbb{E}[(a'\xi_{n,k})^2] = a'\left[\frac{1}{n}\sum_{k=1}^{n} \Sigma_{n,k}\right] a \to a'\Sigma_{\star}a \geq 0.$$

Next, suppose that $a \neq 0$ and let $\varepsilon > 0$. We have that $\{|a'Z_{n,k}| \geq \varepsilon\sqrt{n}\} \subset \{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}/\|a\|\}$ and therefore

$$\sum_{k=1}^{\infty} \mathbb{E}\left[(a'\xi_{n,k})^2 \mathbf{1}\{|a'\xi_{n,k}| \geq \varepsilon\}\right] \leq \|a\|^2 \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\|Z_{n,k}\|^2 \mathbf{1}\{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}/\|a\|\}\right] \to 0,$$

by assumption.[111] Noting the assumed independence, the conditions of Theorem 18.1 of Billingsley (1999) are satisfied and hence

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} a'Z_{n,k} = \sum_{k=1}^{\infty} a'\xi_{n,k} \rightsquigarrow \mathcal{N}(0, a'\Sigma_{\star}a).$$

The claimed result then follows by an application of the Cramér-Wold theorem. □

Remark C.1. Lemma C.1 is, of course, completely standard. I record it here because I have been unable to find a reference for a multivariate CLT for triangular arrays which permits a

[111] In the case that $a = 0$ this limit trivially holds.

positive *semi*-definite limiting variance matrix.

Lemma C.2. *Let $\mathcal{G}$ be a closed subspace of $L_2(P)$ where the latter is separable and let $(g_m)_{m\in\mathbb{N}}$ denote an orthonormal basis in $\mathcal{G}$. Let for $m \in \mathbb{N}$, let Π_m denote the orthogonal projection on $\mathcal{G}_m := \operatorname{lin}\{g_1, \dots, g_m\}$ and let Π denote the orthogonal projection on $\mathcal{G}$. Then, for any $X \in L_2(P)$ we have that $\Pi_m X \to \Pi X$ in $L_2(P)$ as $m \to \infty$.*

Proof. We first note that the formulation in the lemma is well-defined: every subspace of a separable metric space is itself separable (see e.g. Proposition 26, section 9.6 of Royden and Fitzpatrick, 2010, p. 204-205). Since a closed subspace of a Hilbert space is also a Hilbert space (with the same inner product), it follows that $\mathcal{G}$ is separable and therefore possesses an orthonormal basis (e.g. Theorem 11, Section 16.3 of Royden and Fitzpatrick, 2010, p. 317-318). Since any finite dimensional subset of a Hilbert space is closed, the orthogonal projection operators Π_m are well defined. Throughout $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ will denote the inner product in $L_2(P)$.

By proposition I.4.7 in Conway (1985, p. 15) we have that

$$\Pi_m X = \sum_{k=1}^{m} \langle X, g_k \rangle g_k.$$

ΠX is the unique vector in $\mathcal{G}$ such that $\langle X - \Pi X, g \rangle = 0$ for all $g \in \mathcal{G}$ (see e.g. I.2.6 - I.2.8 in Conway, 1985, p. 9-10). Now, let $Y = \sum_{k=1}^{\infty} \langle X, g_k \rangle g_k$ which converges by e.g. lemma I.4.12 in Conway (1985, p. 16). By continuity and linearity of the inner product we then have that for any g_j

$$\langle X - Y, g_j \rangle = \langle X, g_j \rangle - \sum_{k=1}^{\infty} \langle X, g_k \rangle \langle g_k, g_j \rangle = \langle X, g_j \rangle - \langle X, g_j \rangle = 0.$$

Using linearity and continuity of the inner product once more permits the conclusion that $\langle X - Y, g \rangle = 0$ for any $g \in \mathcal{G}$. Hence $Y = \Pi X$. Then, we have $\Pi X - \Pi_m X = \sum_{k=m+1}^{\infty} \langle X, g_k \rangle g_k = Y - \sum_{k=1}^{m} \langle X, g_k \rangle g_k$ which converges to 0 in $L_2(P)$ by the convergence of $\sum_{k=1}^{m} \langle X, g_k \rangle g_k$ to Y. □

Lemma C.3. *Let X be an integrable random variable and Z a random element in a metric space $\mathcal{Z}$, both defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Then $\mathbb{E}[X|Z] = 0$ (P-almost surely) if and only if $\mathbb{E}[Xf(Z)] = 0$ for all square integrable functions $f : \mathcal{Z} \to \mathbb{R}$ such that $Xf(Z)$ is integrable.*

Proof. Suppose that $\mathbb{E}[X|Z] = 0$. We have

$$\mathbb{E}[Xf(Z)] = \mathbb{E}[\mathbb{E}[Xf(Z)|Z]] = \mathbb{E}[\mathbb{E}[X|Z]f(Z)] = 0.$$

Conversely suppose that $\mathbb{E}[Xf(Z)] = 0$ for all square-integrable functions $f : \mathcal{Z} \to \mathbb{R}$ with $Xf(Z)$ integrable. Let Y be any of the conditional expectations $\mathbb{E}[X|Z]$ and let $A \in \sigma(Z)$. There is a set $B \in \mathcal{B}(\mathbb{R})$ such that $A = Z^{-1}(B)$. Put f as the indicator $f(z) := \mathbf{1}\{z \in B\}$. Clearly $\mathbb{E}f(Z)^2 \leq 1$ and $Xf(Z)$ is integrable. Then, by definition,

$$\int_A Y \,\mathrm{dP} = \int_A X \,\mathrm{dP} = \int X f(Z) \,\mathrm{dP} = \mathbb{E}[Xf(Z)] = 0.$$

Now, suppose $\{Y \neq 0\}$ has positive measure. Then one of $\{Y > 0\}$ or $\{Y < 0\}$ must. Say the first, the argument for the latter is analogous. This is $\{Y > 0\} = E = \cup_{n \geq 1} E_n$ for $E_n := \{Y > 1/n\}$. So one E_k at least has positive measure. So $\int_E Y \,\mathrm{dP} \geq \int_{E_k} Y \,\mathrm{dP} \geq \int_{E_k} 1/k \,\mathrm{dP} = \mathrm{P}(E_k)/k > 0$. But this is a contradiction since $E \in \sigma(Z)$. □

Lemma C.4. *Let $\dot\ell$ and $\dot\kappa$ be L- and K- dimensional vectors of functions in $L_2(P)$ respectively. Define $\mathscr{B} := \mathrm{lin}\{\dot\kappa_1, \ldots, \dot\kappa_K\}$ and suppose that $\mathscr{G}$ is a subspace of $L_2(P)$. For any closed subspace $S \subset L_2(P)$, denote the orthogonal projection of $X \in L_2(P)$ on S by $\Pi\left(X \mid S\right)$. Then if $\breve{X} := \Pi\left(X \mid \mathscr{G}^\perp\right)$ we have*

$$\tilde\ell := \Pi\left(\dot\ell \mid [\mathscr{B} + \mathscr{G}]^\perp\right) = \breve\ell - \Pi\left(\breve\ell \mid \mathrm{lin}\{\breve\kappa_1, \ldots, \breve\kappa_K\}\right). \tag{99}$$

Moreover, if $\tilde I := P\left[\tilde\ell\tilde\ell'\right]$ and $\breve J := P\left[\left(\breve\ell', \breve\kappa'\right)'\left(\breve\ell', \breve\kappa'\right)\right]$ and $\breve J_{22}$ is positive-definite then

$$\tilde\ell = \breve\ell - \breve J_{12}\breve J_{22}^{-1}\breve\kappa, \quad \text{and} \quad \tilde I = \breve J_{11} - \breve J_{12}\breve J_{22}^{-1}\breve J_{21}. \tag{100}$$

Proof. The proof of the first claim is as discussed on p. 74 of Bickel et al. (1998). As there, noting that $\mathscr{G} \subset \mathrm{lin}\,\mathscr{B} + \mathscr{G}$ and using their equation (A.2.11) (p. 428) we obtain

$$\begin{aligned}
\tilde\ell &= \dot\ell - \Pi\left(\dot\ell \mid \mathscr{G}\right) - \Pi\left(\dot\ell \mid (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^\perp\right) \\
&= \dot\ell - \Pi\left(\dot\ell \mid \mathscr{G}\right) - \Pi\left(\dot\ell - \Pi\left(\dot\ell \mid \mathscr{G}\right) \mid (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^\perp\right) \\
&= \breve\ell - \Pi\left(\breve\ell \mid (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^\perp\right).
\end{aligned}$$

Now, suppose that $f \in \mathrm{lin}\{\breve\kappa_1, \ldots, \breve\kappa_K\}$. Then we have

$$f = \sum_{k=1}^K a_k \breve\kappa_k = \sum_{k=1}^K a_k \dot\kappa_k - \sum_{k=1}^K a_k \Pi(\dot\kappa_k \mid \mathscr{G}) \in \mathrm{lin}\,\mathscr{B} + \mathscr{G},$$

and moreover, since each $\breve\kappa_k \in \mathscr{G}^\perp$, linearity of the inner product implies the same holds for f. Hence $f \in (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^\perp$. For the reverse containment, suppose that $f \in (\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^\perp$. Then, we have for some $g \in \mathscr{G}$ that

$$f = \sum_{k=1}^K a_k \dot\kappa_k + g.$$

Now, suppose that $g \neq -\sum_{k=1}^{K} a_k \Pi\left(\dot{\kappa}_k \mid \mathscr{G}\right)$, and hence $g = -\sum_{k=1}^{K} a_k \Pi\left(\dot{\kappa}_k \mid \mathscr{G}\right) + h \neq 0$ for some $h \in \mathscr{G}$ with $h \neq 0$. Then

$$\langle f, h\rangle = \sum_{k=1}^{K} a_k \langle \dot{\kappa}_k, h\rangle - \sum_{k=1}^{K} a_k \langle \Pi(\dot{\kappa}_k \mid \mathscr{G}), h\rangle + \langle h, h\rangle = \sum_{k=1}^{K} a_k \langle \breve{\kappa}_k, h\rangle + \langle h, h\rangle = \langle h, h\rangle > 0,$$

which is a contradiction to $f \in \mathscr{G}^{\perp}$. Hence we must have $g = -\sum_{k=1}^{K} a_k \Pi\left(\dot{\kappa}_k \mid \mathscr{G}\right)$ and therefore $f = \sum_{k=1}^{K} a_k \breve{\kappa}_k \in \operatorname{lin}\{\breve{\kappa}_1, \ldots, \breve{\kappa}_K\}$. It follows that $(\mathscr{B} + \mathscr{G}) \cap \mathscr{G}^{\perp} = \operatorname{lin}\{\breve{\kappa}_1, \ldots, \breve{\kappa}_K\}$ which, in conjunction with the first display of the proof, yields (99).

Next, if $\breve{J}_{22}$ is positive definite, then the formulae in in (100) are well-defined. For the left hand side note that we have

$$P\left[\left(\breve{\ell} - \breve{J}_{12}\breve{J}_{22}^{-1}\breve{\kappa}\right)\breve{\kappa}'\right] = \breve{J}_{12} - \breve{J}_{12}\breve{J}_{22}^{-1}\breve{J}_{22} = \breve{J}_{12} - \breve{J}_{12} = 0,$$

implying that $\breve{\ell} - \breve{J}_{12}\breve{J}_{22}^{-1}\breve{\kappa}$ is the orthogonal projection of $\breve{\ell}$ onto the orthocomplement of $\operatorname{lin}\{\breve{\kappa}_1, \ldots, \breve{\kappa}_K\}$ (e.g. Conway, 1985, Theorem I.2.6) and hence satisfies the condition given in (99). The formula on the right hand side of (100) then follows by elementary calculations. □

Lemma C.5. *Suppose that X is an integrable random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ and $\sigma(\sigma(X) \cup \mathcal{H})$ is independent of $\mathcal{G}$. Then, almost surely $\mathbb{E}[X|\sigma(\mathcal{G} \cup \mathcal{H})] = E[X|\mathcal{H}]$.*

Proof. (i) $\mathbb{E}(X|\mathcal{H})$ is $\sigma(\mathcal{G} \cup \mathcal{H})$ measurable since $\mathbb{E}(X|\mathcal{H})$ is $\mathcal{H}$-measurable by definition. (ii) $\mathbb{E}(X|\mathcal{H})$ is integrable by definition of conditional expectation. (iii) We demonstrate that for each $A \in \sigma(\mathcal{G} \cup \mathcal{H})$,

$$\int_A \mathbb{E}(X|\mathcal{H})\, \mathrm{d}P = \int_A X\, \mathrm{d}P.$$

Let $\mathcal{M} = \{B \cap C : B \in \mathcal{G}, C \in \mathcal{H}\}$. This is closed under intersections and contains Ω. Additionally, we have that $\mathcal{G} \cup \mathcal{H} \subset \mathcal{M} \subset \sigma(\mathcal{G} \cup \mathcal{H})$ and therefore, $\sigma(\mathcal{M}) = \sigma(\mathcal{G} \cup \mathcal{H})$. Hence, by Theorem 34.1 in Billingsley (1995) it is sufficient to demonstrate $\int_{B \cap C} \mathbb{E}(X|\mathcal{H})\, \mathrm{d}P = \int_{B \cap C} X\, \mathrm{d}P$ for $B \in \mathcal{G}$ and $C \in \mathcal{H}$. To this end, suppose that $X \geq 0$ (without loss of generality, since the following argument can be applied to the two positive parts $X = X^{+} + X^{-}$ separately and linearity used to conclude otherwise). Then, we have that

$$\int_{B \cap C} X\, \mathrm{d}P = \mathbb{E}(\mathbf{1}_B \mathbf{1}_C X) = \mathbb{E}(\mathbf{1}_B)\mathbb{E}(\mathbf{1}_C X),$$

since $\mathcal{G}$ is independent of $\sigma(\sigma(X) \cup \mathcal{H})$. Additionally,

$$\begin{aligned}\int_{B\cap C} \mathbb{E}[X|\mathcal{H}]\,\mathrm{d}P &= \mathbb{E}\left(\mathbf{1}_B \mathbf{1}_C \mathbb{E}[X|\mathcal{H}]\right) \\ &= \mathbb{E}(\mathbf{1}_B)\mathbb{E}\left[\mathbf{1}_C \mathbb{E}[X|\mathcal{H}]\right] \\ &= \mathbb{E}(\mathbf{1}_B)\mathbb{E}\left[\mathbb{E}[\mathbf{1}_C X|\mathcal{H}]\right] \\ &= \mathbb{E}(\mathbf{1}_B)\mathbb{E}(\mathbf{1}_C X),\end{aligned}$$

using the independence between $\mathcal{G}$ and $\mathcal{H}$, 10.10 in Davidson (1994) and the LIE. □

Lemma C.6 (Cf. Theorem 2 in Andrews, 1987)**.** *Suppose that equations* (1.7) *and* (1.8) *hold. Then,*

$$\|\hat{\mathcal{I}}_{n,\theta_n}^{\dagger} - \tilde{\mathcal{I}}_{\gamma}^{\dagger}\|_2 = o_{P_{\gamma_n}}(1).$$

Proof. Let $r := \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)$ and let $\mathcal{M}$ denote the set of $d_\theta \times d_\theta$ matrices with rank r. Fix $\varepsilon > 0$ and let $\delta > 0$ be small enough that whenever $M \in \mathcal{M}$ is such that $\|\tilde{\mathcal{I}}_\gamma - M\|_2 < \delta$ we have $\|\tilde{\mathcal{I}}_\gamma^{\dagger} - M^{\dagger}\|_2 < \varepsilon$.[112] It follows that for each $n \in \mathbb{N}$,

$$\left\{\|\hat{\mathcal{I}}_{n,\theta_n}^{\dagger} - \tilde{\mathcal{I}}_{\gamma}^{\dagger}\|_2 \geq \varepsilon\right\} \subset \left\{\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_{\gamma}\|_2 \geq \delta\right\} \cup \left\{\operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) \neq r\right\},$$

and so

$$P_{\gamma_n}\left(\|\hat{\mathcal{I}}_{n,\theta_n}^{\dagger} - \tilde{\mathcal{I}}_{\gamma}^{\dagger}\|_2 \geq \varepsilon\right) \leq P_{\gamma_n}\left(\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_{\gamma}\|_2 \geq \delta\right) + P_{\gamma_n}\left(\operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) \neq r\right) \to 0.$$

□

Lemma C.7. *Suppose that equation* (1.7) *holds and* $\tilde{\mathcal{I}}_\gamma \succ 0$*. Then assumption R holds.*

Proof. The function $M \mapsto \operatorname{rank}(M)$ is lower-semicontinuous on the set of matrices of any (fixed) dimension. There is a $\delta > 0$ such that on the set $\{\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_\gamma\|_2 < \delta\}$, $d_\theta \geq \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) \geq \operatorname{rank}(\tilde{\mathcal{I}}_\gamma) - 1/2 > d_\theta - 1$, implying $\operatorname{rank}(\tilde{\mathcal{I}}_\gamma) = d_\theta = \operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n})$. Hence, by (1.7)

$$P_{\gamma_n}\left(\operatorname{rank}(\hat{\mathcal{I}}_{n,\theta_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\gamma)\right) \leq P_{\gamma_n}(\{\|\hat{\mathcal{I}}_{n,\theta_n} - \tilde{\mathcal{I}}_\gamma\|_2 < \delta\}) \to 1.$$

□

Lemma C.8. *Suppose that S is a Polish space and $(P_n)_{n\in\mathbb{N}}$ is a sequence of probability measures which converges in total variation to P, with each P_n and P defined on $(S, \mathcal{B}(S))$. If $(f_n)_{n\in\mathbb{N}}$ is a sequence of non-negative functions in $L_1(P_n)$ such that (a) $f_n \xrightarrow{P} f \in L_1(P)$ and (b) $P_n f_n \to P f$ then $(f_n)_{n\in\mathbb{N}}$ is uniformly P_n-integrable.*

[112] See e.g. section 6.6 in Ben-Israel and Greville (2003).

Proof. Condition (a) and $P_n \xrightarrow{TV} P$ together imply that $Q_n \rightsquigarrow Q$ where Q_n is the pushforward measure of P_n under f_n and Q the same of P under f. Let $h \in C_b(S)$. By change of variables (e.g. Bogachev, 2007, Theorem 3.6.1) $\int h \,\mathrm{d}Q_n = \int h(f_n)\,\mathrm{d}P_n$ and $\int g \,\mathrm{d}Q = \int h(f)\,\mathrm{d}P$. By (a) and the bounded convergence theorem, $\int h(f_n)\,\mathrm{d}P \to \int h(f)\,\mathrm{d}P$. By $P_n \xrightarrow{TV} P$

$$\left|\int h(f_n)\,\mathrm{d}P_n - \int h(f_n)\,\mathrm{d}P\right| \leq 2\bar{h} \sup\left\{\left|\int g\,\mathrm{d}P_n - \int g\,\mathrm{d}P\right|\right\} \to 0,$$

where $|h| \leq \bar{h} \in (0,\infty)$ and the supremum is taken over all measurable g with $0 \leq g \leq 1$. Hence $Q_n \rightsquigarrow Q$ as claimed. This, in conjunction with (b), Theorem 3.6 of Billingsley (1999) and translating terms yields the result. □

D. Tables & figures

D.1. Empirical rejection frequencies (ERF)

SIM

Table D.1: Homoskedastic SIM ERF (%), specification 1

$\epsilon \sim \mathcal{N}(0,1)$, $X_k \sim U(-1,1)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.24	6.58	6.14	15.94	14.38	18.92
400	$\sqrt{1}$	5.38	5.20	5.40	10.28	10.14	13.82
600	$\sqrt{1}$	5.50	5.70	5.14	8.06	7.88	11.22
800	$\sqrt{1}$	4.74	4.76	5.36	6.94	7.78	10.28
200	$\sqrt{2}$	5.46	5.36	5.38	17.62	15.18	19.90
400	$\sqrt{2}$	5.58	5.68	5.58	12.72	10.26	14.58
600	$\sqrt{2}$	4.60	5.48	5.42	10.66	9.14	13.20
800	$\sqrt{2}$	5.20	5.34	5.74	9.20	8.98	10.60
200	$\sqrt{4}$	5.22	5.50	5.62	20.86	19.10	24.62
400	$\sqrt{4}$	4.98	5.86	5.60	14.68	12.62	17.04
600	$\sqrt{4}$	4.92	5.20	5.52	12.80	9.82	15.10
800	$\sqrt{4}$	5.48	4.96	6.02	10.48	9.32	13.08
200	$\sqrt{8}$	5.12	5.34	5.60	16.28	22.52	26.20
400	$\sqrt{8}$	5.98	5.50	5.12	19.48	16.12	19.98
600	$\sqrt{8}$	5.62	5.00	6.48	15.24	14.18	16.94
800	$\sqrt{8}$	4.98	5.54	5.40	13.02	11.76	14.42
200	$\sqrt{16}$	4.82	5.64	5.22	12.28	20.08	21.76
400	$\sqrt{16}$	5.28	5.30	6.02	15.66	18.66	23.66
600	$\sqrt{16}$	4.58	5.46	5.62	19.30	15.68	19.64
800	$\sqrt{16}$	5.30	5.56	5.32	17.02	14.68	17.62

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Table D.2: Homoskedastic SIM ERF (%), specification 2

$\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\text{Beta}(2,3)$, $\xi \sim \text{Bernoulli}(1/2)$, $X_k \sim U(-1,1)$

		$\hat{S}$			W		
n	δ^{-1}	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$	$f=\delta f_1$	$f=\delta f_2$	$f=\delta f_3$
200	$\sqrt{1}$	4.82	5.56	5.94	14.72	12.88	16.98
400	$\sqrt{1}$	5.74	4.96	5.50	10.28	10.68	12.42
600	$\sqrt{1}$	4.78	4.98	5.08	7.98	8.52	10.56
800	$\sqrt{1}$	5.14	4.88	5.34	7.06	7.78	9.58
200	$\sqrt{2}$	4.82	5.84	5.94	17.06	15.38	19.58
400	$\sqrt{2}$	5.14	5.86	5.52	11.86	10.02	14.20
600	$\sqrt{2}$	5.18	5.26	5.46	9.72	9.22	12.84
800	$\sqrt{2}$	5.04	5.12	5.40	8.72	8.60	11.90
200	$\sqrt{4}$	5.26	5.48	5.78	19.84	18.44	22.34
400	$\sqrt{4}$	5.64	5.38	5.62	15.18	12.20	16.02
600	$\sqrt{4}$	6.18	5.66	5.64	10.92	10.18	15.18
800	$\sqrt{4}$	4.88	5.26	4.84	10.12	9.52	13.24
200	$\sqrt{8}$	5.10	5.38	5.08	15.36	20.18	25.64
400	$\sqrt{8}$	4.66	5.58	4.96	19.08	16.20	20.44
600	$\sqrt{8}$	5.22	4.92	5.52	15.14	13.08	16.36
800	$\sqrt{8}$	5.10	4.98	5.66	12.64	11.00	14.78
200	$\sqrt{16}$	5.28	4.76	5.60	12.58	18.62	21.90
400	$\sqrt{16}$	5.54	5.56	5.34	15.38	19.14	23.40
600	$\sqrt{16}$	5.24	5.20	5.32	18.08	14.98	20.26
800	$\sqrt{16}$	4.92	5.30	5.02	17.54	13.60	18.08

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Table D.3: Homoskedastic SIM ERF (%), specification 3

$\epsilon \sim \mathcal{N}(0,1)$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.28	5.56	6.52	14.74	15.76	14.42
400	$\sqrt{1}$	6.20	5.94	5.96	10.62	10.88	10.68
600	$\sqrt{1}$	5.64	5.62	5.70	9.28	9.00	9.06
800	$\sqrt{1}$	5.10	5.80	5.00	7.28	8.78	8.18
200	$\sqrt{2}$	6.14	5.62	5.80	17.74	20.14	16.92
400	$\sqrt{2}$	5.62	5.96	6.52	12.08	14.02	11.02
600	$\sqrt{2}$	5.70	5.26	5.66	9.72	11.16	9.94
800	$\sqrt{2}$	5.38	5.08	5.78	9.68	10.34	9.02
200	$\sqrt{4}$	6.20	5.44	5.32	20.84	25.02	20.26
400	$\sqrt{4}$	5.64	5.62	5.90	15.70	16.82	14.22
600	$\sqrt{4}$	5.24	5.54	5.88	12.20	13.08	11.32
800	$\sqrt{4}$	5.68	5.74	5.38	11.18	13.14	10.62
200	$\sqrt{8}$	5.42	5.88	5.54	15.70	25.26	16.86
400	$\sqrt{8}$	5.82	5.42	5.32	17.24	21.64	17.42
600	$\sqrt{8}$	5.80	5.84	5.94	15.82	16.56	15.24
800	$\sqrt{8}$	5.44	5.68	5.60	13.14	15.14	13.14
200	$\sqrt{16}$	5.52	5.94	5.86	12.32	20.14	12.94
400	$\sqrt{16}$	6.18	5.68	5.58	16.06	24.22	15.98
600	$\sqrt{16}$	5.76	5.72	5.66	17.90	22.20	16.80
800	$\sqrt{16}$	5.24	5.28	5.02	17.40	19.54	15.38

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Table D.4: Homoskedastic SIM ERF (%), specification 4

$\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\mathrm{Beta}(2,3)$, $\xi \sim \mathrm{Bernoulli}(1/2)$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.26	5.92	6.18	14.78	15.28	13.60
400	$\sqrt{1}$	5.50	5.84	5.54	10.44	10.90	9.50
600	$\sqrt{1}$	5.22	5.70	5.36	8.62	9.14	8.28
800	$\sqrt{1}$	5.26	5.32	5.90	8.26	9.72	8.40
200	$\sqrt{2}$	5.96	6.00	6.02	17.62	19.86	15.54
400	$\sqrt{2}$	5.18	5.16	5.96	12.32	14.40	11.10
600	$\sqrt{2}$	5.22	6.02	5.34	10.86	10.58	9.14
800	$\sqrt{2}$	5.38	4.96	6.02	8.94	10.44	8.36
200	$\sqrt{4}$	5.96	6.26	5.58	20.32	24.04	20.48
400	$\sqrt{4}$	5.78	6.40	6.00	15.26	16.46	13.52
600	$\sqrt{4}$	5.30	5.26	5.60	13.16	13.72	11.06
800	$\sqrt{4}$	5.18	5.62	5.04	10.12	12.38	9.56
200	$\sqrt{8}$	5.72	5.78	5.72	15.14	25.52	16.50
400	$\sqrt{8}$	5.24	5.54	6.14	18.22	21.88	17.82
600	$\sqrt{8}$	5.76	4.96	5.10	15.18	17.34	14.70
800	$\sqrt{8}$	5.46	5.48	5.82	14.26	15.30	13.28
200	$\sqrt{16}$	5.66	5.16	5.96	11.42	20.78	12.82
400	$\sqrt{16}$	5.66	5.84	6.00	15.58	24.86	16.28
600	$\sqrt{16}$	5.00	4.78	5.98	17.44	22.06	16.72
800	$\sqrt{16}$	5.60	5.64	5.36	16.78	19.94	15.90

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Table D.5: Heteroskedastic SIM ERF (%), specification 1, optimal weighting

$\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X_k \sim U(-1, 1)$, $\breve{\omega}(X) = \omega(X)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	6.38	6.64	6.20	18.72	16.76	23.46
400	$\sqrt{1}$	6.24	5.84	6.50	12.34	11.70	17.26
600	$\sqrt{1}$	5.78	5.12	5.72	10.38	10.96	14.70
800	$\sqrt{1}$	5.88	5.58	5.92	8.50	9.94	12.76
200	$\sqrt{2}$	5.76	5.76	6.12	22.62	19.30	25.86
400	$\sqrt{2}$	5.96	6.22	6.26	16.30	13.72	20.08
600	$\sqrt{2}$	5.52	5.46	6.26	14.46	11.70	15.70
800	$\sqrt{2}$	5.34	5.94	5.68	11.26	10.14	14.78
200	$\sqrt{4}$	5.32	5.72	5.44	27.12	24.36	30.40
400	$\sqrt{4}$	5.42	5.96	6.12	21.06	16.28	22.48
600	$\sqrt{4}$	5.24	5.52	5.74	15.50	13.38	19.58
800	$\sqrt{4}$	5.74	5.72	5.76	13.74	11.16	17.78
200	$\sqrt{8}$	5.40	5.64	5.46	19.66	25.36	30.08
400	$\sqrt{8}$	6.60	6.22	6.32	25.42	21.10	28.72
600	$\sqrt{8}$	5.50	5.80	6.60	21.34	17.78	23.80
800	$\sqrt{8}$	5.42	5.84	6.06	17.86	15.58	21.08
200	$\sqrt{16}$	5.86	6.26	5.74	14.06	23.96	25.06
400	$\sqrt{16}$	5.52	6.50	6.46	20.32	23.98	29.78
600	$\sqrt{16}$	5.50	5.74	5.08	25.04	22.00	29.20
800	$\sqrt{16}$	5.28	4.82	5.24	22.90	19.90	25.40

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table D.6: Heteroskedastic SIM ERF (%), specification 2, optimal weighting

$\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X_k \sim U(-1, 1)$, $\breve{\omega}(X) = \omega(X)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.40	5.94	6.46	19.10	16.34	18.46
400	$\sqrt{1}$	6.68	6.34	7.42	13.36	11.24	13.72
600	$\sqrt{1}$	5.94	6.14	6.00	10.28	8.74	10.88
800	$\sqrt{1}$	5.86	5.70	5.78	8.86	7.68	9.76
200	$\sqrt{2}$	5.12	5.32	5.70	23.74	19.96	22.58
400	$\sqrt{2}$	5.42	6.28	6.62	15.70	12.92	15.72
600	$\sqrt{2}$	5.92	6.00	5.92	12.66	10.44	12.86
800	$\sqrt{2}$	5.68	5.76	5.78	10.38	9.58	11.90
200	$\sqrt{4}$	5.64	6.50	5.94	23.30	22.86	25.92
400	$\sqrt{4}$	5.48	5.82	6.84	19.76	16.60	18.44
600	$\sqrt{4}$	5.82	5.74	6.24	15.70	13.08	14.30
800	$\sqrt{4}$	5.80	5.82	6.18	13.86	12.16	12.54
200	$\sqrt{8}$	5.98	5.70	5.50	14.74	23.00	28.56
400	$\sqrt{8}$	5.48	6.50	5.78	22.32	20.00	23.70
600	$\sqrt{8}$	5.46	5.76	6.24	20.56	16.76	19.02
800	$\sqrt{8}$	5.36	6.00	6.18	17.94	13.74	16.50
200	$\sqrt{16}$	4.96	6.20	5.42	12.96	18.18	26.24
400	$\sqrt{16}$	5.42	6.50	6.70	12.78	21.82	25.66
600	$\sqrt{16}$	5.20	5.86	5.58	18.30	21.24	23.82
800	$\sqrt{16}$	5.06	5.66	5.92	21.44	18.76	20.98

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table D.7: Heteroskedastic SIM ERF (%), specification 3, optimal weighting

$\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$, $\breve{\omega}(X) = \omega(X)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.08	5.98	6.10	15.22	16.62	16.40
400	$\sqrt{1}$	5.06	5.74	5.54	9.62	11.76	12.00
600	$\sqrt{1}$	5.56	5.94	5.84	8.18	11.02	10.86
800	$\sqrt{1}$	5.02	5.58	5.44	8.00	9.02	9.50
200	$\sqrt{2}$	5.70	5.62	5.50	17.94	19.58	19.94
400	$\sqrt{2}$	5.92	5.80	6.06	12.90	13.24	14.08
600	$\sqrt{2}$	6.20	6.02	5.38	9.52	11.22	11.54
800	$\sqrt{2}$	5.60	5.70	5.48	8.78	10.76	9.78
200	$\sqrt{4}$	5.66	6.02	5.50	20.92	24.00	22.28
400	$\sqrt{4}$	5.90	5.68	5.86	16.50	16.98	17.84
600	$\sqrt{4}$	5.08	5.40	5.92	12.20	14.42	14.44
800	$\sqrt{4}$	5.32	4.88	5.72	10.74	11.96	12.54
200	$\sqrt{8}$	5.62	5.36	5.56	18.02	26.58	17.74
400	$\sqrt{8}$	5.90	5.76	5.44	19.70	21.66	20.64
600	$\sqrt{8}$	5.70	5.86	5.76	16.72	17.70	18.04
800	$\sqrt{8}$	5.42	5.18	5.26	13.30	14.92	14.82
200	$\sqrt{16}$	5.20	5.18	5.30	12.16	21.54	15.70
400	$\sqrt{16}$	5.58	5.26	5.80	17.04	25.38	18.52
600	$\sqrt{16}$	5.68	5.42	5.88	18.78	22.58	20.06
800	$\sqrt{16}$	5.08	5.26	5.46	17.80	19.20	18.82

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table D.8: Heteroskedastic SIM ERF (%), specification 4, optimal weighting

$\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1, 1)$, $\breve{\omega}(X) = \omega(X)$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	4.74	5.34	5.66	14.86	14.12	17.52
400	$\sqrt{1}$	5.60	6.28	6.12	9.34	10.24	10.12
600	$\sqrt{1}$	5.66	6.00	5.48	6.82	7.76	7.62
800	$\sqrt{1}$	5.82	6.42	5.64	6.70	7.54	6.62
200	$\sqrt{2}$	5.16	5.56	5.84	19.24	17.10	20.10
400	$\sqrt{2}$	6.38	6.14	6.38	11.50	11.92	12.28
600	$\sqrt{2}$	5.62	5.08	6.02	8.34	9.38	9.98
800	$\sqrt{2}$	5.50	6.10	5.50	8.14	8.94	7.42
200	$\sqrt{4}$	5.88	5.58	5.48	24.48	22.80	23.50
400	$\sqrt{4}$	6.10	6.04	6.10	15.04	15.08	15.72
600	$\sqrt{4}$	5.98	6.32	5.84	11.54	11.70	11.30
800	$\sqrt{4}$	5.68	5.82	5.62	9.24	10.88	9.78
200	$\sqrt{8}$	5.48	4.96	5.26	24.04	27.46	24.94
400	$\sqrt{8}$	5.42	5.42	5.94	20.26	19.66	20.38
600	$\sqrt{8}$	5.74	5.58	5.76	16.10	15.44	15.22
800	$\sqrt{8}$	5.40	5.08	5.60	12.26	12.80	13.88
200	$\sqrt{16}$	5.50	4.60	5.10	17.84	22.32	22.20
400	$\sqrt{16}$	5.38	5.80	5.66	20.82	23.02	22.36
600	$\sqrt{16}$	5.54	5.86	5.56	19.78	19.58	21.86
800	$\sqrt{16}$	5.70	5.80	5.60	17.88	17.00	17.14

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table D.9: Heteroskedastic SIM ERF (%), specification 1, feasible weighting

$\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X_k \sim U(-1, 1)$, $\breve{\omega}(X) = 1$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	4.86	5.74	5.62	22.22	19.88	23.22
400	$\sqrt{1}$	5.64	5.20	6.04	15.80	13.76	18.40
600	$\sqrt{1}$	5.10	5.72	5.50	12.08	12.14	14.68
800	$\sqrt{1}$	5.32	4.88	5.32	10.82	11.06	13.50
200	$\sqrt{2}$	4.68	5.90	5.98	26.22	23.80	27.42
400	$\sqrt{2}$	5.18	5.72	6.44	19.14	15.68	20.74
600	$\sqrt{2}$	5.26	5.72	5.24	15.98	13.20	17.00
800	$\sqrt{2}$	5.28	5.28	6.16	14.00	12.24	16.22
200	$\sqrt{4}$	5.78	5.18	5.54	29.72	27.44	32.58
400	$\sqrt{4}$	5.88	5.46	6.14	24.32	19.24	24.88
600	$\sqrt{4}$	5.34	5.14	6.18	20.10	15.92	19.62
800	$\sqrt{4}$	5.14	5.28	5.10	17.86	14.08	18.12
200	$\sqrt{8}$	6.02	5.74	6.18	23.12	29.98	32.70
400	$\sqrt{8}$	5.44	5.34	5.94	29.00	26.08	29.76
600	$\sqrt{8}$	5.52	5.72	5.04	25.26	20.50	24.70
800	$\sqrt{8}$	5.16	5.70	6.18	21.74	17.42	22.78
200	$\sqrt{16}$	5.48	5.16	5.40	15.62	25.38	28.04
400	$\sqrt{16}$	5.78	5.50	5.86	23.62	28.28	33.34
600	$\sqrt{16}$	5.02	4.74	6.10	28.38	25.90	30.54
800	$\sqrt{16}$	5.00	5.14	5.28	27.00	21.72	26.24

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table D.10: Heteroskedastic SIM ERF (%), specification 2, feasible weighting

$\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X_k \sim U(-1, 1)$, $\breve{\omega}(X) = 1$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	5.10	5.34	6.56	18.52	18.06	22.40
400	$\sqrt{1}$	5.90	5.52	5.26	13.12	12.60	15.68
600	$\sqrt{1}$	5.28	5.10	5.36	9.94	10.36	13.10
800	$\sqrt{1}$	5.08	5.18	5.06	9.10	9.58	12.78
200	$\sqrt{2}$	5.48	5.86	5.86	21.18	19.64	23.92
400	$\sqrt{2}$	5.64	5.14	5.64	15.58	13.28	18.48
600	$\sqrt{2}$	4.70	5.86	5.52	11.58	11.48	14.84
800	$\sqrt{2}$	5.36	5.34	5.20	11.18	10.54	13.80
200	$\sqrt{4}$	4.84	5.22	5.78	21.96	23.54	27.20
400	$\sqrt{4}$	5.52	6.26	6.32	19.00	16.60	20.88
600	$\sqrt{4}$	5.18	5.76	5.14	15.90	13.58	18.66
800	$\sqrt{4}$	5.34	4.88	5.56	13.58	11.90	16.62
200	$\sqrt{8}$	4.86	5.92	5.30	15.86	23.46	27.62
400	$\sqrt{8}$	4.96	5.36	5.78	22.28	20.46	25.90
600	$\sqrt{8}$	5.22	5.66	5.44	19.80	16.18	21.58
800	$\sqrt{8}$	5.10	5.24	5.28	17.08	15.36	19.78
200	$\sqrt{16}$	5.10	5.42	5.68	12.16	17.86	20.54
400	$\sqrt{16}$	5.50	5.70	5.60	13.54	23.14	27.24
600	$\sqrt{16}$	5.54	5.36	5.98	18.22	20.12	25.44
800	$\sqrt{16}$	4.40	5.38	5.00	20.90	18.50	23.26

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table D.11: Heteroskedastic SIM ERF (%), specification 3, feasible weighting

$\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1, 1)$, $\breve{\omega}(X) = 1$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	6.14	5.46	6.16	17.80	18.88	17.66
400	$\sqrt{1}$	6.24	6.10	5.98	12.54	13.54	12.90
600	$\sqrt{1}$	6.02	5.58	6.44	10.78	11.70	9.96
800	$\sqrt{1}$	5.66	5.42	5.26	10.44	10.90	9.48
200	$\sqrt{2}$	6.08	5.62	5.42	22.22	22.46	20.82
400	$\sqrt{2}$	5.58	5.12	6.00	16.24	16.44	13.68
600	$\sqrt{2}$	5.64	5.66	6.02	12.46	13.22	11.64
800	$\sqrt{2}$	6.08	5.88	5.42	11.96	12.94	10.28
200	$\sqrt{4}$	6.04	5.98	6.12	26.00	28.62	21.70
400	$\sqrt{4}$	5.94	5.60	5.48	19.68	20.80	17.68
600	$\sqrt{4}$	6.10	5.44	5.54	16.54	16.96	14.42
800	$\sqrt{4}$	5.34	5.32	5.74	13.46	15.26	12.44
200	$\sqrt{8}$	5.36	5.72	5.44	19.90	28.44	17.34
400	$\sqrt{8}$	6.36	5.74	5.72	22.72	26.44	20.62
600	$\sqrt{8}$	5.82	5.68	4.98	19.84	20.78	17.80
800	$\sqrt{8}$	4.98	5.36	5.80	17.74	18.96	15.46
200	$\sqrt{16}$	4.90	5.42	5.28	15.04	23.14	15.82
400	$\sqrt{16}$	5.66	5.40	6.06	20.76	28.06	17.40
600	$\sqrt{16}$	5.64	5.26	5.72	22.92	26.58	19.36
800	$\sqrt{16}$	4.84	5.20	5.00	20.84	23.30	18.86

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2 \exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

Table D.12: Heteroskedastic SIM ERF (%), specification 4, feasible weighting

$\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1, 1)$, $\breve{\omega}(X) = 1$

		$\hat{S}$			W		
n	δ^{-1}	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$	$f = \delta f_1$	$f = \delta f_2$	$f = \delta f_3$
200	$\sqrt{1}$	6.20	6.34	5.80	18.88	21.20	19.30
400	$\sqrt{1}$	6.36	5.90	5.40	15.04	16.58	13.84
600	$\sqrt{1}$	5.40	5.74	5.24	12.34	14.32	12.60
800	$\sqrt{1}$	5.54	5.66	5.22	10.64	12.30	10.66
200	$\sqrt{2}$	5.72	5.98	6.70	23.24	25.90	23.48
400	$\sqrt{2}$	5.64	6.10	5.82	16.66	19.28	16.56
600	$\sqrt{2}$	5.28	5.22	5.64	14.12	16.08	13.76
800	$\sqrt{2}$	5.92	5.66	6.02	12.94	14.52	11.92
200	$\sqrt{4}$	5.94	6.46	6.12	29.54	29.14	27.76
400	$\sqrt{4}$	6.08	6.16	5.78	21.66	24.08	20.16
600	$\sqrt{4}$	5.10	5.80	5.56	17.50	18.74	14.90
800	$\sqrt{4}$	5.24	5.76	5.32	16.62	18.08	14.58
200	$\sqrt{8}$	6.30	5.96	5.82	26.38	34.50	25.68
400	$\sqrt{8}$	5.64	5.30	5.84	25.76	28.60	24.70
600	$\sqrt{8}$	5.52	5.84	5.72	22.16	23.56	20.06
800	$\sqrt{8}$	5.20	5.74	5.12	18.92	21.02	17.36
200	$\sqrt{16}$	5.44	5.06	6.18	15.94	28.06	18.10
400	$\sqrt{16}$	5.36	5.80	6.50	26.70	33.90	26.38
600	$\sqrt{16}$	5.04	6.00	5.46	26.72	29.04	24.70
800	$\sqrt{16}$	5.46	5.84	5.46	23.14	26.78	22.62

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the psuedo efficient score test. W is a Wald test based on an Ichimura (1993) type estimator as described in section 1.4.4. $f_1(v) = c_1(1 + \exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$, where the constants c_i $(i = 1, 2, 3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i ($i = 1, 2$) are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null.

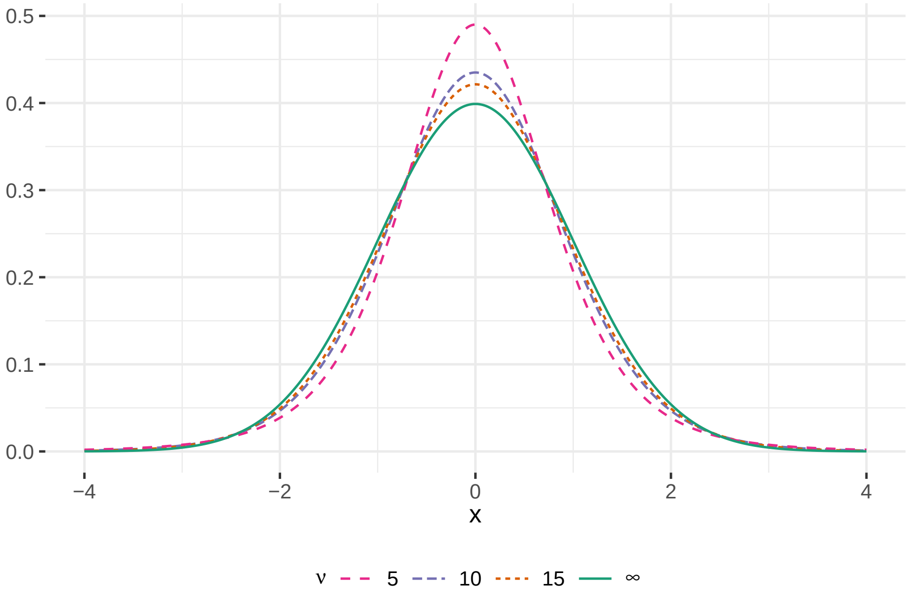
LSEM

Table D.13: True error distributions

	η_1		η_2
a	$\mathcal{N}(0,\,1)$	0 – 1	$\mathcal{N}(0,\,1)$
b	$t'(5)$	1 – 1	$t'(5)$
c	$\mathcal{SN}'(0,\,1,\,4)$	1 – 2	$t'(10)$
–	–	1 – 3	$t'(15)$
–	–	2 – 1	$\mathcal{SN}'(0,\,1,\,4)$
–	–	2 – 2	$\mathcal{SN}'(0,\,1,\,3)$
–	–	2 – 3	$\mathcal{SN}'(0,\,1,\,2)$
–	–	3 – 1	${}^{3}\!/_{4}\mathcal{N}(0,1) + {}^{1}\!/_{4}\mathcal{N}({}^{3}\!/_{2}, {}^{1}\!/_{9})$
–	–	3 – 2	${}^{17}\!/_{20}\mathcal{N}(0,1) + {}^{3}\!/_{20}\mathcal{N}({}^{3}\!/_{2}, {}^{1}\!/_{9})$
–	–	3 – 3	${}^{19}\!/_{20}\mathcal{N}(0,1) + {}^{1}\!/_{20}\mathcal{N}({}^{3}\!/_{2}, {}^{1}\!/_{9})$

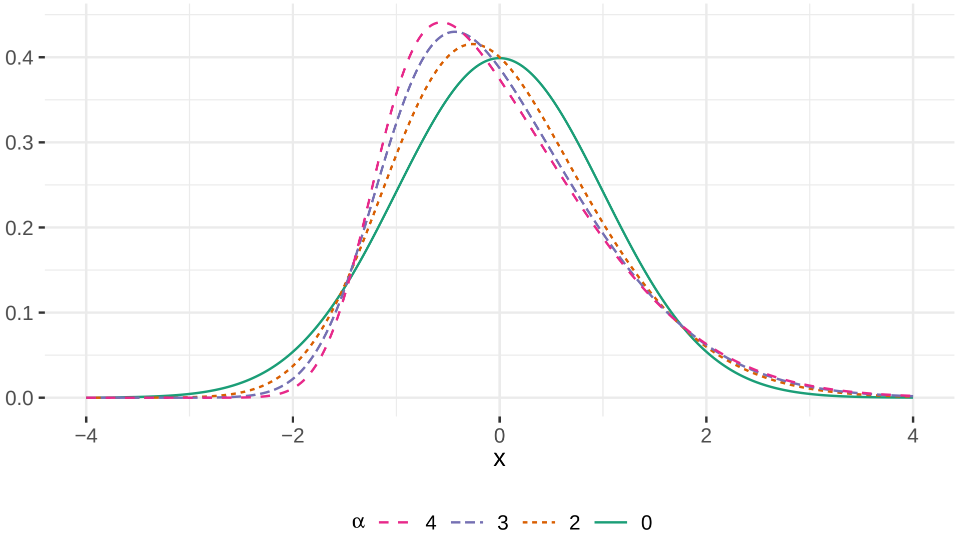
Notes: $\mathcal{SN}(\mu, \sigma, \alpha)$ denotes the skew normal distribution with location μ, scale σ and shape α. t' and $\mathcal{SN}'$ indicate that the corresponding t and skew normal distributions have been normalised to have zero mean and unit variance. The mixutre density in the right hand column is based on the "Skewed bimodal" density in Marron and Wand (1992).

Figure D.1: Density function of $t'(\nu)$



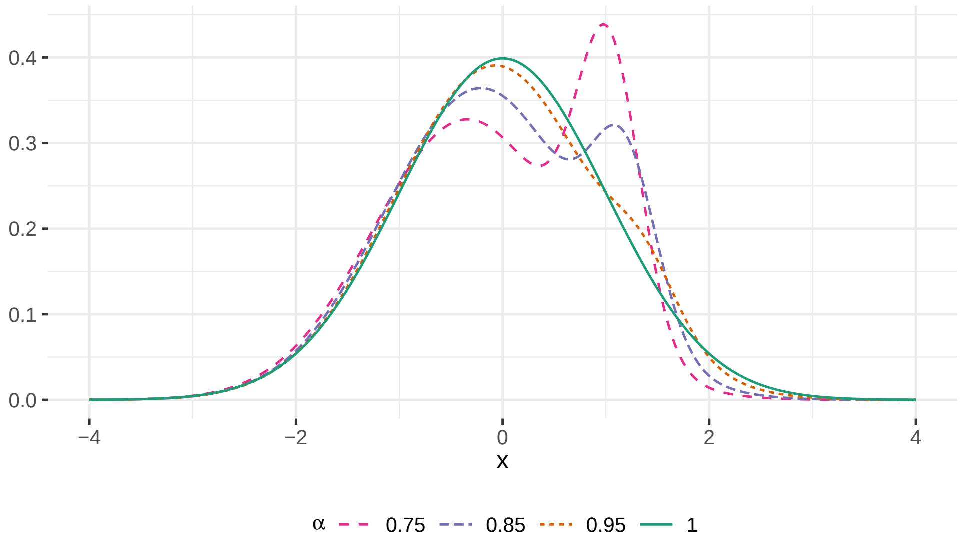
Densities 1 – j for $j = 1, 2, 3$ in table D.13.

Figure D.2: Density function of $\mathcal{SN}'(0, 1, \alpha)$



Densities 2 – j for $j = 1, 2, 3$ in table D.13.

Figure D.3: Density function of $\alpha\mathcal{N}(0,1) + (1-\alpha)\mathcal{N}(3/2, 1/9)$



Densities 3 – j for $j = 1, 2, 3$ in table D.13.

Figure D.4: Density functions for distributions used in LSEM simulation study (ii).

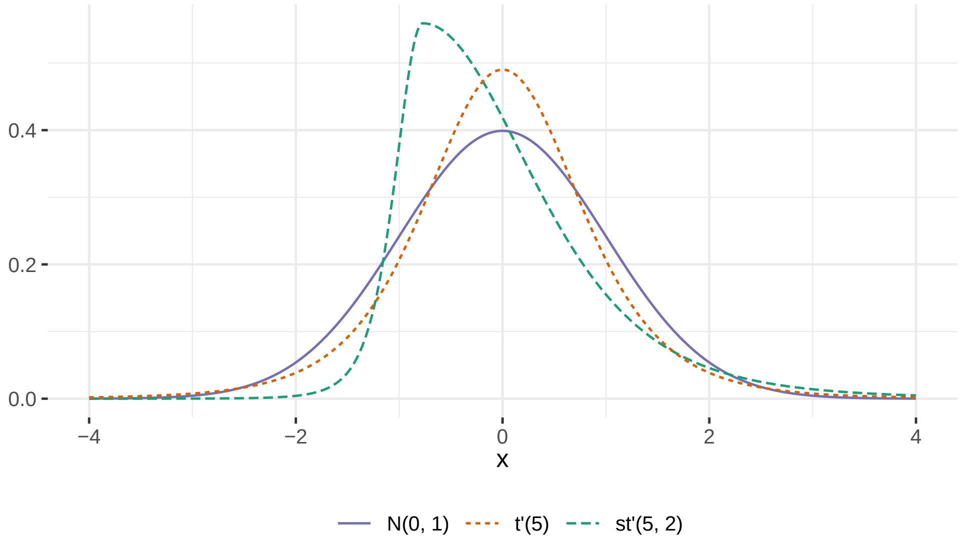


Table D.14: Empirical rejection frequencies (%) for LSEM, $\epsilon_1 \sim \mathcal{N}(0,1)$

n	0 – 1	1 – 1	1 – 2	1 – 3	2 – 1	2 – 2	2 – 3	3 – 1	3 – 2	3 – 3
$\hat{S}$										
200	4.74	5.62	5.90	5.50	3.20	3.86	4.48	2.92	3.58	4.62
400	4.78	5.52	4.44	5.16	2.82	3.68	4.66	1.92	3.58	4.24
600	4.60	4.84	4.20	4.74	2.50	3.42	3.76	2.18	3.34	4.56
800	4.56	4.28	4.48	4.12	2.62	2.94	3.56	2.52	3.86	4.16
$\hat{S}^*$										
200	6.94	6.58	6.76	7.26	6.74	6.78	6.46	7.10	7.00	6.88
400	6.82	6.66	6.44	6.76	8.02	7.36	7.74	5.94	7.12	6.46
600	7.04	7.32	5.86	6.58	8.60	7.80	6.68	6.50	6.74	6.82
800	6.68	6.38	6.48	6.04	8.68	7.50	5.84	5.74	7.20	6.82
$\hat{W}$										
200	33.00	16.76	23.26	25.30	28.84	29.86	30.02	61.02	54.36	40.56
400	32.98	11.74	17.18	21.78	26.40	26.16	27.02	74.60	64.60	44.28
600	33.32	10.02	14.28	18.82	23.92	25.42	26.62	82.70	71.10	44.68
800	33.32	8.98	13.64	16.62	23.78	22.60	24.82	88.16	77.50	47.32
$L\hat{M}$										
200	4.96	4.86	4.90	5.32	5.08	5.32	4.78	5.28	5.44	4.74
400	5.42	4.88	5.08	5.30	4.50	5.88	5.14	5.38	4.86	5.10
600	5.14	5.54	5.34	5.28	5.18	5.36	5.32	4.84	5.08	5.22
800	5.14	4.80	4.60	4.82	4.44	4.84	4.78	4.68	5.36	5.42
$\tilde{W}$										
200	27.38	32.18	30.20	29.80	28.28	29.48	28.76	23.10	24.40	25.50
400	25.26	30.24	28.76	27.92	27.60	27.88	26.58	21.94	22.70	22.60
600	23.82	28.14	27.54	28.14	26.02	26.12	26.74	18.76	20.78	21.68
800	23.14	26.86	26.94	25.86	26.62	26.54	25.64	16.88	20.26	20.86
$L\tilde{M}$										
200	30.52	35.66	32.88	31.76	31.16	32.52	31.28	22.90	24.66	28.04
400	21.64	27.26	23.34	23.56	22.84	22.74	22.38	14.66	15.86	17.74
600	16.64	22.86	18.76	19.58	18.30	18.46	20.16	9.10	11.02	14.36
800	14.72	19.68	15.72	15.88	16.60	16.70	16.50	6.84	8.52	11.52

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. Ŵ, LM̂ denote the Wald and LM tests based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). W̃ and LM̃ denote the Wald and LM tests based on a GMM estimator inspired by Lanne and Luoto (2021). Columns 2 – 14 denote the choice of density for ϵ_2, as in Table D.13.

Table D.15: Empirical rejection frequencies (%) for LSEM, $\epsilon_1 \sim t'(5)$

n	0 – 1	1 – 1	1 – 2	1 – 3	2 – 1	2 – 2	2 – 3	3 – 1	3 – 2	3 – 3
$\hat{S}$										
200	6.16	7.58	6.10	6.00	3.96	4.92	5.40	3.20	4.32	5.74
400	5.40	6.76	5.72	5.86	3.54	4.50	5.06	4.06	3.90	5.34
600	4.96	5.58	5.32	6.06	3.52	4.18	4.82	3.26	4.10	5.50
800	5.04	5.48	5.32	5.58	3.70	4.34	4.78	3.20	4.14	4.80
$\hat{S}^*$										
200	7.24	7.20	6.52	6.88	7.70	7.56	6.92	6.92	7.00	7.20
400	6.38	7.22	6.18	6.52	7.74	6.96	6.70	6.74	6.24	6.52
600	5.64	6.04	5.96	6.72	7.08	6.68	6.28	5.30	5.60	6.42
800	6.12	6.50	6.10	6.32	6.74	7.18	6.40	5.58	5.44	5.68
$\hat{W}$										
200	13.28	10.88	11.38	11.98	13.16	13.62	12.52	21.20	18.76	15.16
400	10.32	8.24	8.42	8.66	9.24	9.36	9.40	16.90	13.94	10.42
600	7.84	7.52	7.38	7.90	7.96	8.22	7.96	15.80	12.20	9.26
800	7.34	6.80	6.42	7.00	7.62	7.44	8.38	13.54	11.72	8.48
$\hat{LM}$										
200	5.20	4.72	4.70	5.00	5.24	5.24	5.46	5.46	5.60	5.42
400	5.40	5.10	5.04	4.80	5.34	4.98	5.30	5.84	5.62	5.14
600	4.78	4.64	4.44	5.18	4.94	5.02	5.22	5.48	5.40	5.12
800	4.82	5.04	5.50	5.40	5.28	4.84	4.38	5.72	5.66	4.48
$\tilde{W}$										
200	24.94	32.26	27.54	26.12	26.34	26.00	25.92	19.88	22.06	22.20
400	20.18	27.78	22.60	21.02	21.04	21.20	20.68	17.38	16.50	19.62
600	17.98	24.62	20.32	19.84	19.52	19.02	17.94	14.96	14.64	16.90
800	16.16	22.20	18.88	18.16	17.66	17.70	16.70	13.42	13.82	15.66
$\tilde{LM}$										
200	37.10	44.10	39.78	39.18	39.44	39.34	37.88	30.94	33.26	34.90
400	29.16	36.58	30.58	29.46	30.74	29.78	29.38	25.06	24.26	27.80
600	23.56	31.82	27.36	26.44	26.52	25.60	24.58	21.06	21.64	23.62
800	21.62	28.30	23.90	23.16	23.22	23.64	21.80	19.20	20.46	21.22

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\text{W}}$, $\hat{\text{LM}}$ denote the Wald and LM tests based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). $\tilde{\text{W}}$ and $\tilde{\text{LM}}$ denote the Wald and LM tests based on a GMM estimator inspired by Lanne and Luoto (2021). Columns 2 – 14 denote the choice of density for ϵ_2, as in Table D.13.

Table D.16: Empirical rejection frequencies (%) for LSEM, $\epsilon_1 \sim \mathcal{SN}'(0,1,4)$

n	0 – 1	1 – 1	1 – 2	1 – 3	2 – 1	2 – 2	2 – 3	3 – 1	3 – 2	3 – 3
$\hat{S}$										
200	4.90	5.84	5.56	5.48	3.88	4.62	5.16	3.58	4.16	5.08
400	5.22	5.70	5.14	5.00	3.38	4.52	4.92	3.88	4.18	4.54
600	5.18	5.72	4.98	5.52	3.46	4.22	4.78	3.00	4.10	5.08
800	5.10	5.02	5.12	5.22	3.76	3.78	5.08	4.02	3.84	5.02
$\hat{S}^*$										
200	6.02	6.68	6.26	6.40	7.44	6.56	6.74	6.42	6.32	6.18
400	6.34	6.42	6.12	5.84	6.96	6.82	6.46	7.16	6.88	6.32
600	6.34	6.44	5.94	6.40	6.64	6.74	6.26	5.88	6.18	6.16
800	5.58	6.12	6.12	5.86	7.66	5.82	6.42	6.08	5.54	6.40
$\hat{W}$										
200	28.96	15.94	20.38	22.70	26.36	25.26	26.94	53.34	47.80	35.26
400	27.76	11.36	15.36	18.42	22.64	22.46	23.50	62.94	51.70	35.58
600	25.48	9.02	13.54	16.44	20.22	20.10	20.34	68.34	56.94	34.72
800	24.38	9.04	11.48	13.62	18.52	18.68	19.58	73.12	59.92	35.18
$L\hat{M}$										
200	4.84	4.74	5.46	4.36	4.80	5.34	5.46	5.42	5.26	5.16
400	5.44	4.94	5.10	4.26	5.50	5.12	4.26	4.82	5.66	5.42
600	5.02	4.80	5.40	5.30	5.18	4.66	4.88	5.14	5.04	5.02
800	4.98	5.20	4.90	5.58	5.66	4.80	5.70	4.84	5.04	4.90
$\tilde{W}$										
200	27.76	34.48	31.56	29.22	31.88	30.72	30.84	23.16	23.90	26.04
400	24.48	32.28	29.04	28.04	27.94	27.70	27.48	17.80	18.94	23.36
600	20.88	29.54	26.58	24.60	25.72	24.32	23.58	14.38	15.06	18.48
800	20.42	27.94	26.74	23.54	25.42	24.08	23.52	12.50	13.26	16.72
$L\tilde{M}$										
200	35.10	39.54	37.00	36.14	38.18	37.32	38.24	28.98	29.82	33.76
400	27.72	29.62	27.98	28.28	27.94	27.52	27.54	18.90	21.02	25.62
600	21.22	24.22	23.04	21.74	22.74	22.24	22.80	15.42	16.26	19.70
800	20.18	22.34	20.74	18.36	20.48	20.52	21.18	12.18	13.64	17.50

Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. Ŵ, LM̂ denote the Wald and LM tests based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). W̃ and LM̃ denote the Wald and LM tests based on a GMM estimator inspired by Lanne and Luoto (2021). Columns 2 – 14 denote the choice of density for ϵ_2, as in Table D.13.

Table D.17: Empirical rejection frequencies (%) for LSEM

η_1, η_2	n	$\hat{S}$	$\hat{S}^*$	$\dot{S}$	$\dot{S}^*$
1	200	5.20	7.52	7.24	11.84
1	400	4.80	7.24	7.92	12.66
1	600	4.32	6.86	7.58	11.94
1	800	4.32	6.30	7.38	10.76
2	200	7.42	7.68	6.14	9.92
2	400	6.46	6.92	5.48	8.60
2	600	5.56	6.42	5.48	7.98
2	800	5.32	6.24	4.96	7.86
3	200	4.26	7.18	9.10	13.20
3	400	4.06	7.28	8.42	12.68
3	600	3.52	6.90	7.84	12.04
3	800	4.06	7.36	7.56	11.98

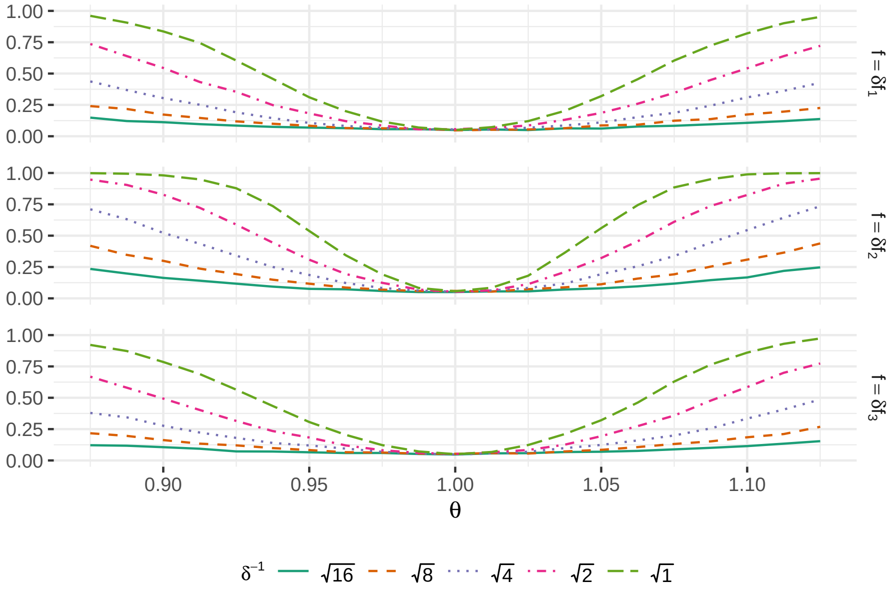
Notes: Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\dot{S}$ and $\dot{S}^*$ are score tests based on the score function for the Euclidean parameters using OLS estimates and 1-step updates respectively. The first column denotes the choice of density for both ϵ_1 and ϵ_2 as in the left colum of Table D.13.

D.2. Power curves

SIM

Figure D.5: Homoskedastic SIM power curve, specification 1

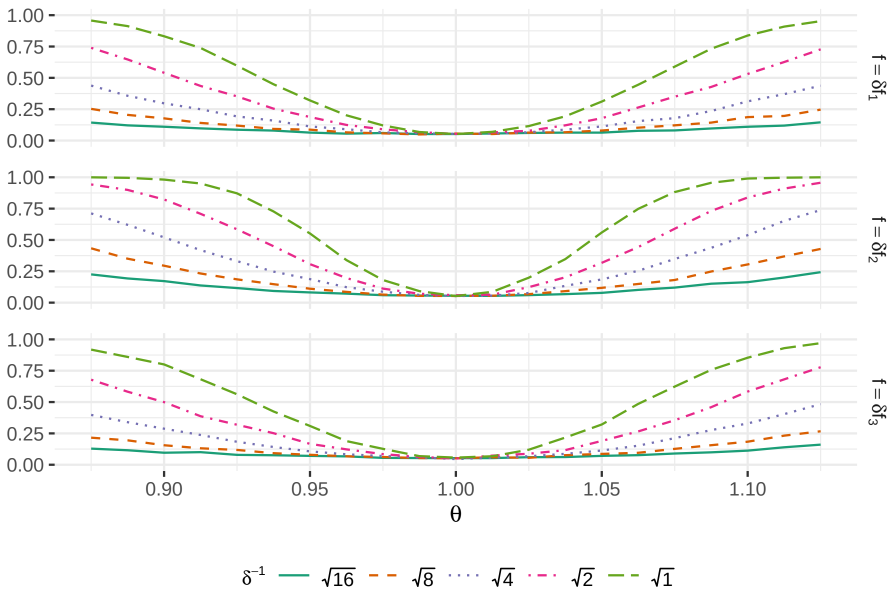
$\epsilon \sim \mathcal{N}(0,1)$, $X_k \sim U(-1,1)$



Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3 v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Figure D.6: Homoskedastic SIM power curve, specification 2

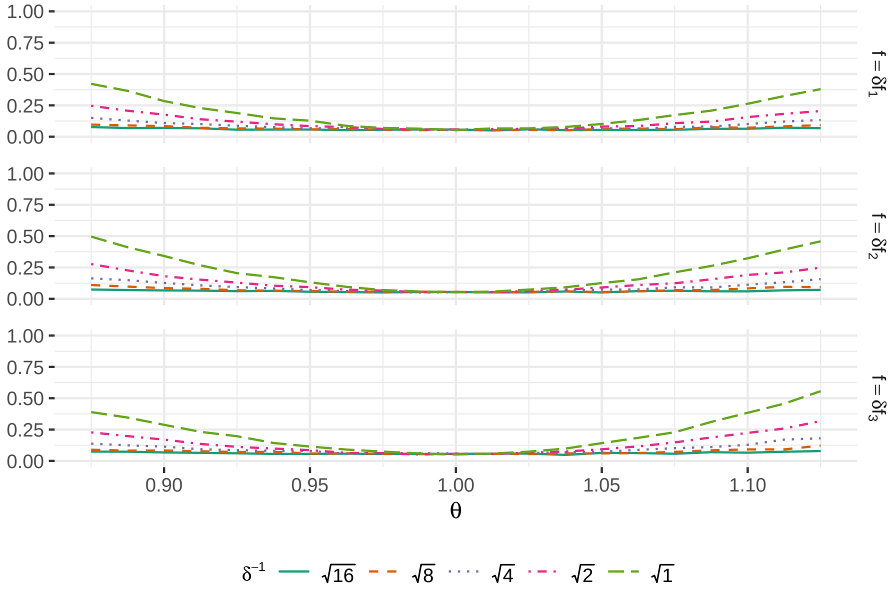
$\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\mathrm{Beta}(2,3)$, $\xi \sim \mathrm{Bernoulli}(1/2)$, $X_k \sim U(-1,1)$



Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Figure D.7: Homoskedastic SIM power curve, specification 3

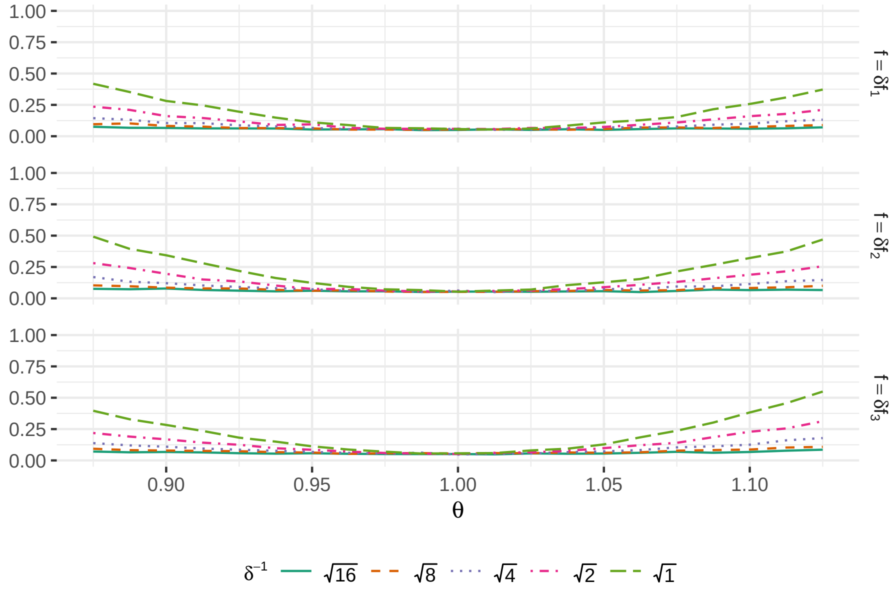
$\epsilon \sim \mathcal{N}(0,1)$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$



Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Figure D.8: Homoskedastic SIM power curve, specification 4

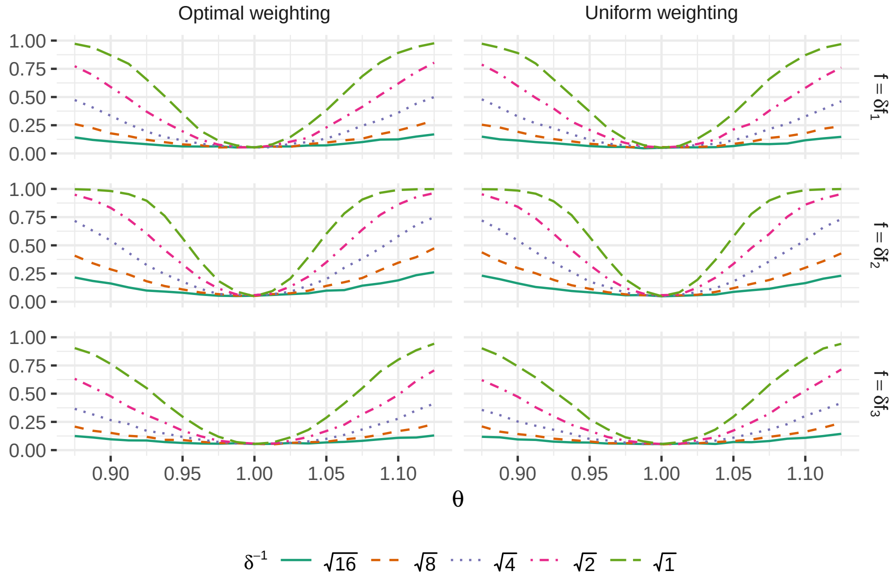
$\epsilon|\xi \sim \sqrt{5}(-1)^{\xi}\,\mathrm{Beta}(2,3)$, $\xi \sim \mathrm{Bernoulli}(1/2)$ $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1,1)$



Based on 5000 Monte carlo replications with a sample size of $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$, where the constants c_i $(i = 1,2,3)$ are chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null.

Figure D.9: Heteroskedastic SIM power curve, specification 1

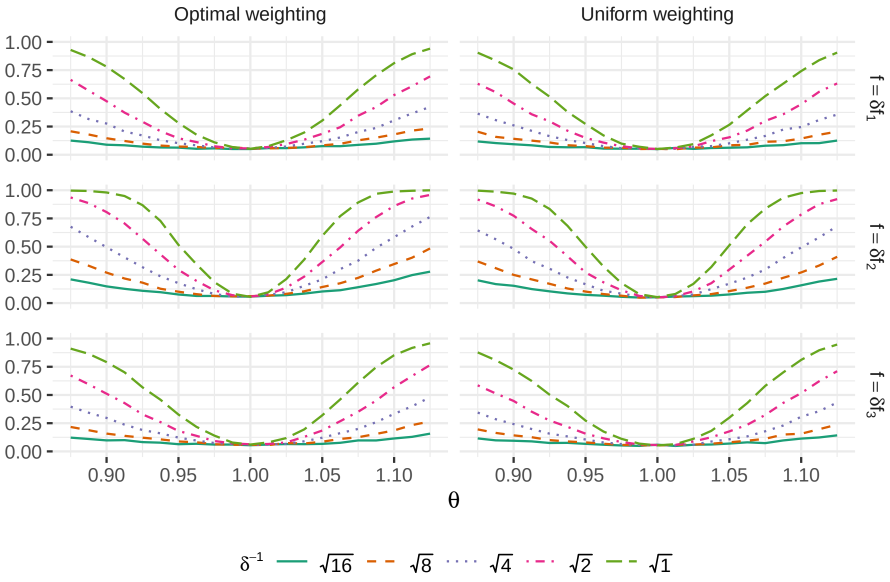
$\epsilon \sim \mathcal{N}(0, s_1 \log(2 + (X_1 + X_2\theta)^2))$, $X_k \sim U(-1,1)$



Based on 5000 Monte carlo replications; $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$; c_i $(i = 1,2,3)$ chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1,2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null. Uniform weighting: $\breve{\omega}(X) = 1$; Optimal weighting: $\breve{\omega}(X) = \omega(X)$.

Figure D.10: Heteroskedastic SIM power curve, specification 2

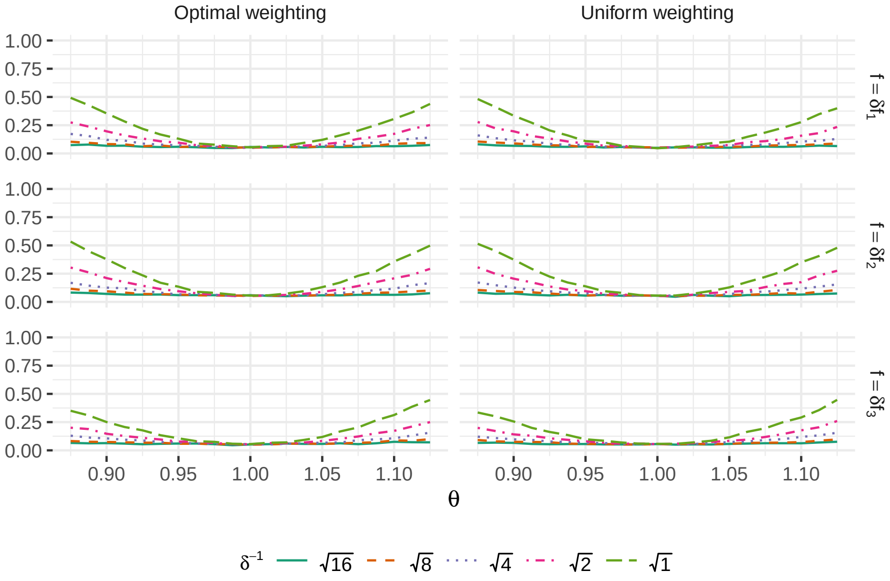
$$\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2)),\ X_k \sim U(-1, 1)$$



Based on 5000 Monte carlo replications; $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$; c_i $(i = 1, 2, 3)$ chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null. Uniform weighting: $\breve{\omega}(X) = 1$; Optimal weighting: $\breve{\omega}(X) = \omega(X)$.

Figure D.11: Heteroskedastic SIM power curve, specification 3

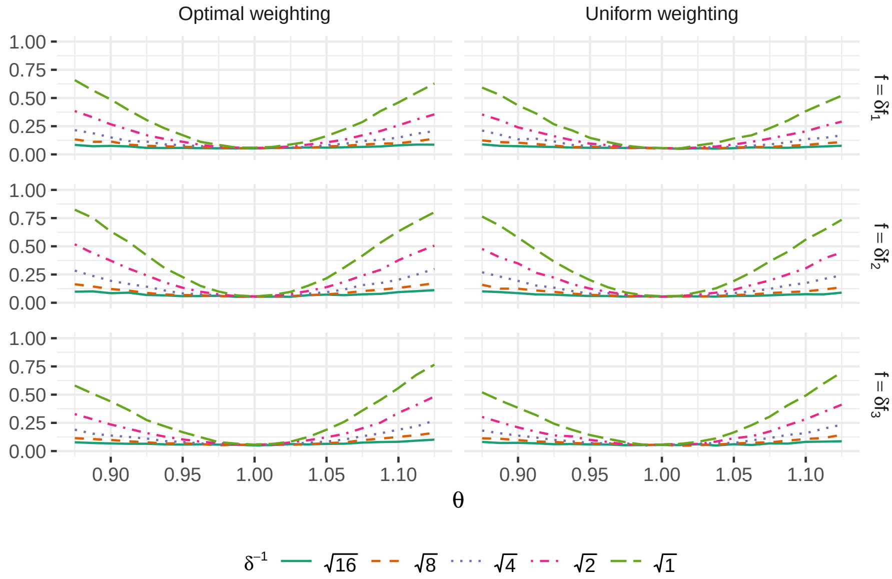
$$\epsilon \sim \mathcal{N}(0, s_1\log(2 + (X_1 + X_2\theta)^2)),\ X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8),\ Z_k \sim U(-1, 1)$$



Based on 5000 Monte carlo replications; $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$; c_i $(i = 1, 2, 3)$ chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null. Uniform weighting: $\breve{\omega}(X) = 1$; Optimal weighting: $\breve{\omega}(X) = \omega(X)$.

Figure D.12: Heteroskedastic SIM power curve, specification 4

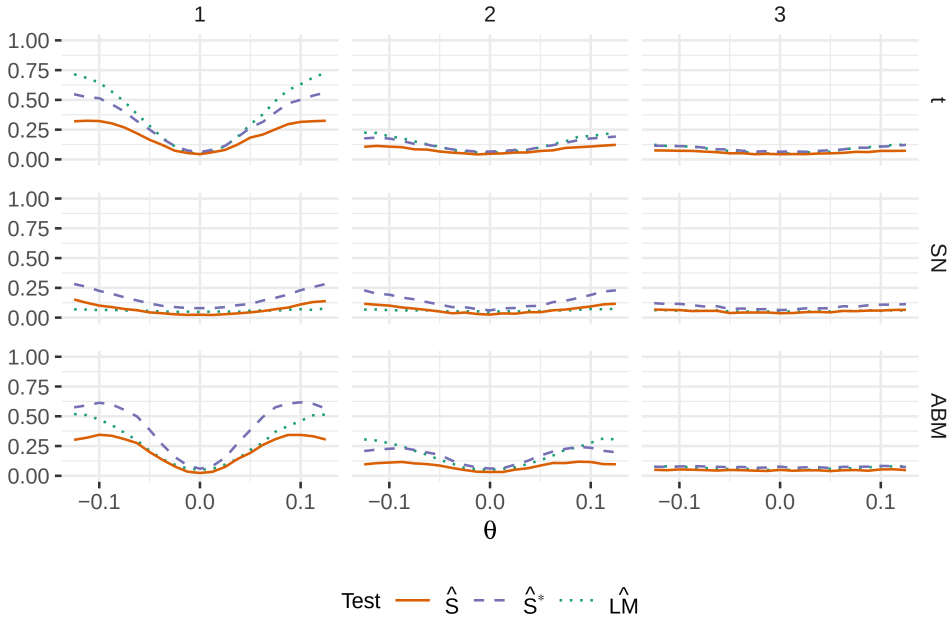
$\epsilon \sim \mathcal{N}(0, s_2(1 + 5\sin(X_1)^2))$, $X = (Z_1, 0.2Z_1 + 0.4Z_2 + 0.8)$, $Z_k \sim U(-1, 1)$



Based on 5000 Monte carlo replications; $n = 800$. $f_1(v) = c_1(1+\exp(-v))^{-1}$, $f_2(v) = c_2\exp(-v^2)$, $f_3(v) = c_3v^2$; c_i $(i = 1, 2, 3)$ chosen to ensure $\mathbb{V}(f_i(V_\theta)) = 4$ under the null. Similarly the constants s_i $(i = 1, 2)$ are chosen to ensure that $\mathbb{V}\epsilon = 1$ under the null. Uniform weighting: $\breve{\omega}(X) = 1$; Optimal weighting: $\breve{\omega}(X) = \omega(X)$.

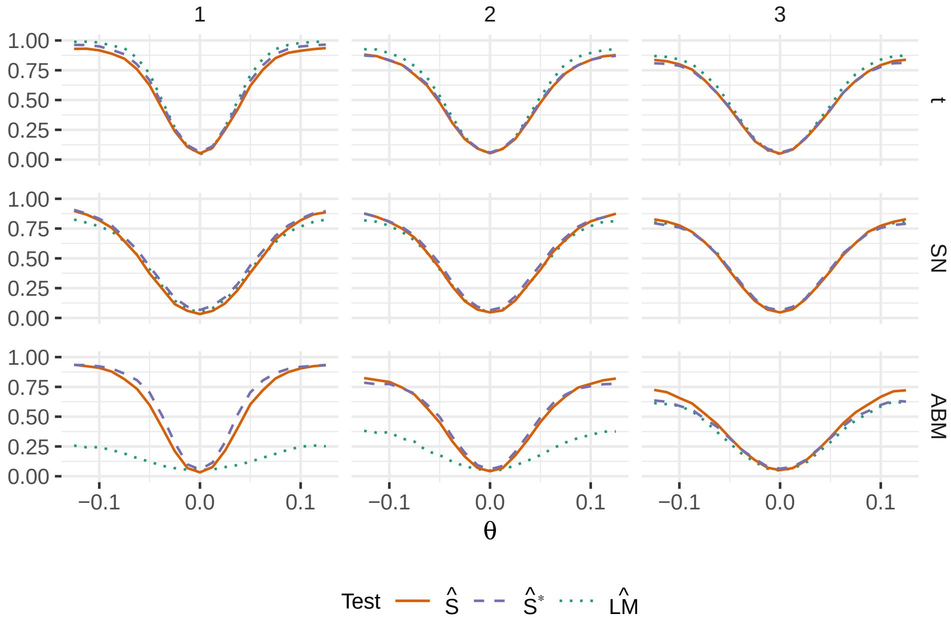
LSEM

Figure D.13: Power curves for LSEM (i), $\epsilon_1 \sim \mathcal{N}(0,1)$



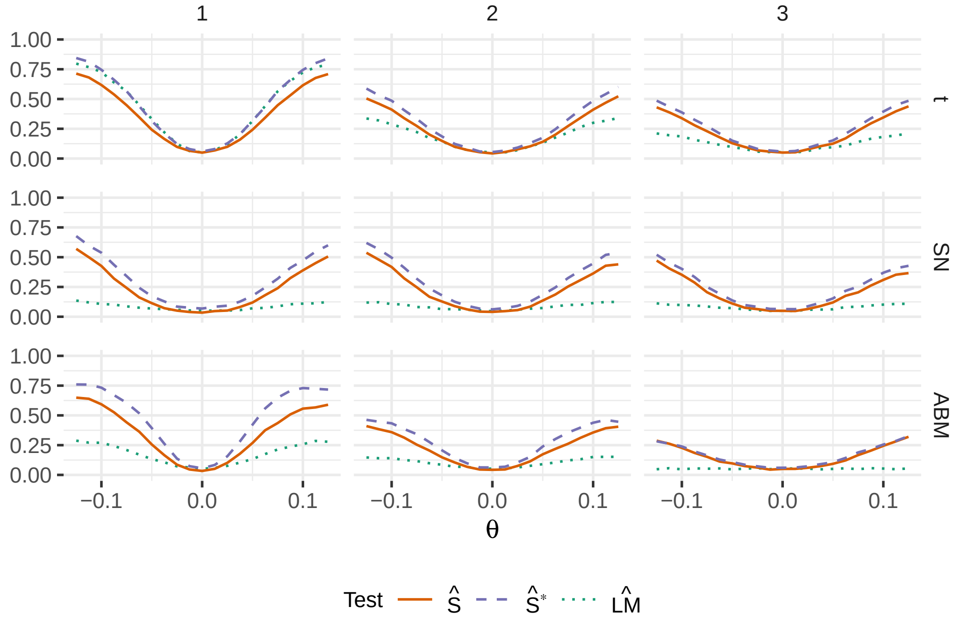
Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\text{LM}}$ denotes the LM test based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). The distribution for ϵ_2 in the $(i,j) - th$ panel has distribution i – j in table D.13.

Figure D.14: Power curves for LSEM (i), $\epsilon_1 \sim t'(5)$



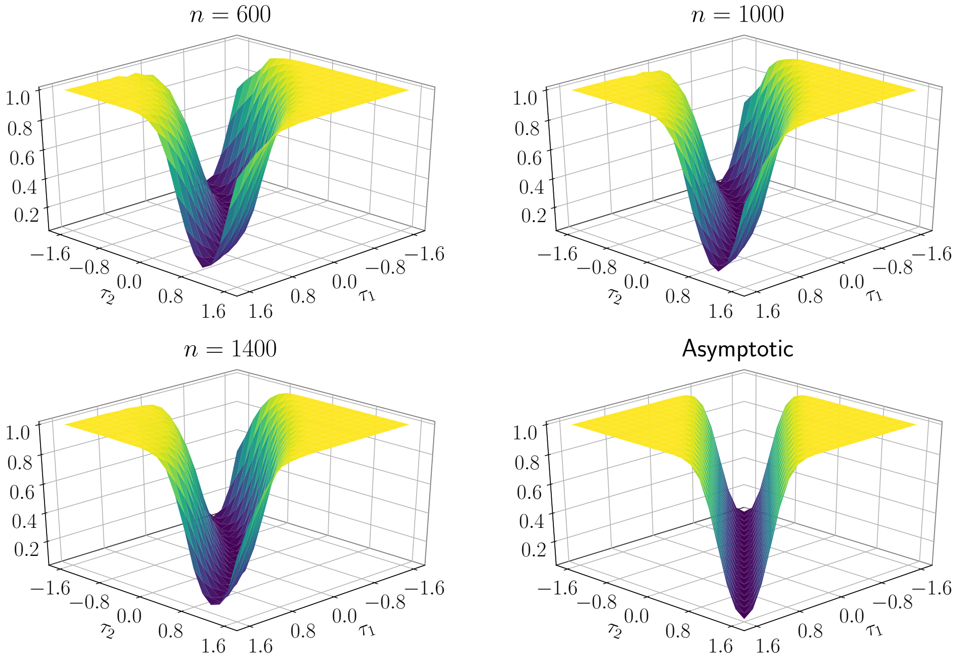
Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. $\hat{\text{LM}}$ denotes the LM test based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). The distribution for ϵ_2 in the $(i,j) - th$ panel has distribution i – j in table D.13.

Figure D.15: Power curves for LSEM (i), $\epsilon_1 \sim \mathcal{SN}'(0,1,4)$



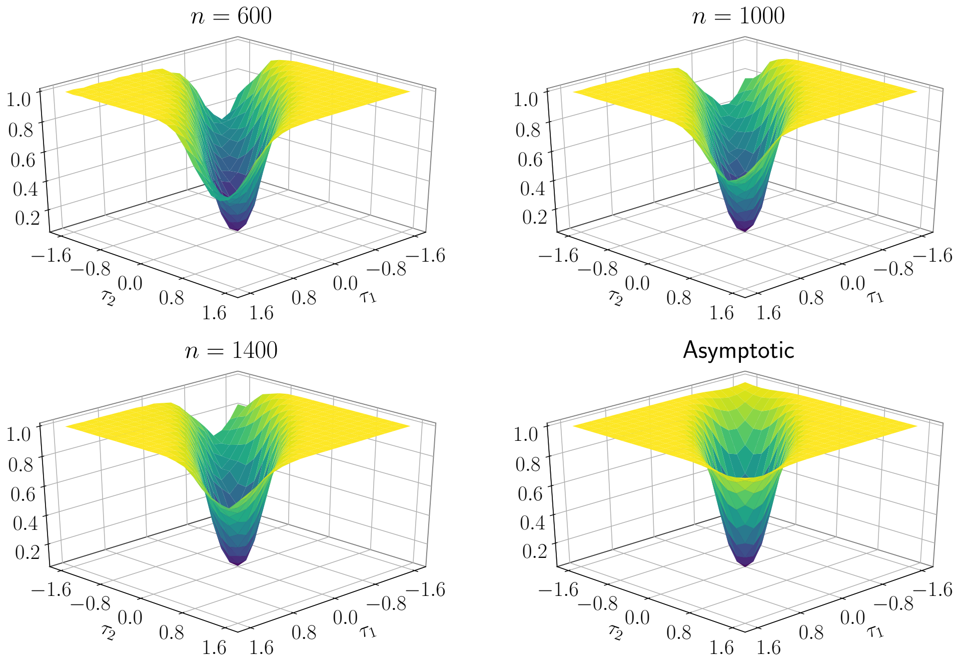
Based on 5000 Monte carlo replications. $\hat{S}$ is the efficient score test computed using OLS estimates of β; $\hat{S}^*$ is the efficient score test computed using 1-step updates from the OLS estimates. LM̂ denotes the LM test based on a psuedo-maximum likelihood estimator inspired by Gouriéroux et al. (2017). The distribution for ϵ_2 in the $(i, j) - th$ panel has distribution i – j in table D.13.

Figure D.16: Power surfaces for LSEM (ii), $\eta_1 \sim \mathcal{N}(0,1)$, $\eta_2 \sim \mathcal{N}(0,1)$



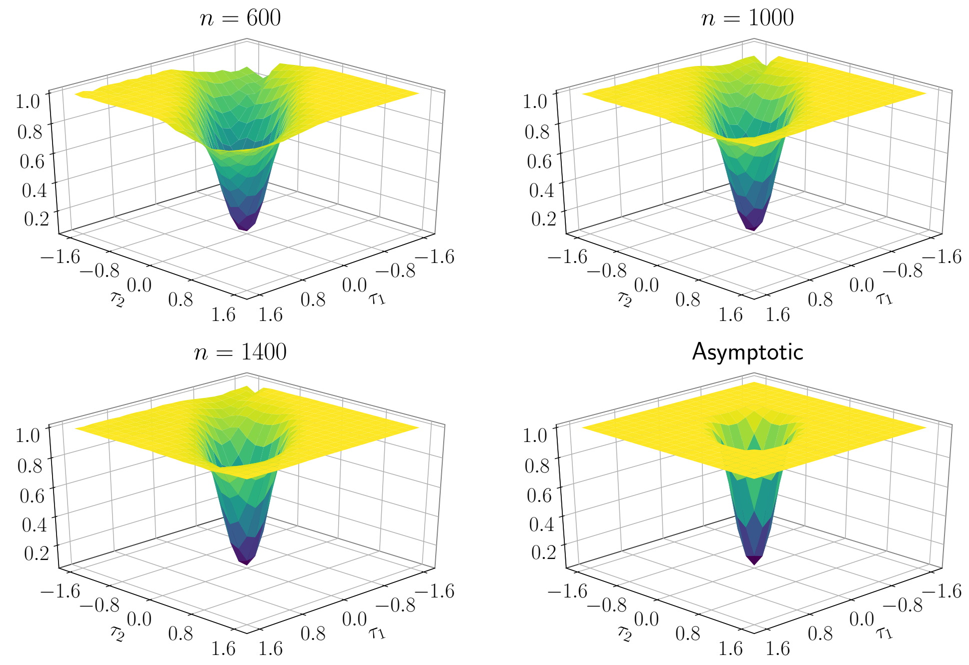
The bottom right panel depicts the asymptotic power surface based on (1.18) and (1.51) with $\theta = (a, b) = (1/2, 1/4)$ and $\sigma_1 = \sigma_2 = 1$. The top-left, top-right and bottom-left panels are Monte Carlo version based on 5000 replications of the efficient score test as described in section 1.5.5, with $n = 600, 1000, 1400$ respectively.

Figure D.17: Power surfaces for LSEM (ii), $\eta_1 \sim t'(5)$, $\eta_2 \sim t'(5)$



The bottom right panel depicts the asymptotic power surface based on (1.18) and (1.51) with $\theta = (a, b) = (1/2, 1/4)$ and $\sigma_1 = \sigma_2 = 1$. The top-left, top-right and bottom-left panels are Monte Carlo version based on 5000 replications of the efficient score test as described in section 1.5.5, with $n = 600, 1000, 1400$ respectively. $\eta_k \sim t'(5)$ indicates that each ϵ_k is drawn from a (standardised) t distribution with 5 degrees of freedom.

Figure D.18: Power surfaces for LSEM (ii), $\eta_1 \sim st'(5,2)$, $\eta_2 \sim st'(5,2)$



The bottom right panel depicts the asymptotic power surface based on (1.18) and (1.51) with $\theta = (a, b) = (1/2, 1/4)$ and $\sigma_1 = \sigma_2 = 1$. The top-left, top-right and bottom-left panels are Monte Carlo version based on 5000 replications of the efficient score test as described in section 1.5.5, with $n = 600, 1000, 1400$ respectively. $\eta_k \sim st'(5, 2)$ indicates that each ϵ_k is drawn from a (standardised) skew t distribution, as in Fernandez and Steel (1998) with 5 degrees of freedom and skewness parameter 2.

Chapter 2

Robust inference for non-Gaussian linear simultaneous equations models

This chapter was co-authored with Geert Mesters.

2.1. Introduction

The linear simultaneous equations model (LSEM) is a benchmark model used to analyze general equilibrium relationships in economics. It was formalized in its modern form by Haavelmo (1943, 1944), building on Frisch (1933) and Tinbergen (1939) among others. As is well known, without further restrictions, not all parameters of the LSEM can be uniquely identified from the first and second moments of the observed data series, see Dhrymes (1994) for an in-depth discussion.

Interestingly, this identification problem vanishes (up to permutation and scale) when the underlying structural shocks are independent and at most one of them follows a Gaussian distribution (e.g. Comon, 1994). This identification approach has a long history in the statistics and signal processing literatures where it is often referred to as independent components analysis, see Hyvärinen et al. (2001) for a textbook treatment. More recently, the econometrics literature has started investigating this approach and developing the corresponding methodology for conducting inference on the parameters of various LSEMs based on non-Gaussian identification.[1]

[1] See for instance: Lanne and Lütkepohl (2010), Moneta et al. (2013), Lanne et al. (2017), Maxand (2018), Lanne and Luoto (2021), Gouriéroux et al. (2017, 2019), Tank et al. (2019), Herwartz (2019), Herwartz et al. (2019), Bekaert et al. (2019, 2020), Fiorentini and Sentana (2022), Velasco (2020), Guay (2020), Moneta and Pallante (2020), Drautzburg and Wright (2021), Sims (2021) and Davis and Ng (2022).

Unfortunately, if in the true data generating process multiple structural shocks follow a Gaussian distribution some structural parameters may be under- or un-identified and standard inference methods that aim to exploit non-Gaussian distributions may fail to control size. Moreover, as is typical in models with points of identification failure, such behavior is also observed if the true distributions of the shocks are sufficiently to close to Gaussianity, relative to the sampling variation. Intuitively, in such *weakly non-Gaussian* settings local identification deteriorates leading to coverage distortions when using standard inference methods, such as maximum likelihood and moment methods.

Similar (weak) identification problems occur in many other econometric models, e.g. instrumental variable models, nonlinear regression models and many others, see Andrews and Cheng (2012, 2013) for numerous examples. The key difference between this existing literature and the non-Gaussian LSEM is that, in the latter, the parameters responsible for the possible identification failure are density functions, i.e. infinite dimensional parameters. Therefore, whilst conceptually the identification problem is the same, providing robust inferential methods requires a new approach which is capable of handling identification failure caused by infinite dimensional nuisance parameters.

To this extent, this paper develops a robust approach for conducting inference in LSEMs that is inspired by the identification robust methods developed in econometrics (e.g. Stock and Wright, 2000; Kleibergen, 2005; Andrews and Mikusheva, 2015) and the general semiparametric statistical theory that is discussed in Bickel et al. (1998) and van der Vaart (2002). In brief, we treat the LSEM as a semiparametric model, where the densities of the independent structural shocks are treated non-parametrically, and we construct confidence bands for the possibly unidentified structural parameters of interest by inverting semiparametric score tests. The approach efficiently exploits non-Gaussianity when it is present in the data and yields correct coverage regardless of the true distribution of the shocks.

Intuitively, the efficient score test that we propose is the semi-parametric analog of Neyman's $C(\alpha)$ test (e.g. Neyman, 1979; Hall and Mathiason, 1990). In the conventional $C(\alpha)$ test the scores of the parameter of interest are orthogonalized with respect to the scores of the *finite dimensional* nuisance parameters. In our setting the nuisance parameter includes the densities of the shocks, i.e. an *infinite dimensional* parameter. While such nuisance functions result in the orthogonal projection being more technically demanding to derive, the main idea of Neyman (1979) continues to apply.

We evaluate the finite sample performance of the semiparametric score test in a large simulation study. This shows that regardless of how close the errors are to the Gaussian distribution our test is correctly sized. In contrast, tests that are based on the sampling variation of (pseudo)-maximum likelihood or GMM estimators have large size distortions in weakly non-Gaussian settings. Further, for moderate sample sizes the power of the semiparametric test is comparable to the parametric score test that relies on knowing the functional form of the density. When the parametric density of the (pseudo)-maximum likelihood score test is misspecified the semi-parametric test is always found to be preferable.

To showcase the empirical value of our methodology we consider the estimation of the coefficients in a production function (e.g. Marschak and Andrews, 1944; Hoch, 1958; Olley and Pakes, 1996; Leeb and Pötscher, 2003; Ackerberg et al., 2015). In contrast to the more recent literature, we explicitly model the correlation between the error term and the production function inputs; capital and labor (e.g. Hoch, 1958), and we exploit non-Gaussianity to identify the product function coefficients. We adopt this strategy for a large sample of manufacturing firms.

Overall, we find that this approach is able to accurately pin down the production function coefficients. We estimate the coefficient for labor between 0.4 and 0.8 and the coefficient for capital is between 0.2 and 0.5. These estimates are (i) robust across a variety of model specifications and (ii) vastly different from standard OLS estimates, potentially indicating a strongly endogenous relationship.

Throughout this paper we retain the assumption that the structural shocks are independent which may not be the case in practice, see the discussions in Matteson and Tsay (2017), Davis and Ng (2022) and Montiel Olea et al. (2022). Therefore, in our empirical study we test the independence of the structural shocks following the approach of Matteson and Tsay (2017) and find that for our empirical application we cannot reject the independence assumption.

The remainder of this paper is organized as follows. In the next section we provide a simple example that illustrates the identification problem and intuitively discusses our solution. Section 2.3 presents the main LSEM model and provides the implementation details for the efficient score test. Section 2.4 discusses the main theoretical results including the required assumptions. Sections 2.5 and 2.6 summarize the results from the simulation and empirical studies. Section 2.7 concludes. Unless otherwise mentioned all proofs are provided in the Appendix.

2.2. Illustrative example

In this section we use a simple example to illustrate: (i) the identification problem in LSEMs, (ii) why conventional inference methods suffer from size distortions when the structural shocks have densities close to Gaussian and (iii) how our proposed approach circumvents such distortions.

THE IDENTIFICATION PROBLEM Consider the simple bi-variate model

$$Y_i = R'\epsilon_i \ , \qquad i = 1, \ldots, n \ , \tag{2.1}$$

where Y_i is a vector of observable variables, R is rotation matrix (i.e. $R'R = I_2$) and ϵ_i is a vector with independent structural shocks $\epsilon_{i,k}$, for $k = 1, 2$, that have mean zero, unit variance and common density η. For concreteness, we will parameterize the rotation matrix as follows

$$R = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} , \tag{2.2}$$

where $\alpha \in [0, 2\pi]$ and we let α_0 denote the true parameter.[2]

Model (2.1) has two parameters: the parameter of interest α and the infinite dimensional nuisance parameter η. Suppose for now that η is known and let the log likelihood function for Y_i be denoted by $\ell_\alpha(\cdot)$. α is locally identified if the expected score of $\ell_\alpha(Y_i)$ with respect to α is non-zero for all $\alpha \neq \alpha_0$ in a neighborhood of α_0.

Whether local identification occurs turns out to depend crucially on η. To illustrate, consider the case where η is equal to the Gaussian density. Since ϵ_i is normalized we have

$$\mathbb{E}\ell_\alpha(Y_i) \propto -\frac{1}{2}\mathbb{E}(RY_i)'(RY_i) = -1$$

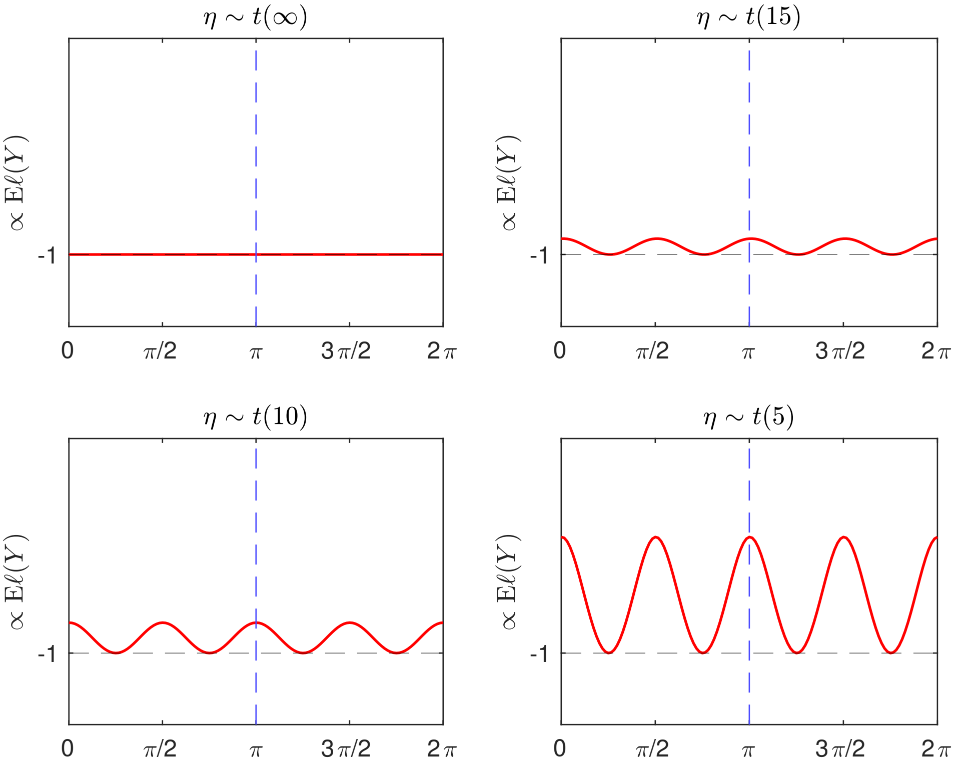
and hence the expected loglikelihood takes the same value irrespective of α. This is plotted in the top left panel of Figure 2.1, where we show the expected likelihood $\mathbb{E}\ell_\alpha(Y_i)$ as a function of α with $\alpha_0 = \pi$ as the true parameter (an arbitrary choice). This illustrates the standard identification problem in linear simultaneous equations models: without additional identifying restrictions, the impact effects of the structural shocks are not identifiable when the structural shocks follow a Gaussian distribution.

The other plots in Figure 2.1 show that this is no longer the case when we move away from the Gaussian distribution. In each case the expected gradient becomes non-zero at values $\alpha \neq \alpha_0$ in the vicinity of α_0, i.e. local identification occurs. While for the Student's t distribution with five degrees of freedom (i.e. $t(5)$) the change in the value of the expected likelihood is substantial it is easy to see that for more modest deviations from Gaussianity (e.g. $t(15)$) the difference is less pronounced. Further, note that non-Gaussian densities do not imply that α is globally identified, instead identification is only up to permutation and sign of the shocks.

FINITE SAMPLE SIZE DISTORTIONS In population α is always locally identified when all but one component of η is non-Gaussian (Comon, 1994), but this is not sufficient for good

[2] Note that in general a researcher may consider $Y_i = \Sigma^{1/2}R'\epsilon_i$, where $\Sigma^{1/2}$ is lower triangular. However, the elements of $\Sigma^{1/2}$ can be identified from the variance of Y_i and pose no difficulty. Therefore we set the variance of Y_i to unity and exclude $\Sigma^{1/2}$ for simplicity.

Figure 2.1: (Weak) Non-Gaussian Identification



Notes: In the figure we show the expected log likelihood (red line) as a function of α (the true value is $\alpha_0 = \pi/4$).

performance of standard testing procedures in finite samples. In particular, if the structural errors are too close to Gaussian, the available identifying information may be small relative to the sampling variability. Standard asymptotic approximations are not reliable in this setting and, as a result, testing procedures based on these approximations may fail to provide reliable inference.

To illustrate how the density η affects standard inference methods in finite sample consider figure 2.2 which depicts the finite sample distribution of the t-statistic for the hypothesis $H_0 : \alpha = \alpha_0$, based on the maximum likelihood estimator under the assumption that η is known. The blue dashed lines show the $\mathcal{N}(0, 1)$ density. As can clearly be seen in this figure, the quality of the approximation provided by the standard Normal depends crucially on the underlying density, η. For a given sample size, the approximation deteriorates substantially the closer η is to a standard Gaussian density.

This deterioration results in poor size control of standard tests. Table 2.1 shows the empirical rejection frequencies for three standard tests in the same setting: Wald (W), likelihood ratio (LR) and Lagrange multiplier (LM) (or score) tests, all computed under the assumption that η is known. Specifically we drew 5000 samples $\{Y_i\}_{i=1}^n$ from model (2.1) for different η's using different sample sizes $n = 250, 500, 750$. The empirical rejection frequencies correspond to the test for $H_0 : \alpha = \alpha_0$ with nominal size $a = 0.05$, where the critical values are based on the standard $\chi^2(1)$ asymptotic approximation.

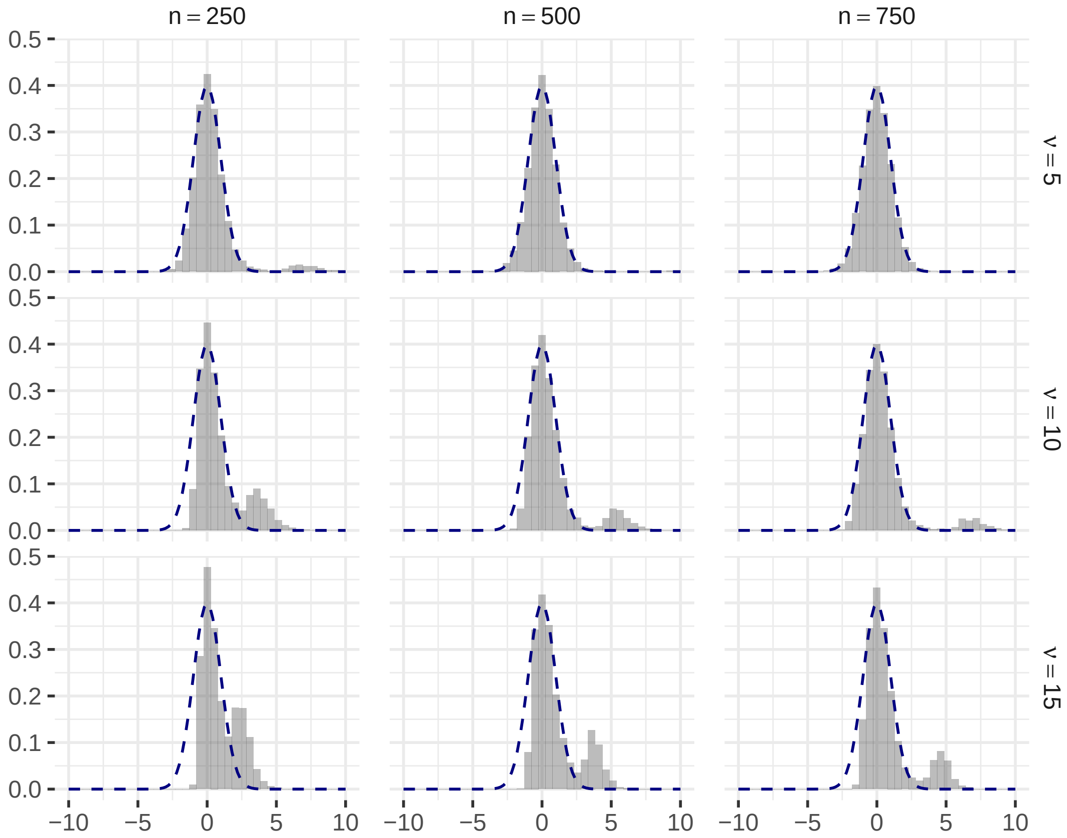
We find that the Wald test is severely size distorted for η close to Gaussian; in view of the poor quality of asymptotic approximation depicted in Figure 2.2 this is not surprising. As η gets closer to Gaussianity, the likelihood ratio test starts to under-reject as when α is poorly identified the likelihood values are very similar. Both of these tests are based on estimates of α and, in weakly identified settings, such estimates will be inaccurate. In contrast, the score test (LM) shows correct size as it fixes $\alpha = \alpha_0$ under the null and α does not need to be (well) identified for this test to be correctly sized.

Table 2.1: Empirical rejection frequencies for ML tests close to Gaussianity

	t(15)			t(10)			t(5)		
n	W	LM	LR	W	LM	LR	W	LM	LR
250	25.26	4.42	3.74	20.56	4.24	4.04	8.88	4.84	4.08
500	21.76	4.54	4.52	13.10	4.38	3.60	6.38	4.42	4.92
750	17.12	4.96	3.94	9.90	4.88	3.42	6.12	5.28	5.64

Notes: The table shows the empirical rejection frequencies for the three maximum likelihood tests, under the assumption that η is known and based on 5000 Monte Carlo replications for the baseline model $Y_i = R'\epsilon_i$. The test has nominal size $a = 0.05$.

Figure 2.2: Poor asymptotic approximation close to Gaussianity



Notes: In the figure we show the finite sample distribution of the t-statistic based on the maximum likelihood estimator of α (the true value is $\alpha_0 = \pi/4$) for different sample sizes (n) and different degrees of freedom (ν) in the (standardised) t distribution, all based on 5000 replications.

TOWARDS A SEMI-PARAMETRIC SCORE TEST Now in practice, η will be unknown and needs to be estimated. To build up to our semi-parametric approach, consider first the case where η is known up to a finite dimensional parameter vector, say β (for example β may include the degrees of freedom of the Student's t distribution). For this case Neyman (1979) proposed a convenient extension of the standard score test, that amounts to first orthogonalizing the scores for α with respect to the scores for β and then computing a quadratic form of the score statistic. To illustrate let $\dot{\ell}(Y_i) = (\dot{\ell}_\alpha(Y_i), \dot{\ell}_\beta(Y_i))'$, $\dot{\ell}_\alpha(Y_i) = \nabla_\alpha \ell(Y_i)$, $\dot{\ell}_\beta(Y_i) = \nabla_\beta \ell(Y_i)$ and $\hat{I} = \frac{1}{n}\sum_{i=1}^n \dot{\ell}(Y_i)\dot{\ell}(Y_i)'$, denote the score and information matrix for α and β. Neyman's $C(\alpha)$ test statistic is given by

$$C(\alpha) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}(Y_i)\right)' \hat{\mathcal{I}}^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}(Y_i)\right) ,$$

with

$$\hat{\kappa}(Y_i) = \dot{\ell}_\alpha - \hat{I}_{\alpha\beta}\hat{I}_{\beta\beta}^{-1}\dot{\ell}_\beta \qquad \text{and} \qquad \hat{\mathcal{I}} = \hat{I}_{\alpha\alpha} - \hat{I}_{\alpha\beta}\hat{I}_{\beta\beta}^{-1}\hat{I}_{\beta\alpha} ,$$

where $\hat{I}_{\cdot\cdot}$ denote the corresponding blocks of $\hat{I}$.[3] The (estimated) orthogonalized scores $\hat{\kappa}(\cdot)$ are often referred to as the (estimates of the) efficient scores and $\hat{\mathcal{I}}$ is the corresponding (estimate of the) efficient information matrix. When evaluating $C(\alpha)$ at $\alpha = \alpha_0$ and $\hat{\beta}$, some $\sqrt{n}$ consistent estimate for β, this statistic will converge to a standard χ^2 limit under the null provided that $\hat{\mathcal{I}}$ is invertible.[4] Tests based on $C(\alpha)$ retain correct size regardless whether α is well identified as α is fixed under H_0, making them attractive for settings where identification failure due to finite dimensional nuisance parameters is a concern (e.g. Andrews and Mikusheva, 2015).

In the present paper, we will not impose that the parametric form of η is known up to finite dimensional parameters but instead treat η non-parametrically. Despite this change, our approach is similar to that sketched above. We will first orthogonalize the score for α with respect to the scores for η and obtain a semi-parametric analog of the conventional Neyman $C(\alpha)$ test. This requires technical adjustments as the scores with respect to η need to be defined differently and the projection with respect to η scores requires more care. For this we follow the semi-parametric literature as outlined in the textbooks of Bickel et al. (1998) and van der Vaart (2002).

[3] This is numerically equivalent to the "usual" score test provided the nuisance parameter β is estimated by (restricted) maximum likelihood under the null hypothesis (Kocherlakota and Kocherlakota, 1991).

[4] In our general framework below we explicitly allow $\hat{\mathcal{I}}$ to be singular and rely on an eigenvalue truncated generalized inverse, see also Andrews (1987), Lütkepohl and Burda (1997) and Andrews and Guggenberger (2019).

2.3. Robust inference for LSEMs

In this section we discuss the implementation of the semi-parametric score test for a general class of linear simultaneous equations models.

2.3.1. General model and objectives

We consider the linear simultaneous equations model for a random sample of the $K \times 1$ endogenous variables Y_i, the $d \times 1$ exogenous variables $X_i = (1, \tilde{X}_i')'$ and the $K \times 1$ structural shocks ϵ_i. Specifically,

$$Y_i = BX_i + A^{-1}\epsilon_i \ , \qquad i = 1, \dots, n \ , \tag{2.3}$$

where the matrices B and A^{-1} map the explanatory variables and the structural shocks to the endogenous variables. The density functions of the components of $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{iK})'$ are denoted by $(\eta_1, \dots, \eta_K)$ and the density of $\tilde{X}_i$ is given by η_0. We set $\eta = (\eta_0, \eta_1, \dots, \eta_K)$.

As illustrated in the previous section, depending on the shapes of $\eta_1, \dots, \eta_K$ we may not be able to identify all parameters in A. To model this we let $A = A(\alpha, \sigma)$, where $A(\alpha, \sigma)$ is a function of the possibly unidentified parameters α and parameters σ which can be always identified from the variance of $Y_i - BX_i$. We let $\alpha \in \mathcal{A} \subset \mathbb{R}^{L_\alpha}$ and set $\beta = (\sigma, b) \in \mathcal{B} \subset \mathbb{R}^{L_\sigma} \times \mathbb{R}^{L_b} = \mathbb{R}^{L_\beta}$, with $b = \operatorname{vec}(B)$. The following two examples illustrate possible parametrizations for $A(\alpha, \sigma)$ that are of practical interest.

Example 1 (Rotation matrix)**.** *Let $A(\alpha, \sigma)^{-1} = \Sigma^{1/2}R'$, where $\Sigma^{1/2}$ is lower triangular and R is a rotation matrix. In this setting we can take $\sigma = \operatorname{vech}(\Sigma^{1/2})$ and α parametrizes R using the trigonometric transformation (as in Section 2.2) or the Cayley or exponential transformation of a skew-symmetric matrix (e.g. Gouriéroux et al., 2017; Magnus et al., 2020).*

Example 2 (Supply and demand)**.** *For $K = 2$ let Y_{i1} denote the quantity of some good and Y_{i2} its price. A simple model (omitting covariates for convenience) is given by*

$$\begin{aligned} Y_{i1}^d &= aY_{i2} + \sigma_1\epsilon_{i1} \qquad \textit{(demand)} \\ Y_{i1}^s &= bY_{i2} + \sigma_2\epsilon_{i2} \qquad \textit{(supply)} \end{aligned}$$

where ϵ_{i1} and ϵ_{i2} are independent demand and supply shocks, and in equilibrium we have $Y_{i1}^d = Y_{i2}^s$. We can accommodate this set up by letting $\alpha = (a, b)$, $\beta = (\sigma_1, \sigma_2)$ and defining the mapping $A(\alpha, \sigma)$ according to

$$A(\alpha, \sigma) = \begin{bmatrix} \sigma_1^{-1} & 0 \\ 0 & \sigma_2^{-1} \end{bmatrix} \begin{bmatrix} 1 & -a \\ 1 & -b \end{bmatrix} .$$

In the remainder we leave the precise mapping $A(\alpha, \sigma)$ unspecified, but we will require that it satisfies certain smoothness conditions.

The general LSEM (2.3) depends on the triplet of parameters $\theta = (\alpha, \beta, \eta)$, which includes the possibly unidentified parameters α, the finite dimensional nuisance parameters $\beta = (\sigma, b)$ and the infinite dimensional nuisance parameters η. We will refer to β as nuisance parameters as our main interest is in conducting inference on α, but clearly β could also be an object of interest. To conduct inference on α without making a priori assumptions on the identification strength of α, i.e. without assuming that sufficiently many η_k's are non-Gaussian, we consider hypothesis tests of the form

$$H_0 : \alpha = \alpha_0 \qquad \text{against} \qquad H_1 : \alpha \neq \alpha_0 \; . \tag{2.4}$$

Such test statistics can then be inverted to yield confidence intervals for α with correct coverage.

The problem formulation reflects that we aim for a procedure that is valid for all densities η_k, for $k = 1, \dots, K$, Gaussian or not. A related set-up is found in Risk et al. (2019) and Jin et al. (2019) who assume that the structural shocks can be separated into *exactly* Gaussian and non-Gaussian shocks. We do not impose such structure, but we note that if indeed shocks can be separated in this way our approach will remain valid, but likely less efficient when compared to Risk et al. (2019).

2.3.2. Efficient score test for LSEMs

Next, we provide a step by step implementation guide for the semi-parametric score test, with the theoretical justification postponed to the next section.

EFFICIENT SCORE AND INFORMATION MATRIX ESTIMATES As a first step, let $\hat{\ell}_\gamma(V_i)$ denote the estimates for efficient scores of the finite dimensional parameters $\gamma = (\alpha, \beta)$ of the LSEM (2.3) evaluated at $V_i = Y_i - BX_i$ and γ. Intuitively, these are the estimates for the scores of the parameters γ that are obtained after projecting out the infinite dimensional nuisance parameter η. As we show in the appendix, consistent estimates for the components of $\hat{\ell}_\gamma(V_i)$ are given by

$$\hat{\ell}_\gamma(V_i) = \begin{bmatrix} \hat{\ell}_{\gamma,\alpha}(V_i) \\ \hat{\ell}_{\gamma,\beta}(V_i) \end{bmatrix} = \begin{bmatrix} \{\hat{\ell}_{\gamma,\alpha_l}(V_i)\}_{l=1}^{L_\alpha} \\ \{\hat{\ell}_{\gamma,\beta_l}(V_i)\}_{l=1}^{L_\beta} \end{bmatrix} \quad \text{with} \quad \hat{\ell}_{\gamma,\beta}(V_i) = \begin{bmatrix} \hat{\ell}_{\gamma,\sigma}(V_i) \\ \hat{\ell}_{\gamma,b}(V_i) \end{bmatrix} = \begin{bmatrix} \{\hat{\ell}_{\gamma,\sigma_l}(V_i)\}_{l=1}^{L_\sigma} \\ \{\hat{\ell}_{\gamma,b_l}(V_i)\}_{l=1}^{L_b} \end{bmatrix},$$

and

$$\begin{aligned}
\hat{\ell}_{\gamma,\alpha_l}(V_i) &= \sum_{j,k=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j}\hat{\phi}_k(A_{k\bullet}V_i)A_{j\bullet}V_i + \sum_{k=1}^{K}\zeta^{\alpha}_{l,k,k}\left[\hat{\tau}_{k,1}A_{k\bullet}V_i + \hat{\tau}_{k,2}\kappa(A_{k\bullet}V_i)\right] \\
\hat{\ell}_{\gamma,\sigma_l}(V_i) &= \sum_{j,k=1,j\neq k}^{K} \zeta^{\sigma}_{l,k,j}\hat{\phi}_k(A_{k\bullet}V_i)A_{j\bullet}V_i + \sum_{k=1}^{K}\zeta^{\sigma}_{l,k,k}\left[\hat{\tau}_{k,1}A_{k\bullet}V_i + \hat{\tau}_{k,2}\kappa(A_{k\bullet}V_i)\right] \\
\hat{\ell}_{\gamma,b_l}(V_i) &= \sum_{k=1}^{K}[-A_{k\bullet}D_{b,l}][(X_i-\bar{X}_n)\hat{\phi}_k(A_{k\bullet}V_i) - \bar{X}_n(\hat{\varsigma}_{k,1}A_{k\bullet}V_i + \hat{\varsigma}_{k,2}\kappa(A_{k\bullet}V_i))]
\end{aligned} \tag{2.5}$$

where $A_{k\bullet}$ denotes the kth row of A, $\kappa(z) = 1 - z^2$, $\zeta^{\alpha}_{l,k,j} := [D_{\alpha,l}]_{k\bullet}A^{-1}_{\bullet j}$, $\zeta^{\sigma}_{l,k,j} := [D_{\sigma,l}]_{k\bullet}A^{-1}_{\bullet j}$, $D_{\alpha,l} = \partial A(\alpha,\sigma)/\partial\alpha_l$, $D_{\sigma,l} = \partial A(\alpha,\sigma)/\partial\sigma_l$, $D_{b_l} = \partial B/\partial b_l$ and $\bar{X}_n = n^{-1}\sum_{i=1}^{n}X_i$. The coefficients $\hat{\tau}_k = (\hat{\tau}_{k,1},\hat{\tau}_{k,2})'$ and $\hat{\varsigma}_k = (\hat{\varsigma}_{k,1},\hat{\varsigma}_{k,2})'$ are given, for $k = 1,\ldots,K$, by

$$\hat{\tau}_k = \hat{M}_k^{-1}\begin{pmatrix}0\\-2\end{pmatrix},\ \hat{\varsigma}_k = \hat{M}_k^{-1}\begin{pmatrix}1\\0\end{pmatrix},\ \hat{M}_k = \begin{pmatrix}1 & \frac{1}{n}\sum_{i=1}^{n}(A_{k\bullet}V_i)^3 \\ \frac{1}{n}\sum_{i=1}^{n}(A_{k\bullet}V_i)^3 & \frac{1}{n}\sum_{i=1}^{n}(A_{k\bullet}V_i)^4 - 1\end{pmatrix}. \tag{2.6}$$

Finally, the efficient score estimates (2.5) depend on $\hat{\phi}_k(\cdot)$: the estimate for the log density score $\phi_k(x) = \partial\eta_k(x)/\partial x$. Such estimates can be obtained in different ways and our preferred approach is based on using B-splines as in Jin (1992) and Chen and Bickel (2006). We can define such estimates as

$$\hat{\phi}_k(x) = \hat{\gamma}_k' b_k(x) \qquad \text{with} \quad \hat{\gamma}_k = -\left[\sum_{i=1}^{n} b_k(A_{k\bullet}V_{k,i})b_k(A_{k\bullet}V_{k,i})'\right]^{-1}\sum_{i=1}^{n}c_k(A_{k\bullet}V_{k,i}), \tag{2.7}$$

where $b_k(x) = (b_{k,1}(x),\ldots,b_{k,B_k}(x))'$ is a collection of B_k cubic B-splines and $c_k(x) = (c_{k,1}(x),\ldots,c_{k,B_k}(x))'$ are their derivatives: $c_{k,i}(x) = \frac{\mathrm{d}b_{k,i}(x)}{\mathrm{d}x}$ for each $i = 1,\ldots,B_k$, see de Boor (2001) for more details on B-splines. In practice we rely on equally spaced knots with upper and lower end points taken to be the 95th and 5th percentile of the samples $\{A_{k\bullet}V_{i,k}\}_{i=1}^{n}$ adjusted by $\log(\log(n))$. We use $B_k = 6$ splines in our main simulations below and investigate the sensitivity of this choice.

Given the estimates of the efficient scores we estimate the efficient information matrix, which is the variance matrix of the efficient score function, as

$$\hat{I}_\gamma = \frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_\gamma(V_i)\hat{\ell}_\gamma(V_i)' \qquad \text{with partitioning} \quad \hat{I}_\gamma = \begin{bmatrix}\hat{I}_{\gamma,\alpha\alpha} & \hat{I}_{\gamma,\alpha\beta} \\ \hat{I}_{\gamma,\beta\alpha} & \hat{I}_{\gamma,\beta\beta}\end{bmatrix}. \tag{2.8}$$

EFFICIENT SCORE STATISTIC To compute the efficient semi-parametric score statistic for testing $H_0 : \alpha = \alpha_0$ we first orthogonalize the efficient scores for α with respect to those for $\beta = (\sigma, b)$. Since, β is finite dimensional the estimates of the resulting orthogonalized scores and information for α are given by

$$\hat{\kappa}_\gamma(V_i) = \hat{\ell}_{\gamma,\alpha}(V_i) - \hat{I}_{\gamma,\alpha\beta}\hat{I}^{-1}_{\gamma,\beta\beta}\hat{\ell}_{\gamma,\beta}(V_i) \qquad \text{and} \qquad \hat{\mathcal{I}}_\gamma = \hat{I}_{\gamma,\alpha\alpha} - \hat{I}_{\gamma,\alpha\beta}\hat{I}^{-1}_{\gamma,\beta\beta}\hat{I}_{\gamma,\beta\alpha} \,. \tag{2.9}$$

These are estimates of the population efficient score and efficient information matrix. Importantly, the latter may not be positive definite in our setting. For instance, when the densities η_k correspond to the Gaussian density, $\mathcal{I}_\gamma$ is singular, see Lemma D.1 in the supplementary material.

With $\hat{\kappa}_\gamma(V_i)$ and $\hat{\mathcal{I}}_\gamma$ we can define the efficient score statistic for the LSEM model as function of $\gamma = (\alpha, \beta)$ and $V_i = Y_i - BX_i$ by

$$\hat{S}_\gamma = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_\gamma(V_i)\right)' \hat{\mathcal{I}}_\gamma^{t,\dagger} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_\gamma(V_i)\right) \,, \tag{2.10}$$

where $\hat{\mathcal{I}}_\gamma^{t,\dagger}$ denotes the generalized inverse of the eigenvalue truncated efficient information matrix $\hat{\mathcal{I}}_\gamma$ (e.g. Lütkepohl and Burda, 1997). Formally,

$$\hat{\mathcal{I}}_\gamma^{t} = \hat{U}_n\hat{\Lambda}_n(\nu_n)\hat{U}_n' \,, \tag{2.11}$$

where $\hat{\Lambda}_n(\nu_n)$ is a diagonal matrix with the ν_n-truncated eigenvalues of $\hat{\mathcal{I}}_\theta$ on the main diagonal and $\hat{U}_n$ is the matrix of corresponding orthonormal eigenvectors. To be specific, let $\{\hat{\lambda}_{n,i}\}_{i=1}^{L}$ denote the non-increasing eigenvalues of $\hat{\mathcal{I}}_\theta$, then the (i,i)th element of $\hat{\Lambda}_n(\nu_n)$ is given by $\hat{\lambda}_{n,i}\mathbf{1}(\hat{\lambda}_{n,i} \geq \nu_n)$.

Equations (2.5)-(2.11) define the semi-parametric score statistic for the LSEM model (2.3) for a given parameter vector $\gamma = (\alpha, \beta)$. To test the null hypothesis (2.4) we will evaluate this test statistic at $\alpha = \alpha_0$, i.e. fixing the possibly unidentified parameters under the null, and at $\hat{\beta}$, which can be any $\sqrt{n}$ consistent estimate for β. In our simulations, we use ordinary least squares estimates for σ and $b = \mathrm{vec}(B)$, or one-step efficient estimates following van der Vaart (2002, Section 7.2). Let $\hat{\gamma} = (\alpha_0, \hat{\beta})$, in our theoretical section below we show that under suitable assumptions the score statistic will converge to a χ^2 limit. Specifically, we prove that under H_0 for any $a \in (0,1)$ we have

$$\lim_{n\to\infty} P(\hat{S}_{\hat{\gamma}} > c_n) \leq a \,, \tag{2.12}$$

where c_n is the $1-a$ quantile of the $\chi^2_{r_n}$ distribution with $r_n = \mathrm{rank}(\hat{\mathcal{I}}^t_{\hat{\gamma}})$. Importantly, as we show in section 2.4 this result does not rely on any assumptions regarding the shape of the densities η, i.e. we do not need to assume that η is non-Gaussian. Only conventional moment assumptions and some regularity conditions on the densities are required. The following algorithm summarizes the complete implementation.

```
Algorithm: Efficient score test for LSEM
1 Obtain $\sqrt{n}$-consistent estimates $\hat{\beta} = (\hat{\sigma}, \hat{b})$ and residuals $\hat{V}_i = Y_i - \hat{B}X_i$;
2 For $k = 1, \ldots, K$, compute $\hat{\phi}_k(\hat{A}_{k\bullet}\hat{V}_i)$ from (2.7) with $\hat{A} = A(\alpha_0, \hat{\sigma})$;
3 Compute the efficient scores $\hat{\ell}_{\hat{\gamma}}(\hat{V}_i)$ from (2.5) and the information matrix $\hat{I}_{\hat{\gamma}}$ from (2.8) using $\hat{\gamma} = (\alpha_0, \hat{\beta})$;
4 Compute $\hat{\kappa}_{\hat{\gamma}}(\hat{V}_i)$ and $\hat{\mathcal{I}}_{\hat{\gamma}}$ from (2.9).
5 Compute the score statistic $\hat{S}_{\hat{\gamma}}$ from (2.10) and reject $H_0 : \alpha = \alpha_0$ if $\hat{S}_{\hat{\gamma}} > c_n$, where $c_n$ is the $1 - a$ quantile of the $\chi^2_{r_n}$ distribution with $r_n = \mathrm{rank}(\hat{\mathcal{I}}^t_{\hat{\gamma}})$.
```

The algorithm highlights that the computational cost for evaluating the semi-parametric score statistic $\hat{S}_{\hat{\gamma}}$ is modest; effectively one only needs to compute K B-spline regressions to obtain the log density scores. Importantly, this implies that the algorithm can be implemented without relying on numerical optimization routines. Confidence sets for α can be constructed by inverting the score statistic over a range of values for α_0.

2.4. Asymptotic theory

In this section we present our main theoretical results and discuss the required underlying assumptions.

2.4.1. Assumptions

We assume that we observe a random sample $\{(Y_i, \tilde{X}_i)\}_{i=1}^n$ from model (2.3) where the underlying components satisfy the following.

Assumption 2.4.1. *For $\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,K})'$ in model* (2.3)*, each component $\epsilon_{i,k}$ has a continuously differentiable root density (with respect to Lebesgue measure on $\mathbb{R}$). We write the density as η_k with log density score $\phi_k(x) = \partial \log \eta_k(x)/\partial x$. We assume that for all $k = 1, \ldots, K$ and some $\delta > 0$*

1. $\mathbb{E}\epsilon_{i,k} = 0$, $\mathbb{E}\epsilon_{i,k}^2 = 1$, $\mathbb{E}\epsilon_{i,k}^{4+\delta} < \infty$, $\mathbb{E}(\epsilon_{i,k}^4) - 1 > \mathbb{E}(\epsilon_{i,k}^3)^2$, *and* $\mathbb{E}\phi_k^{4+\delta}(\epsilon_{i,k}) < \infty$*;*
2. $\mathbb{E}\phi_k(\epsilon_{i,k}) = 0$, $\mathbb{E}\phi_k(\epsilon_{i,k})\epsilon_{i,k} = -1$, $\mathbb{E}\phi_k(\epsilon_{i,k})\epsilon_{i,k}^2 = 0$ *and* $\mathbb{E}\phi_k(\epsilon_{i,k})\epsilon_{i,k}^3 = -3$*;*
3. *$\epsilon_{i,k}$ is independent of $\epsilon_{i,l}$ for all $k \neq l$;*

4. *$\eta_0 \in \mathscr{X}$ is a density function (with respect to Lebesgue measure on $\mathbb{R}^{d-1}$) such that if $\tilde{X}_i \sim \eta_0$, then $\mathbb{E}\tilde{X}_i\tilde{X}_i'$ is positive definite and $\mathbb{E}[|\tilde{X}_{i,l}|^{4+\delta}] < \infty$ for all $l = 1, \ldots, d-1$;*

5. *ϵ_i and $\tilde{X}_i$ are independent.*

The first part normalizes the errors to have mean zero, variance one and finite four+δ moments,[5] hence ruling out heavy tailed errors.[6] Additionally, we require the log density scores $\phi_k(x) = \partial \log \eta_k(x)/\partial x$ evaluated at the errors to have finite four+δ moments. The second part simplifies the construction of the efficient score functions. Whilst this may at first glance appear a strong condition, Lemma D.2 in the supplementary material shows that if the first part holds, then a simple sufficient condition is that the tails of the densities η_k converge to zero at a polynomial rate.[7] The third part imposes that the components of ϵ_i are independent. Part four imposes some structure on $\tilde{X}_i$ that allows us to identify B; notably positive definite second moments and four+δ finite moments are required. Part five requires the explanatory variables and errors to be independent. This can be relaxed by requiring the moment assumptions in 2.4.1 to hold conditional on $\tilde{X}_i$. In this setup, our general theory as outlined in this section would continue to be valid though the resulting efficient score function would take a different form.

Most important is what is *not* in Assumption 2.4.1: there is no condition that imposes that a certain number of components of ϵ_i have a (sufficiently) non-Gaussian distribution.

The second assumption that we impose is only required for the estimation of the log density scores $\phi(x) = \partial\eta(x)/\partial x$ using B-spline regressions and can be appropriately replaced when a different density score estimator is used. For notation purposes, let $\Xi_{k,n}^L$ and $\Xi_{k,n}^U$ denote the lower and upper endpoints of the cubic B-splines for $\phi_k(x)$ for $k = 1, \ldots, K$.[8]

Assumption 2.4.2. *Define ν_n according to $\nu_{n,p}^2 = o(\nu_n)$ with $p := \min\{1 + \delta/4, 2\}$ and $\nu_{n,p} = n^{(1-p)/p}$ if $p \in (1,2)$ or $\nu_{n,p} = n^{-1/2}\log(n)^{1/2+\rho}$, for some $\rho > 0$, if $p = 2$. Let $\phi_{k,n} := \phi_k \mathbf{1}_{[\Xi_{k,n}^L, \Xi_{k,n}^U]}$ and $\Delta_{k,n} := \Xi_{k,n}^U - \Xi_{k,n}^L$ and suppose that for , $[\Xi_{k,n}^L, \Xi_{k,n}^U] \uparrow \tilde{\Xi} \supset \operatorname{supp}(\eta_k)$ and $\delta_{k,n} \downarrow 0$ such that*

(I) *$P(\epsilon_{i,k} \notin [\Xi_{k,n}^L, \Xi_{k,n}^U]) = o(\nu_n^2)$;*

(II) *For some $\iota > 0$, $n^{-1}\Delta_{k,n}^{2+2\iota}\delta_{k,n}^{-(8+2\iota)} = o(\nu_n)$;*

(III) *η_k is bounded ($\|\eta_k\|_\infty < \infty$) and differentiable, with a bounded derivative: $\|\eta_k'\|_\infty <$*

[5] $\mathbb{E}(\epsilon_{i,k}^4) - 1 \geq \mathbb{E}(\epsilon_{i,k}^3)^2$ always holds; this is known as Pearson's inequality. See e.g. result 1 in Sen (2012). Assuming that $\mathbb{E}(\epsilon_{i,k}^4) - 1 > \mathbb{E}(\epsilon_{i,k}^3)^2$ rules out (only) cases where $1, \epsilon_{i,k}$ and $\epsilon_{i,k}^2$ are linearly dependent when considered as elements of L_2. See e.g. Theorem 7.2.10 in Horn and Johnson (2013).

[6] Heavy tailed errors in ICA and SVAR models have recently been considered in Davis and Ng (2022) and Davis and Fernandes (2022), but an inferential theory remains to be developed.

[7] See Example 3 in the supplementary material for an explicit example of a density which satisfies the first part of the assumption but not the second.

[8] In practice, we select these points as the lower 5th and upper 95th percentiles of the samples $\{V_{i,k}\}_{i=1}^n$ adjusted by $\log\log n$, see the implementation section 2.3.

∞;

(IV) *For each n, $\phi_{k,n}$ is three-times continuously differentiable on $[\Xi^L_{k,n}, \Xi^U_{k,n}]$ and $\|\phi^{(3)}_{k,n}\|^2_\infty \delta^6_{k,n} = o(\nu_n)$;*[9]

(V) *There are $c > 0$ and $N \in \mathbb{N}$ such that for $n \geq N$ we have $\inf_{t\in[\Xi^L_{k,n}, \Xi^U_{k,n}]} |\eta_k(t)| \geq c\delta_{k,n}$.*

First, the assumption makes explicit the truncation rate ν_n that is needed for the truncation of the eigenvalues in (2.11). This rate is split into two parts. The "slow" rate $n^{(1-p)/p}$ (for $p \in (1, 2)$) is always sufficient given assumption 2.4.1, but if ϵ_k has finite eighth moments the faster rate applies.

Part (i) imposes that the tails of $\epsilon_{i,k}$ decay to zero sufficiently fast.[10] Part (ii) ensures that the number of knots does not grow to fast relative to the sample size (and the truncation rate). Part (iii) requires the density and its derivative to be bounded. Part (iv) requires the existence of the third derivatives of ϕ_k and that the rate of increase of the third derivative is not too great. Part (v) ensures that the density is bounded away from zero on $[\Xi^L_{k,n}, \Xi^U_{k,n}]$. Overall, these assumptions are similar as in Chen and Bickel (2006), with two key differences.[11] Firstly, Chen and Bickel (2006) require the conditions to hold for the functions $v \mapsto \phi_k(A_{k\bullet}v)$ (rather than ϕ_k), uniformly over shrinking balls (at rate $n^{-1/2}$) around A. In our setting we are only interested in testing as consistent estimation is ruled out by the possible lack of identification, hence we only require the conditions to hold for the functions ϕ_k. Secondly, unlike Chen and Bickel (2006), we require convergence at a rate ν_n which satisfies certain decay conditions. This is due to the fact that we may have a singular efficient information matrix and in order to obtain a consistent estimate of the Moore – Penrose inverse of this matrix, we require knowledge of the rate of convergence of our estimate.

2.4.2. Main result

In this section we formally state our main result for the efficient score test $\hat{S}_{\hat{\gamma}}$. To do so, instead of evaluating the efficient score test at the $\sqrt{n}$-consistent estimates $\hat{\gamma} = (\alpha_0, \hat{\beta})$ we will evaluate the score test at its discretized version $\bar{\gamma} = (\alpha_0, \bar{\beta}_n)$. Formally, let $B_n = n^{-1/2}C\mathbb{Z}^{L_\beta}$ for some $C > 0$ and define $\bar{\beta}_n$ as a new version of $\hat{\beta}$ that replaces its value with the closest point in B_n. Note that this changes each coordinate of $\hat{\beta}$ by a quantity which is at most $O(n^{-1/2})$, hence the $\sqrt{n}$-consistency is retained by discretization. Since the constant C can be chosen arbitrarily large this change has no practical relevance for the implementation of the test.

[9] The differentiability and continuity requirements at the end-points are one-sided.

[10] The required speed of decay is linked to the truncation rate.

[11] Cf. their conditions C3, C5 – C7, p. 2834.

The advantage of relying on discretized estimates is that it simplifies the proof of the main result. Specifically, it removes the need to show uniform convergence between the efficient scores evaluated at $\hat{\beta}$ and β. The discretization trick is due to Le Cam (1960) and is widely used in statistics, see the detailed discussion in Le Cam and Yang (2000, Section 6.3), or van der Vaart (1998, page 72). It has also been adopted in econometrics, see Cattaneo et al. (2012) for instance.

With this modification we have the following result.

Theorem 2.4.1. *Suppose that Assumptions 2.4.1 and 2.4.2 hold, that $(\alpha,\sigma)\mapsto A(\alpha,\sigma)$ is continuously differentiable and the maps $(\alpha,\sigma)\mapsto\zeta^{\alpha}_{l,k,j}$ and $(\alpha,\sigma)\mapsto\zeta^{\sigma}_{l,k,j}$ are Lipschitz continuous. Let $r_n=\operatorname{rank}(\hat{\mathcal{I}}^t_{\bar{\gamma}})$ and denote by c_n the $1-a$ quantile of the $\chi^2_{r_n}$ distribution, for any $a\in(0,1)$. Then, under H_0*

$$\lim_{n\to\infty}P_{\theta_0}(\hat{S}_{\bar{\gamma}}>c_n)\le a,$$

with inequality only if $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_0})=0$ where $\gamma_0=(\alpha_0,\beta)$.

The proposition shows that semi-parametric score test $\hat{S}_{\bar{\gamma}}$ has correct asymptotic size for all densities η that satisfy the requirements in Assumptions 2.4.1 and 2.4.2. The requirements that $(\alpha,\sigma)\mapsto A(\alpha,\sigma)$ is continuously differentiable and $(\alpha,\sigma)\to\zeta^{\alpha}_{l,k,j}$, $(\alpha,\sigma)\to\zeta^{\sigma}_{l,k,j}$ are Lipschitz continuous are easily verified for Examples 1 and 2. The choice for the estimator $\hat{\beta}$ is left open to the researcher. Possible choices include using OLS estimates or one-step efficient estimators (e.g. van der Vaart, 2002, Section 7.2). Our simulation study explores the finite sample differences between these two estimators.

It follows from Choi et al. (1996) that for non-singular information matrices tests based on $\hat{S}_{\bar{\gamma}}$ are asymptotically uniformly most powerful within the class of rotation invariant tests. This implies that asymptotically when testing the hypothesis $H_0:\alpha=\alpha_0$, the power of the test is the greatest possible in the class of rotationally invariant tests. This makes tests based on $\hat{S}_{\bar{\gamma}}$ attractive for scenarios where there is no explicit direction in which one want to maximize power. When such directions are given alternative test statistics, also based on the efficient score function, can be considered (e.g. Bickel et al., 2006). Uniformity results and minimax optimality results which permit singular information matrices can be found in Lee (2022) for efficient score tests in general semi-parametric models.

2.5. Simulation results

In this section we study the finite sample properties of the singularity and identification robust score test $\hat{S}_{\bar{\gamma}}$. We study the size and power of the test under different data generating processes and compare its performance to several alternatives that have been proposed in the literature. We first study the simple model of section (2.2) after which we consider the general linear simultaneous equations model (2.3). The supplementary material provides additional results.

2.5.1. Baseline model

We start by drawing independent samples from model (2.1), which we restate for convenience

$$Y_i = R'\epsilon_i \ , \qquad i = 1, \ldots, n \ .$$

We take Y_i to be $K \times 1$ and consider $K = 2, 3$ and $K = 5$. The sample size is taken as $n = 200, 500$ or $n = 1000$. We fix $\epsilon_{i,1}$ to have a standard Gaussian density and consider different densities for $\epsilon_{i,k}$, with $k = 2, \ldots, K$. The non-Gaussian densities are either Student's t or mixtures of normals taken from Marron and Wand (1992). Figure E.3 provides an overview.

The matrix of interest $R = R(\alpha)$ is orthogonal and parametrized by the Cayley transformation of a skew-symmetric matrix (e.g. Gouriéroux et al., 2017):

$$R(\alpha) = (I - \Omega(\alpha))(I + \Omega(\alpha))^{-1} \ ,$$

where $\Omega(\alpha)$ is a skew-symmetric matrix (i.e. $\Omega(\alpha)' = -\Omega(\alpha)$) parameterized by α which we sample at random from $\alpha \sim N(0, I_{L_\alpha})$.

In this setting there are no additional nuisance parameters which allows us to concentrate on the consequences of weak non-Gaussianity on the efficient score test and some alternative tests that have been proposed in the literature. In the simulation designs below we include additional finite dimensional nuisance parameters (i.e. $\beta = (\sigma, b)$) and investigate whether their inclusion alters the size and power of the test.

For each specification we simulate $S = 5,000$ datasets and for each we compute the efficient score statistic $\hat{S}_{\hat{\gamma}}$ as defined in equation (2.10) following the Algorithm given in Section 2.3.[12] We implement the log density score estimator (2.7) using $B = 4, 6$ or 8 cubic splines.

In Table E.2 we show the empirical rejection frequencies corresponding to the $S_{\hat{\gamma}}$ test with nominal size 0.05. The columns correspond to the different choices for the densities ϵ_k for $k \geq 2$.

The first column corresponds to the case where all densities are Gaussian and the expected likelihood takes the same value for all $\alpha \in \mathbb{R}^{L_\alpha}$, i.e. α is unidentified. Nonetheless, we find that the empirical rejection frequency of the score test is always close to the nominal size.

[12] To be specific, since the model does not contain any finite dimensional nuisance parameters step 1 in the algorithm can be skipped and the score statistic is simply evaluated at α_0.

This holds regardless of the sample size n, the dimension of the model K and the number of cubic splines B.

Second, when the densities for $k \geq 2$ are non-Gaussian the size remains correct. Specifically, columns 2-4 show the results for the case where $\epsilon_{i,k}$ follows a Student's t distribution with decreasing degrees of freedom ($\nu = 15, 10, 5$). No matter how close we get to the Gaussian density the size remains correct. Columns 5-10 show similarly correct size for a variety of mixture distributions. Even for complicated skewed bi-modal densities (e.g. columns 8-10) the $S_{\hat{\gamma}}$ test has size close to nominal regardless of the sample size.

Third, overall the number of cubic splines used has little influence on the results. A close inspection reveals that when the number of cubic splines is equal to four the test becomes mildly conservative for some densities, therefore we use $B = 6$ cubic splines in the remaining exercises.

Overall, the asymptotic approximation in Theorem 2.4.1 seems to provide a good approximation for the finite sample behavior of the semiparametric score test, at least for the distributions shown in Figure E.3.

2.5.2. Comparison to alternative approaches

Next, we compare our semiparametric testing approach to different parametric approaches based on (psuedo) maximum likelihood and the generalized method of moments. We concentrate on evaluating different tests based on size and power in the vicinity of Gaussianity.[13]

ALTERNATIVE TESTS Conceptually, there are two types of alternative tests that we consider: (i) tests that rely on estimates for α and (ii) tests that fix $\alpha = \alpha_0$ under the null. Clearly, from our intuitive discussion in Section 2.2 it follows that we expect tests that fix α under the null to perform relatively well.

In category (i) we consider the standard maximum likelihood Wald ($\mathrm{W}^{\mathrm{mle}}$) and likelihood ratio ($\mathrm{LR}^{\mathrm{mle}}$) tests based on the Student's t density for ϵ_k. For densities 2-4 in Figure E.3 these tests correspond to exact maximum likelihood tests, with the caveat that when the degrees of freedom increases the parameters α become weakly identified, or not-identified. For all other densities these tests are mis-specified.

In addition, we consider the psuedo-maximum likelihood Wald test ($\mathrm{W}^{\mathrm{pmle}}$) from Gouriéroux et al. (2017). This test is asymptotically valid for a broader range of true distribution functions and amount to fixing the functional form of the densities $\eta_1, \ldots, \eta_K$. We follow the implementation of Gouriéroux et al. (2017) and choose the Students t density

[13] The recent simulation studies of Herwartz et al. (2019) and Moneta and Pallante (2020) provide further simulation evidence for existing methods, also focusing on estimation accuracy.

with five degrees of freedom as the pseudo-likelihood and compute the Wald statistic based on this density.

Finally, we consider the recently developed GMM method of Lanne and Luoto (2021), which relies on higher order moments to identify the parameters α. We use $\mathbb{E}\epsilon_{i,k}^2\epsilon_{i,j} = 0$, $\mathbb{E}\epsilon_{i,k}^3\epsilon_{i,j} = 0$ and $\mathbb{E}\epsilon_{i,k}^2\epsilon_{i,j}^2 = 1$ as moment conditions for all $j \neq k$ and $j, k = 1, \ldots, K$. The GMM likelihood ratio test is then computed as the rescaled difference between the unrestricted and restricted J-statistics, based on the 2-step GMM estimator ($\mathrm{LR}^{\mathrm{gmm}}$), see Lanne and Luoto (2021) for details.[14]

In category (ii) we consider tests which fix $\alpha = \alpha_0$ under the null. Specifically, we include the standard LM test ($\mathrm{LM}^{\mathrm{mle}}$) based on the Student's t density where the degrees of freedom parameter is estimated from the data. Second, we consider the pseudo-maximum likelihood version of the LM test ($\mathrm{LM}^{\mathrm{pmle}}$) based on Gouriéroux et al. (2017), which fixes the degrees of freedom at five. Finally, we consider the GMM-based identification robust S-statistic ($\mathrm{S}^{\mathrm{gmm}}$) of Stock and Wright (2000), which was recently considered in Drautzburg and Wright (2021) in the context of structural VAR models with non-Gaussian errors. We use the same moment conditions as considered in Drautzburg and Wright (2021) for the $\mathrm{LM}^{\mathrm{gmm}}$ test.

Size comparison We compare the size of the different tests for the simulation designs described in Section 2.5.1. The empirical rejection frequencies are shown in Table E.3 for the case where $K = 2$ and $n = 200, 500, 1000$. Overall we find, perhaps not surprisingly, that all tests in category (i) do not have correct size when the true density is close to Gaussian nor when the corresponding method is based on a mis-specified model. This shows that tests based on estimates for α are generally unreliable. Second, tests in category (ii) overall control the size of the test well.

More specifically, we find that the Wald tests ($\mathrm{W}^{\mathrm{mle}}$ and $\mathrm{W}^{\mathrm{pmle}}$) tend to over-reject quite severely whilst the standard likelihood ratio test ($\mathrm{LR}^{\mathrm{mle}}$) tends to be undersized for most densities, especially in the vicinity of the Gaussian density, as ought to be expected given the earlier evidence in shown in Figure 2.1. Finally, the GMM likelihood ratio test ($\mathrm{LR}^{\mathrm{gmm}}$) is also over-sized, which confirms findings in Lanne and Luoto (2021) where the $\mathrm{LR}^{\mathrm{gmm}}$ also over-rejects when the densities of the structural shocks are close to Gaussian.

In the second category the semi-parametric score test $\hat{S}_{\hat{\gamma}}$ (as proposed in this paper) and the pseudo maximum likelihood LM test ($\mathrm{LM}^{\mathrm{pmle}}$), inspired by Gouriéroux et al. (2017), both have near perfect size across all densities. The standard LM test ($\mathrm{LM}^{\mathrm{mle}}$) also performs reasonably well, but when the functional form of the true densities is very different from the Student's t density (e.g. separate bi-modal, column 9) the test tends to under-reject.[15]

[14] Note that lower order moments are not required as the baseline model $Y_i = R'\epsilon_i$ implies that the observations have mean zero and unit variance.

[15] Recall here that this test is based on a misspecified density.

Finally, the GMM based robust S test ($\mathrm{S}^{\mathrm{gmm}}$) tends to be over-sized for small samples, but for large samples it generally shows correct size except for densities with moderately heavy tails such as the $t(5)$ density (column 4). In these cases the $\mathrm{S}^{\mathrm{gmm}}$ is over-sized which can be understood when realizing that the GMM approach requires eight finite moments for inference when based on fourth-order moment restrictions. The $t(5)$ density does not have eight finite moments.

In sum, we recommend avoiding statistics that are based on estimates for α as these are overall unreliable when the shock distributions are close to Gaussian. All tests that fix α under the null perform at least reasonably well. In the next section we compare these tests based on their finite sample power.

POWER COMPARISON We compare the power of all tests that fix α under the null, that is $\hat{S}_{\gamma}$, $\mathrm{LM}^{\mathrm{mle}}$, $\mathrm{LM}^{\mathrm{pmle}}$ and $\mathrm{S}^{\mathrm{gmm}}$.

We consider the case where $K = 2$ and $n = 1000$.[16] In this setting α is a scalar parameter and we fixed the true value at 0 (an arbitrary choice). Figure E.4 shows the empirical rejection frequencies when we vary α around $\alpha = 0$. Each point on the curve is based on $S = 5,000$ simulations.

Two main findings stand out. First, for the Student's t densities $t(15)$, $t(10)$ and $t(5)$ (panels 2-4) the standard LM test ($\mathrm{LM}^{\mathrm{mle}}$) shows the highest power. This is not surprising as for these data generating processes the $\mathrm{LM}^{\mathrm{mle}}$ test is correctly specified and hence takes advantage of fitting the true densities using only a scalar parameter. That said, the semi-parametric score test ($\hat{S}_{\hat{\gamma}}$) and the pseudo maximum likelihood LM test ($\mathrm{LM}^{\mathrm{pmle}}$) come reasonably close in terms of power.

Second, for all other densities, i.e. different mixtures of normals in panels 5 – 10, the semi-parametric score test ($\hat{S}_{\hat{\gamma}}$) shows the highest power. Sometimes the difference with the other tests is not very large, but for instance for bi-modal densities (panels 8-10) the differences are substantial. Overall, the good power of the $\hat{S}_{\hat{\gamma}}$ test corresponds to the theoretical finding that for non-singular information matrices the test is asymptotically uniformly most powerful in the class of unbiased tests.[17]

Besides the $\hat{S}_{\hat{\gamma}}$ test, we note that the pseudo maximum likelihood LM test and the GMM based S test shows quite promising power for most of the densities considered. None of these dominates the other. The caveat for the GMM test is that it is size-distorted for moderately heavy tails (panel 4).

[16] Power comparisons for different n can be found in the supplementary material.

[17] Cf. Choi et al. (1996).

2.5.3. Linear simultaneous equations model

Next, we discuss the simulation results for the general linear simultaneous equations model (2.3). The dimensions of the design are similar as above with the addition that we consider $d = 2, 3$ for the number of covariates. We now parametrize $A(\alpha, \sigma)^{-1} = \Sigma^{1/2}(\beta_1)R(\alpha)$ as in example 1, where $\Sigma^{1/2}$ is lower triangular and the rotation matrix R remains to be specified by the Cayley transform. The explanatory variables are drawn from the standard normal distribution.

The vector of finite dimensional nuisance parameters β now includes $\sigma = \text{vech}(\Sigma^{1/2})$ and $b = \text{vec}(B)$. Our main theoretical result in Theorem 2.4.1 shows that β can be approximated by any $\sqrt{n}$-consistent estimate. Obviously, ordinary least squares estimates are attractive for their simplicity, but given the non-normality of the structural shocks these estimates may be improved. Therefore we also consider estimating β by one-step-efficient estimates (e.g. van der Vaart, 2002, Section 7.2), which are easy to compute here since the efficient score of β is computed anyway to construct the score test.

Similar, as before the first error $\epsilon_{i,1}$ follows a Gaussian distribution and the different densities from Figure E.3 are assigned to the other error terms. For each specification we simulate $S = 5,000$ datasets and for each sample we compute the semi-parametric score statistic using the Algorithm in Section 2.3.

SIZE RESULTS The empirical rejection frequencies are shown in Tables E.4 and E.5 for the OLS and one-step efficient estimates for β, respectively.

We find that for all the rejection frequencies of the $\hat{S}_{\hat{\gamma}}$ test are generally close to the nominal size. That said, there is more variation in the empirical rejection frequencies compared to Table E.2, indicating that the estimation of the finite dimensional nuisance parameters does have consequences.

Starting with Table E.4 where $\hat{\beta}$ is estimated by OLS. We find that the size of $\hat{S}_{\hat{\gamma}}$ is the same regardless of how close the densities of $\epsilon_{i,k}$ are to the Gaussian density. Specifically, moving from columns 1-4 (i.e. from Gaussian to $t(5)$) we see virtually no changes in the rejection frequencies. This holds for all specifications considered and highlights the main point of this paper: the semi-parametric score test yields reliable inference even when α is not, or poorly, identified.

Depending on the dimension of β we do find size distortions for small sample sizes, most notably when $K = 5$ and $n = 200$. In this setting β is of dimension 20 or 25 depending on $d = 2, 3$, and we see that the test is often over-sized. This does not hold for all densities considered, but for Gaussian, Student's t and kurtotic unimodal densities the test over-rejects. When n increases the over-rejection vanishes and the test appears correctly sized.

For the one-step efficient estimator for β the results are shown in Table E.5. We find that on average the empirical rejection frequencies are larger when compared to the OLS estimator. Notably, when n is small over-rejection becomes more severe. Again, we find that this holds uniformly across densities, i.e. distortions do no depend on being close to Gaussian, and the sizes improve when n increases.

POWER RESULTS Next, we investigate the power of the $\hat{S}_{\hat{\gamma}}$ test for the LSEM model. We again consider the case where $K = 2$, $d = 2$ and $n = 1000$, which allows us to compare the results with those for the baseline model. The power curves are shown in Figure E.5 for both OLS and one-step estimates for β.

First, when comparing Figure E.5 to the case without nuisance parameters (i.e. Figure E.4) we find that the power of the test is reduced when we include nuisance parameters. Second, the power of the test using the one-step efficient estimates (dotted blue line) is higher when compared to the same test evaluated at OLS estimates. This holds for all densities considered.

Based on these results we recommend using OLS estimates for β when the sample size is small (e.g. $n = 200, 500$), but for larger sample sizes the one-step efficient estimates are preferable.

2.6. Testing production function coefficients

In this section we explore whether non-Gaussian distributions can help to identify the coefficients in the production function of a firm. Fittingly, the very first contributions in this literature highlighted the identification problem in this setting using simultaneous equations (e.g. Marschak and Andrews, 1944; Hoch, 1958). This generated a large number of works that aim to address the simultaneity problem in different ways. Prominent examples include using panel data methods (e.g. Arellano and Bond, 1991; Blundell and Bond, 1998) or proxy variable methods (e.g. Olley and Pakes, 1996; Leeb and Pötscher, 2003; Ackerberg et al., 2015).

To study how non-Gaussian distributions may assist in the quest for identification we consider the baseline Cobb-Douglas production function

$$O_i = e^{c_1} L_i^{\alpha_1} K_i^{\alpha_2} e^{\epsilon_{i,1}} ,$$

where O_i, L_i, K_i denote output, labor and capital, respectively, and $\epsilon_{i,1}$ captures unobserved factors that determine output. Our interest is in the coefficients α_1 and α_2 that determine the contributions of labor and capital to output. The, well known, difficulty for learning about α_1 and α_2 is that the inputs L_i, K_i are typically choice variables of the firm. Allocations are made to maximize profits and hence will generally depend on unobservables $\epsilon_{i,1}$.

To address this simultaneity problem we consider a simultaneous equations approach that allows for correlation among L_i, K_i, ϵ_1, and exploits possible non-Gaussianity in the errors to identify the parameters α_1 and α_2.

To be specific, the models that we consider are defined for $Y_i = (\log O_i, \log L_i, \log K_i)'$, and are of the form

$$S(\alpha, \sigma)Y_i = BX_i + D(\sigma)\epsilon_i \ , \tag{2.13}$$

where X_i includes a constant and any other additional exogenous variables such as the age of the firm. We adopt the following specification for the matrices S and D.

$$S(\alpha, \sigma) = \begin{bmatrix} 1 & -\alpha_1 & -\alpha_2 \\ -\sigma_1 & 1 & -\alpha_3 \\ -\sigma_2 & -\sigma_3 & 1 \end{bmatrix} \quad \text{and} \quad D(\sigma) = \begin{bmatrix} \sigma_4 & 0 & 0 \\ 0 & \sigma_5 & 0 \\ 0 & 0 & \sigma_6 \end{bmatrix} .$$

We note that parameters in σ can be recovered from the variance of $Y_i - BX_i$ and we will simultaneously test $\alpha = \alpha_0$, where $\alpha = (\alpha_1, \alpha_2, \alpha_3)'$, for different choices of α_0 to obtain the confidence sets. The positioning of α_3 is arbitrary in our setting as it is not a parameter of interest, but it can also not be identified from the variance alone. The confidence sets for α_1 and α_2 that we report are obtained by taking the minimum and maximum values for α_1 and α_2 that are not rejected by the score test.[18] Finally, to pin down the desired rotation we impose that α_1 and α_2 are positive and the correlations between L_i, K_i and $\epsilon_{i,1}$ are non-negative. In other words, positive shocks to output do not decrease labor and capital, a mild sign restriction that corresponds with most economic models (e.g. Hoch, 1958).

We use a sample of $115,000$ manufacturing firms that are observed from 2000 until 2017.[19] We perform two exercises. First, to illustrate our methodology we consider the cross section of firms that exist in 2017 and investigate in detail the output of the methodology. Second, we repeat the exercise for different years and assess the changes in α_1 and α_2 over time.

RESULTS We first illustrate the methodology using the manufacturing firms that existed in 2017. We have $n = 1247$ firms with observations for output, labor and capital. We consider model (2.13) with a constant and possibly the age of the firm as a control variable (e.g. Olley and Pakes, 1996).

The 95% confidence bounds for the production function coefficients α_1 (labor) and α_2 (capital) are shown in Table E.6. We find that these coefficients are generally well identified empirically. In particular, with 95% confidence, α_1 lies between 0.41 and 0.68, while α_2 lies between 0.27 and 0.50, for all choices of the control variables. The joint confidence

[18] We note that this projection approach is conservative and refinements along the lines of Kaido et al. (2019) may improve the current findings.

[19] The data are obtained from CompuStat.

region for (α_1, α_2) is shown in the top left panel of Figure E.6. It shows that we cannot reject that $\alpha_1 + \alpha_2 = 1$ as the confidence region exactly lies on this line.

To understand where the identification in the LSEM is coming from, the other panels in Figure E.6 show the empirical densities of the residuals $\hat{\epsilon}_i = \hat{A}(Z_i - \hat{B}X_i)$, where $\hat{A}$ corresponds to the choice for α that minimizes the score statistic. We find that the empirical densities are indeed different from the normal density, notably for the first density. Overall, we can reject the null hypothesis that the errors are normally distributed for the first and second errors using a Jarque-Bera test. For the third error we cannot reject normality.

Given our simulation results such mild deviations from Gaussianity may cause problems for standard inference methods. This is true for the alternative methods that which were found not robust to weak deviations from Gaussianity; they tend to give much smaller confidence bands. This suggests that whilst non-Gaussianity may be a useful tool for identification, robust methods need to be adopted for the approach to be used reliably. We emphasize that besides the sign restrictions that ensure that the correlations between L, K and ϵ_1 are non-negative no further structural assumptions or instruments are needed.

Table E.6 also reports the baseline OLS estimates as obtained by regressing log output on the controls and log labor and log capital. We find that these estimates are very different and the confidence intervals do not overlap with those of the LSEM. This highlights that there may indeed be endogeneity in the form of correlation between labor, capital and the error term $\epsilon_{i,1}$.

To verify whether this conclusion is justified we need to test whether the underlying assumption regarding the independence of the underlying structural shocks is indeed true (e.g. Montiel Olea et al., 2022). To do so, we adopt the permutation test for independent components as proposed in Matteson and Tsay (2017). We implement their test on the sample $\{\hat{\epsilon}_i\}$ as defined above.[20] The results are shown in the bottom row of Table E.6. Depending on whether age is included as a control variable, the p-values are 0.12 and 0.16 indicating that there is not substantial evidence against independence.

Next, to highlight that the year 2017 was in no way exceptional we repeat the previous exercise for the years 2000-2017. The results for the model that includes age as a control variable are shown in Figure E.7. Overall, the findings are very stable. We do notice a modest decline in the labor input coefficient and an increase of the coefficient on capital towards the end of the sample.

[20] The test was implemented using the R package steadyICA using the function permTest.

2.7. Conclusion

In this paper we highlighted a weak identification problem that arises when non-Gaussianity is used to identify coefficients in LSEMs. The consequence of this problem is that several existing inference methods suffer from size distortions when the true distributions are close to Gaussian.

To remedy this problem we proposed an identification robust semi-parametric score statistic for testing hypotheses in LSEMs. Under mild regularity conditions we showed that the score test retains correct asymptotic size regardless of the shape of the true density functions. A simulation study shows that our asymptotic theory provides an accurate approximation to the finite sample performance of our test.

While we have restricted our treatment to models where the observations were independently distributed across entities, we note that a similar approach may be considered for dynamic models, but this will require extending our results to allow for non-i.i.d. data. Similarly, dynamic panel data models could be considered pending a novel strategy for handling the initial conditions. These extensions are left for future work.

Acknowledgements

I am grateful to my co-author & advisor Geert Mesters for the suggestion of this problem and co-authoring the chapter. I would also like to thank Christian Brownlees, Bjarni G. Einarsson, Kirill Evdokimov, Lukas Hoesch, Katerina Petrova, Barbara Rossi, André B. M. Souza and Philipp Tiozzo for helpful comments and discussions. I am also grateful to participants at the 2022 Workshop on Data Analytics for Business (Verona) for helpful comments and discussions.

Appendices

In this appendix we provide the proof for Theorem 2.4.1. The proof is structured as follows. We first provide a general approach for conducting identification and singularity robust hypothesis tests in semiparametric models. This general theory is subsequently applied to prove Theorem 2.4.1.

Throughout the appendix we often use the empirical process notation: $Pf = \mathbb{E}f(X_i)$, $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(Y_i)$ and $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f$. Further, G_k denotes the law on $\mathbb{R}$ corresponding to η_k and ϵ_k is distributed according to G_k. Similarly G_0 denotes the law on $\mathbb{R}^{d-1}$ corresponding to η_0 and $\tilde{X}$ is distributed according to G_0.

A. General theory

We expound a general approach for conducting identification robust hypothesis tests in semi-parametric models. The LSEM model of Section 2.3 constitutes as a special case of the model considered in this section.

Let $Y \in \mathcal{Y} \subset \mathbb{R}^K$ by a random vector defined on some underlying probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with its distribution on $\mathcal{Y}$ specified by the law P_{θ_0} that depends on parameters $\theta_0 \in \Theta$. The parameter space Θ has the form $\Theta = \mathcal{A} \times \mathcal{B} \times \mathcal{H}$, where $\mathcal{A} \subset \mathbb{R}^{L_\alpha}$, $\mathcal{B} \subset \mathbb{R}^{L_\beta}$ and $\mathcal{H}$ a metric space. We write a typical element of Θ as $\theta = (\alpha, \beta, \eta)$, where it is understood that $\alpha \in \mathcal{A}$, $\beta \in \mathcal{B}$ and $\eta \in \mathcal{H}$.

The model that the researcher considers is the collection

$$\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\} \,, \tag{14}$$

where each $P_\theta \ll \mu$ for some σ-finite measure μ on $\mathcal{Y}$. We define $\gamma = (\alpha, \beta)$ and $\Gamma = \mathcal{A} \times \mathcal{B}$, which implies that $\Gamma \subset \mathbb{R}^L$ with $L = L_\alpha + L_\beta$, and $P_\theta = P_{(\gamma,\eta)}$.

In general, we assume that the nuisance parameters β and η do not suffer from identification problems, but α may. In particular, for different points $\beta \in \mathcal{B}$ and $\eta \in \mathcal{H}$ the vector α may be strongly identified, weakly identified or completely unidentified. To conduct inference on α without making a priori assumptions on the identification of α we consider hypothesis tests of the form

$$H_0 : \alpha = \alpha_0 \,,\ \beta \in \mathcal{B} \,,\ \eta \in \mathcal{H} \qquad \text{against} \qquad H_1 : \alpha \neq \alpha_0 \,,\ \beta \in \mathcal{B} \,,\ \eta \in \mathcal{H} \,. \tag{15}$$

To derive our tests, we first define the scores of model (14) following the definition in van der Vaart (2002).

Definition A.1 (Cf. Definition 1.6 in van der Vaart, 2002)**.** *A* differentiable path *is a map $t \mapsto P_t$ from a neighborhood of $0 \in [0, \infty)$ to $\mathcal{P}_\Theta$ such that for some measurable function $s : \mathcal{Y} \to \mathbb{R}$,*

$$\int \left[\frac{\sqrt{p_t} - \sqrt{p}}{t} - \frac{1}{2} s \sqrt{p} \right]^2 \mathrm{d}\mu \to 0 \ , \tag{16}$$

where p_t and p respectively denote the densities of P_t and P relative to μ. The map $t \to \sqrt{p_t}$ is the root density path and s is the ***score function*** *of the submodel $\{P_t : t \geq 0\}$ at $t = 0$.*

In words we say that a differentiable path is a parametric submodel $\{P_t : 0 \leq t < \epsilon\}$ that is differentiable in quadratic mean at $t = 0$ with score function s. If we let $t \mapsto P_t$ range over a collection of submodels, indexed by $\mathcal{V}$, we will obtain a collection of score functions, say s_i for $i \in \mathcal{V}$. This collection, $\{s_i : i \in \mathcal{V}\}$, will be denoted by $\mathcal{T}_{P,\mathcal{V}}$ and as we only consider models with linear spaces we refer to it as a *tangent space*. For the semiparametric model (14) we define tangent spaces along restricted paths concerning the two parts of the parameter $\theta = (\gamma, \eta)$ separately.

Assumption A.1. *The map $t \mapsto P_{\gamma+tg,\eta_t(\gamma,\eta,h)}$ is a differentiable path for each $(g, h) \in \mathbb{R}^L \times H =: \mathcal{J}$. The tangent space $\mathcal{T}_{P_\theta,\mathcal{J}}$ has the form*

$$\mathcal{T}_{P_\theta,\mathcal{J}} = \mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} + \mathcal{T}^{\eta|\gamma}_{P_\theta,H} \ , \tag{17}$$

where $\mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} = \{g'\dot{\ell}_\theta : g \in \mathbb{R}^L\}$, for $\dot{\ell}_\theta$ a L-vector of measurable functions from $\mathcal{Y} \to \mathbb{R}$, is the tangent space for γ and $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ is the tangent space for η.

The assumption defines the tangent spaces for the semiparametric model (14) and imposes that the tangent space of the complete model is the sum of the tangent spaces of the parametric and non-parametric parts of the model. The assumption is mild and can typically be satisfied by imposing that the square root of the density function is continuously differentiable almost everywhere with respect to the parameters θ.[21]

For the parametric part of the model we note that $\dot{\ell}_\theta$ is simply the $L \times 1$ vector of scores of γ evaluated at $\theta = (\gamma, \eta)$, and the tangent space of γ is simply the span of $\dot{\ell}_\theta$, i.e. $\mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} = \{g'\dot{\ell}_\theta : g \in \mathbb{R}^L\}$. The tangent space of the non-parametric part, i.e. $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$, is formed by scores corresponding to paths of the form $t \mapsto P_{(\gamma,\eta_t(\gamma,\eta,h))}$ for $h \in H$, where the choice for $\eta_t(\gamma, \eta, h)$ depends on η such that $\eta_t(\gamma, \eta, h)|_{t=0} = \eta$.

Having defined the tangent spaces of γ and η, let Π_θ be the orthogonal projection from $L_2(P_\theta)$ onto the closure of $\mathcal{T}^{\eta|\alpha}_{P_\theta,H}$, i.e. $\mathrm{cl}\,\mathcal{T}^{\eta|\alpha}_{P_\theta,H}$. The *efficient score function* for γ is defined as (e.g. Definition 2.15 in van der Vaart, 2002)

$$\tilde{\ell}_\theta := \dot{\ell}_\theta - \Pi_\theta \dot{\ell}_\theta \ , \tag{18}$$

[21] See e.g. Lemma 7.6 in van der Vaart (1998), Lemma 1.8 in van der Vaart (2002) or Proposition 2.1.1 in Bickel et al. (1998).

where the projection is understood to apply componentwise. The accompanying *efficient information matrix* for γ is given by

$$\tilde{I}_\theta := \mathbb{E}_\theta \tilde{\ell}_\theta \tilde{\ell}_\theta' \, . \tag{19}$$

When η is finite dimensional the efficient score is equivalent to the population residual of the regression of $\dot{\ell}_\theta$ on the scores of η and the efficient information matrix is the variance of this residual (e.g. Neyman, 1979; Choi et al., 1996).

To obtain the efficient score function for α which is the part of $\gamma = (\alpha, \beta)$ that is of interest, note that the previous two displays imply the partitioning

$$\tilde{\ell}_\theta = \left(\tilde{\ell}_{\theta,\alpha}', \tilde{\ell}_{\theta,\beta}'\right)' \quad \text{and} \quad \tilde{I}_\theta = \begin{bmatrix} \tilde{I}_{\theta,\alpha\alpha} & \tilde{I}_{\theta,\alpha\beta} \\ \tilde{I}_{\theta,\beta\alpha} & \tilde{I}_{\theta,\beta\beta} \end{bmatrix} \, . \tag{20}$$

If $\tilde{I}_{\theta,\beta\beta}$ is nonsingular,[22] we can (orthogonally) project once more to obtain the efficient score function for α:

$$\tilde{\kappa}_\theta := \tilde{\ell}_{\theta,\alpha} - \tilde{I}_{\theta,\alpha\beta} \tilde{I}_{\theta,\beta\beta}^{-1} \tilde{\ell}_{\theta,\beta} \, , \tag{21}$$

which has corresponding efficient information matrix

$$\tilde{\mathcal{I}}_\theta := \tilde{I}_{\theta,\alpha\alpha} - \tilde{I}_{\theta,\alpha\beta} \tilde{I}_{\theta,\beta\beta}^{-1} \tilde{I}_{\theta,\beta\alpha} \, . \tag{22}$$

Building tests or estimators based on the efficient score function $\tilde{\kappa}_\theta$ is attractive as efficiency results are well established, see Choi et al. (1996), Bickel et al. (1998) and van der Vaart (2002).

It follows from (18) and Lemma 1.7 in van der Vaart (2002) that at $\theta_0 = (\alpha_0, \beta, \eta)$, where $\beta \in \mathcal{B}$ and $\eta \in \mathcal{H}$, we have

$$\mathbb{E}_{\theta_0} \tilde{\kappa}_{\theta_0} = 0 \, . \tag{23}$$

To construct test statistics we assume that we observe n independent and identically distributed copies of the vector Y that are denoted by $\{Y_i\}_{i=1}^n$. These observations satisfy the following high level assumption.

Assumption A.2. *Let $\gamma_0 = (\alpha_0, \beta)$ and $\theta_0 = (\alpha_0, \beta, \eta)$ for any $(\beta, \eta) \in \mathcal{B} \times \mathcal{H}$. Additionally, let $\gamma_n = \{(\alpha_0, \beta_n)\}_{n\in\mathbb{N}}$ be a deterministic sequence such that $\sqrt{n}(\gamma_n - \gamma_0) = O(1)$ and define $\theta_n = (\gamma_n, \eta)$ for each $n \in \mathbb{N}$. Suppose that*

1. $\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{\theta_0}(Y_i) \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{I}_{\theta_0})$ *under P_{θ_0} where $\tilde{I}_{\theta_0,\beta\beta}$ is nonsingular*

[22] If $\tilde{I}_{\theta,\beta\beta}$ is singular, we may drop components from $\tilde{\ell}_{\theta,\beta}$ until the remaining components form a linearly independent collection which span the same subspace of $L_2(P_\theta)$ as $\tilde{\ell}_{\theta,\beta}$. The corresponding variance matrix of this smaller vector will be non-singular and $\tilde{\ell}_{\theta,\beta}$ can be replaced throughout by this smaller vector.

2. *We have an array of estimates* $\{\hat{\ell}_{\gamma_n}(Y_i)\}_{n\geq 1, i\leq n}$ *such that:*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\gamma_n}(Y_i) - \tilde{\ell}_{\theta_n}(Y_i)\right) = o_{P_{\theta_n}}(n^{-1/2})$$

3. *For some sequence of estimates* $\{\hat{I}_{\gamma_n}\}_{n\geq 1}$ *and some sequence* $\{\nu_n\}_{n\geq 1}$ *with* $0 \leq \nu_n \to 0$

$$\|\hat{I}_{\gamma_n} - \tilde{I}_{\theta_0}\|_2 = o_{P_{\theta_n}}(\nu_n)$$

4. *We have that*

$$\int \left\|\tilde{\ell}_{\theta_n} p_{\theta_n}^{1/2} - \tilde{\ell}_{\theta_0} p_{\theta_0}^{1/2}\right\|^2 \,\mathrm{d}\mu \to 0.$$

We note that the estimates for the efficient scores $\hat{\ell}_{\gamma_n}(Y_i)$ and information matrix $\hat{I}_{\gamma_n}$ no longer depend on η, hence they are only indexed by γ_n. Based on Assumption A.2-parts 2 and 3 we define the following estimators for the efficient score and information matrix for α:

$$\hat{\kappa}_\gamma := \hat{\ell}_{\gamma,\alpha} - \hat{I}_{\gamma,\alpha\beta}\hat{I}_{\gamma,\beta\beta}^{-1}\hat{\ell}_{\gamma,\beta}\ , \qquad \text{and} \qquad \hat{\mathcal{I}}_\gamma := \hat{I}_{\gamma,\alpha\alpha} - \hat{I}_{\gamma,\alpha\beta}\hat{I}_{\gamma,\beta\beta}^{-1}\hat{I}_{\gamma,\beta\alpha}\ . \tag{24}$$

Given ν_n from Assumption A.2-part 3, we define a truncated eigenvalue version of the information matrix estimate as

$$\hat{\mathcal{I}}_\gamma^t = \hat{U}_n\hat{\Lambda}_n(\nu_n)\hat{U}_n'\ , \tag{25}$$

where $\hat{\Lambda}_n(\nu_n)$ is a diagonal matrix with the ν_n-truncated eigenvalues of $\hat{\mathcal{I}}_\gamma$ on the main diagonal and $\hat{U}_n$ is the matrix of corresponding orthonormal eigenvectors. To be specific, let $\{\hat{\lambda}_{n,i}\}_{i=1}^{L}$ denote the non-increasing eigenvalues of $\hat{\mathcal{I}}_\gamma$, then the (i,i)th element of $\hat{\Lambda}_n(\nu_n)$ is given by $\hat{\lambda}_{n,i}\mathbf{1}(\hat{\lambda}_{n,i} \geq \nu_n)$.

Based on this we define the singularity and identification robust score statistic as a function of $\gamma = (\alpha, \beta)$ as follows.

$$\hat{S}_\gamma := \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_\gamma(Y_i)\right)' \hat{\mathcal{I}}_\gamma^{t,\dagger}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_\gamma(Y_i)\right). \tag{26}$$

where $\hat{\mathcal{I}}_\gamma^{t,\dagger}$ is the Moore-Penrose psuedo-inverse of $\hat{\mathcal{I}}_\gamma^t$. The limiting distribution of $\hat{S}_\gamma$ is characterized in the following theorem, which implies that we can use the estimated rank of $\hat{\mathcal{I}}_\gamma^t$ to compute the critical value for $\hat{S}_\gamma$.

Theorem A.1. *Let* $\gamma_0 = (\alpha_0, \beta)$ *for any* $\beta \in \mathcal{B}$. *Suppose that* $\hat{\beta}_n$ *is a* $\sqrt{n}$*-consistent estimator of* β *under* P_{θ_0}. *Let* $B_n = n^{-1/2}C\mathbb{Z}^{L_\beta}$ *for some* $C > 0$ *and let* $\bar{\beta}_n$ *be a discretised version of* $\hat{\beta}_n$ *which replaces its value with the closest point in* B_n. *Suppose assumptions A.1 and A.2 hold and let* $\bar{\gamma}_n = (\alpha_0, \bar{\beta}_n)$. *Let* $r_n = \operatorname{rank}(\hat{\mathcal{I}}_{\bar{\gamma}_n}^t)$ *and denote by* c_n *the* $1-a$

quantile of the $\chi^2_{r_n}$ *distribution for any* $a \in (0,1)$.[23] *Then*

$$\lim_{n\to\infty} P_{\theta_0}\left(\hat{S}_{\bar{\gamma}_n} > c_n\right) \le a,$$

with inequality only if $\operatorname{rank}(\tilde{\mathcal{I}}_{\gamma_0}) = 0$.

The proof for Theorem A.1 is given below. This theorem provides the main building block for the proof of Theorem 2.4.1 for the LSEM model.

B. Proof of Theorem 2.4.1

We note that the LSEM model (2.3) can be viewed as a semi-parametric model defined by

$$\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta\} \tag{27}$$

where $\Theta = \mathcal{A} \times \mathcal{B} \times \mathcal{H}$, with $\mathcal{A} \subset \mathbb{R}^{L_\alpha}$, $\mathcal{B} \subset \mathbb{R}^{L_\beta}$ and $\mathcal{H} = \mathscr{Z} \times \prod_{k=1}^K \mathscr{H}$, where $\mathscr{Z}$ is the space of density functions η_0 with $\tilde{X}_i \sim \eta_0$ and $\mathscr{H}$ is the space of density functions η_k, i.e.

$$\begin{aligned}\mathscr{H} := \Bigg\{ g \in L_1(\lambda) \cap \mathcal{C}^1(\lambda) : g(z) \ge 0, \int g(z)\,\mathrm{d}z = 1, \int z g(z) dz = 0, \int \kappa(z) g(z)\,\mathrm{d}z = 0,\\ \int |z|^{4+\delta} g(z)\,\mathrm{d}z < \infty, \int \left|\left(g'(z)/g(z)\right)\right|^{4+\delta} g(z)\,\mathrm{d}z < \infty,\\ \int z^4 g(z)\,\mathrm{d}z > 1 + \left[\int z^3 g(z)\,\mathrm{d}z\right]^2 \Bigg\},\end{aligned}$$

where λ denotes Lebesgue measure on $\mathbb{R}$, $\mathcal{C}^1(\lambda)$ is the class of real functions on $\mathbb{R}$ which are continuously differentiable λ-a.e. and $\kappa(z) = z^2 - 1$. We denote by $\mathcal{H}_0 \subset \mathcal{H}$ the set with elements $\eta = (\eta_0, \ldots, \eta_K)$ such that each η_k satisfies the requirements imposed by assumption 2.4.1. Finally, P_θ is the law on $\mathcal{Y} \times \mathcal{X}$, with $Y_i \in \mathcal{Y} \subset \mathbb{R}^K$ and $\tilde{X}_i \in \mathcal{X} \subset \mathbb{R}^{d-1}$, defined by the density

$$p_\theta(y, \tilde{x}) := |\det A| \prod_{k=1}^K \eta_k(A_{k\bullet} y) \times \eta_0(\tilde{x})\,, \tag{28}$$

where $A_{k\bullet}$ denotes the kth row of $A = A(\alpha, \sigma)$.

With these formalities established we give three useful lemmas whose proofs are deferred to the web-appendix. The first lemma defines the tangent spaces for the LSEM and effectively ensures that the LSEM model satisfies the high-level assumption A.1 in the general theory.

Lemma B.1. *Given Assumption 2.4.1, if* $(\alpha, \sigma) \mapsto A(\alpha, \sigma)$ *is continuously differentiable,*

[23] If $r_n = 0$ we take $c_n = 0$.

we have that for any $\theta \in \Theta$ there is a $\delta > 0$ small enough such that the path $t \mapsto P_{\theta_t(\theta,g,h)}$ from $[0,\delta)$ to (a subset of) $\mathcal{P}_\Theta$ is a differentiable path with score function $y \mapsto g'\dot{\ell}_\theta(y,\tilde{x}) + h_0(\tilde{x}) + \sum_{k=1}^K h_k(A_{k\bullet}v)$, where $v = y - Bx$. In particular,

$$\mathcal{T}_{P_\theta,\mathcal{J}} = \left\{ y \mapsto g'\dot{\ell}_\theta(y,\tilde{x}) + h_0(\tilde{x}) + \sum_{k=1}^K h_k(A_{k\bullet}v) : g \in \mathbb{R}^L, h \in H \right\} = \mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} + \mathcal{T}^{\eta|\gamma}_{P_\theta,H},$$

and $\mathcal{T}_{P_\theta,\mathcal{J}}$ is a tangent space to the model at P_θ.

The next lemma presents the efficient score functions (18) for the LSEM model.

Lemma B.2. *Given Assumption 2.4.1, if $(\alpha,\sigma) \mapsto A(\alpha,\sigma)$ is continuously differentiable, the components of the efficient score function $\tilde{\ell}_\theta$ for the semiparametric linear simultaneous equations model $\mathcal{P}_\Theta$ in* (27) *at any $\theta = (\gamma,\eta)$ with $\gamma = (\alpha,\beta)$, $\alpha \in \mathcal{A}$, $\beta = (\sigma,b) \in \mathcal{B}$ and $\eta \in \mathcal{H}_0$ are given by*

$$\begin{aligned}
\tilde{\ell}_{\theta,\alpha_l}(y,\tilde{x}) &= \sum_{k=1}^K \sum_{j=1,j\neq k}^K \zeta^\alpha_{l,k,j}\phi_k(A_{k\bullet}v)A_{j\bullet}v + \sum_{k=1}^K \zeta^\alpha_{l,k,k}\left[\tau_{k,1}A_{k\bullet}v + \tau_{k,2}\kappa(A_{k\bullet}v)\right] \\
\tilde{\ell}_{\theta,\sigma_l}(y,\tilde{x}) &= \sum_{k=1}^K \sum_{j=1,j\neq k}^K \zeta^\sigma_{l,k,j}\phi_k(A_{k\bullet}v)A_{j\bullet}v + \sum_{k=1}^K \zeta^\sigma_{l,k,k}\left[\tau_{k,1}A_{k\bullet}v + \tau_{k,2}\kappa(A_{k\bullet}v)\right] \\
\tilde{\ell}_{\theta,b_l}(y,\tilde{x}) &= \sum_{k=1}^K [-A_{k\bullet}D_{b,l}]\left[(x - \mathbb{E}x)\phi_k(A_{k\bullet}v) - \mathbb{E}x\left(\varsigma_{k,1}A_{k\bullet}v + \varsigma_{k,2}\kappa(A_{k\bullet}v)\right)\right]
\end{aligned}$$

with $v = y - Bx$, $x = (1,\tilde{x}')'$, $\zeta^\alpha_{l,k,j} := [D_{\alpha,l}]_{k\bullet}A^{-1}_{\bullet j}$, $\zeta^\sigma_{l,k,j} := [D_{\sigma,l}]_{k\bullet}A^{-1}_{\bullet j}$, $D_{\alpha,l} = \partial A(\alpha,\sigma)/\partial\alpha_l$, $D_{\sigma,l} = \partial A(\alpha,\sigma)/\partial\sigma_l$ and $D_{b,l} = \partial B/\partial b_l$. Further,

$$\tau_k := M_k^{-1}\begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \varsigma_k := M_k^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{where } M_k := \begin{pmatrix} 1 & \mathbb{E}_\theta(A_{k\bullet}v)^3 \\ \mathbb{E}_\theta(A_{k\bullet}v)^3 & \mathbb{E}_\theta(A_{k\bullet}v)^4 - 1 \end{pmatrix}.$$

The proof of this lemma follows from Amari and Cardoso (1997) for $\tilde{\ell}_{\theta,\alpha_l}(y,\tilde{x})$ and $\tilde{\ell}_{\theta,\sigma_l}(y,\tilde{x})$, and for $\tilde{\ell}_{\theta,b_l}(y,\tilde{x})$ the derivations are similar to those found in, for example, Bickel et al. (1998) or Newey (1990).

The final lemma summarizes which conditions a log density score estimator should satisfy. We will apply this lemma for different choices of $W_{i,n}$ to verify our main result.

Lemma B.3. *Given assumptions 2.4.1 and 2.4.2, let $\{\beta_n\}_{n\geq 1}$ be any deterministic sequence in $\mathcal{B}$ with $\sqrt{n}(\beta_n - \beta) = O(1)$ and let $\theta_n = (\alpha_0, \beta_n, \eta)$ for some $\eta \in \mathcal{H}_0$. The log density score estimates $\hat{\phi}_k$ defined in* (2.7) *satisfy*

$$\frac{1}{n}\sum_{i=1}^n \left[\hat{\phi}_k(A_{n,k\bullet}(Y_i - B_nX_i)) - \phi_k(A_{n,k\bullet}(Y_i - B_nX_i))\right] W_{i,n} = o_{P_{\theta_n}}(n^{-1/2}), \quad (29)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_k(A_{n,k\bullet}(Y_i - B_nX_i)) - \phi_k(A_{n,k\bullet}(Y_i - B_nX_i))\right] W_{i,n}\right)^2 = o_{P_{\theta_n}}(\nu_n). \quad (30)$$

where $\{W_{i,n}\}_{n\geq 1, i\leq n}$ *is such that for each* $n \in \mathbb{N}$*, under* P_{θ_n}*, the* $W_{i,n}$ *are i.i.d. with marginal distribution given by* G_w*, with zero-mean, finite second moments and independent of each* $A_{n,k}Y_j$*.*

Proof of Theorem 2.4.1. We verify assumptions A.1 and A.2 for the LSEM under Assumptions 2.4.1 and 2.4.2.

First, the technical assumption A.1 is verified in Lemma B.1, as given above. Next, we verify each part of Assumption A.2 separately. First, we note that assumption A.2-part 1 follows by the CLT since our data is iid and the efficient score $\tilde{\ell}_{\theta_0}$ as derived in Lemma B.2 lies in $L_2(P_0)$ by construction. Next, let $\theta_n = (\alpha_0, \beta_n, \eta)$ and note that under P_{θ_n}, each $A_{n,k}(Y_i - B_nX_i) \simeq \epsilon_{i,k} \sim \eta_k$ where $A_n = A(\alpha_0, \sigma_n)$ and $A_{n,k}$ denotes the kth row of A_n. Hence we can compute certain properties of the efficient score using the equality in distribution:

$$\tilde{\ell}_{\theta_n,\alpha_l}(Y_i, \tilde{X}_i) \simeq \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j,n}\phi_k(\epsilon_{i,k})\epsilon_{i,j} + \sum_{k=1}^{K}\zeta^{\alpha}_{l,k,k,n}\left[\tau_{k,1}\epsilon_{i,k} + \tau_{k,2}\kappa(\epsilon_{i,k})\right] \quad (31)$$

$$\tilde{\ell}_{\theta_n,\sigma_l}(Y_i, \tilde{X}_i) \simeq \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\sigma}_{l,k,j,n}\phi_k(\epsilon_{i,k})\epsilon_{i,j} + \sum_{k=1}^{K}\zeta^{\sigma}_{l,k,k,n}\left[\tau_{k,1}\epsilon_{i,k} + \tau_{k,2}\kappa(\epsilon_{i,k})\right] \quad (32)$$

$$\tilde{\ell}_{\theta_n,b_l}(Y_i, \tilde{X}_i) \simeq \sum_{k=1}^{K}[-A_{n,k\bullet}D_{b_l}]\left[(X_i - \mathbb{E}X_i)\phi_k(\epsilon_{i,k}) - \mathbb{E}X_i\left(\varsigma_{k,1}\epsilon_{i,k} + \varsigma_{k,2}\kappa(\epsilon_{i,k})\right)\right] \quad (33)$$

where we note that the same observation implies that $\tau_{k,n} = \tau_k$ and $\varsigma_{k,n} = \varsigma_k$ for each n.[24] By our assumptions on the map $(\alpha, \sigma) \mapsto A(\alpha, \sigma)$, we have $\zeta^{\alpha}_{l,k,j,n} \to \zeta^{\alpha}_{l,k,j,\infty} := [D_{\alpha,l}(\gamma_0)]_{k\bullet}A(\gamma_0)^{-1}_{\bullet j}$ and $\zeta^{\sigma}_{l,k,j,n} \to \zeta^{\alpha}_{l,k,j,\infty} := [D_{\sigma,l}(\gamma_0)]_{k\bullet}A(\gamma_0)^{-1}_{\bullet j}$ for $\gamma = (\alpha_0, \beta)$. Note that the entries of $D_{b,l}$ are all zero except for entry l (corresponding to b_l) which is equal to one.

We verify assumption A.2-part 2 for each component of the efficient score (31)-(33), but we note that (31) and (32) are identical hence we concentrate on (31). For (31) and $v_n = y - B_nx$, we define

$$\varphi_{1,n}(v_n) := \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j,n}\phi_k(A_{n,k\bullet}v_n)A_{n,j\bullet}v_n\ ,$$

[24] In the preceding display we have written $\zeta^{\alpha}_{l,k,j,n}$ and $\zeta^{\sigma}_{l,k,j,n}$ rather than $\zeta^{\alpha}_{l,k,j}$ and $\zeta^{\sigma}_{l,k,j}$ to indicate their dependence on β_n. $\zeta^{\alpha}_{l,k,j,\infty}$ and $\zeta^{\sigma}_{l,k,j,\infty}$ corresponds to evaluation at the point (α_0, β).

and

$$\hat{\varphi}_{1,n}(v_n) := \sum_{k=1}^{K} \sum_{j=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j,n} \hat{\phi}_k(A_{n,k\bullet} v_n) A_{n,j\bullet} v_n \ ,$$

Let $\overline{\zeta}^{\alpha}_n := \max_{l\in[L],j\in[K],k\in[K]} |\zeta^{\alpha}_{l,j,k,n}|$ which converges to $\overline{\zeta}^{\alpha} := \max_{l\in[L],j\in[K],k\in[K]} |\zeta^{\alpha}_{l,j,k,\infty}| < \infty$. We have that

$$\sqrt{n}\mathbb{P}_n(\hat{\varphi}_{1,n} - \varphi_{1,n}) \leq \sqrt{n} \sum_{k=1}^{K} \sum_{j=1,j\neq k}^{K} \overline{\zeta}^{\alpha}_n \left| \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_k(V_{i,k,n}) V_{i,j,n} - \phi_k(V_{i,k,n}) V_{i,j,n} \right| \ ,$$

with $V_{i,j,n} = A_{n,j\bullet}(Z_i - B_n X_i)$. Since each $\left| \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_{k,n}(V_{i,k,n}) V_{i,j,n} - \phi_k(V_{i,k,n}) V_{i,j,n} \right| = o_{P_{\theta_n}}(n^{-1/2})$ by applying Lemma B.3-part (29) with $W_{i,n} = V_{i,j,n}$ (noting that under P_{θ_n}, $V_{i,k,n} \simeq \epsilon_{k,i}$ and $V_{i,j,n} \simeq \epsilon_{j,i}$ are independent with $\mathbb{E}_{\theta_n} V^2_{i,j,n} = 1$ by Assumption 2.4.1) and the outside summations are finite, it follows that

$$\sqrt{n}\mathbb{P}_n(\hat{\varphi}_{1,n} - \varphi_{1,n}) = o_{P_{\theta_n}}(1). \tag{34}$$

Next, we note that $\hat{\tau}_{k,n} - \tau_k \to 0$ and $\hat{\varsigma}_{k,n} - \varsigma_k \to 0$ in P_{θ_n}-probability by Lemma B.7 where $\hat{\tau}_{k,n}$ and $\hat{\varsigma}_{k,n}$ are defined in (2.6).

Now, consider $\varphi_{2,\tau,n}(v_n)$ defined by

$$\varphi_{2,\tau,n}(v_n) := \sum_{k=1}^{K} \zeta^{\alpha}_{l,k,k,n} \left[\tau_{k,1} A_{n,k\bullet} v_n + \tau_{k,2} \kappa(A_{n,k\bullet} v_n) \right].$$

Since sum is finite and each $|\zeta^{\alpha}_{l,k,k,n}| \to |\zeta^{\alpha}_{l,k,k,\infty}| < \infty$ it is sufficient to consider the convergence of the summands. In particular we have that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\hat{\tau}_{k,n,1} - \tau_{k,1}\right] V_{i,k,n} = \left[\hat{\tau}_{k,n,1} - \tau_{k,1}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{i,k,n} = o_{P_{\theta_n}}(1) \times O_{P_{\theta_n}}(1) = o_{P_{\theta_n}}(1),$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\hat{\tau}_{k,n,2} - \tau_{k,2}\right] \kappa(V_{i,k,n}) = \left[\hat{\tau}_{k,n,2} - \tau_{k,2}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \kappa(V_{i,k,n}) = o_{P_{\theta_n}}(1) \times O_{P_{\theta_n}}(1) = o_{P_{\theta_n}}(1).$$

since $V_{i,k,n} \simeq \epsilon_{k,i} \sim \eta_k$ under P_{θ_n} and $(\epsilon_{i,k})_{i\geq 1}$ and $(\kappa(\epsilon_{i,k}))_{i\geq 1}$ are i.i.d. mean-zero sequences with finite second moments such that the CLT holds. Together these yield that

$$\sqrt{n}\mathbb{P}_n(\varphi_{2,\hat{\tau}_n,n} - \varphi_{2,\tau,n}) = o_{P_{\theta_n}}(1). \tag{35}$$

Putting (34) and (35) together yields the required convergence for components of the type (31) , since $\tilde{\ell}_{\theta_n,\alpha_l} = \varphi_{1,n} + \varphi_{2,\tau,n}$ and $\hat{\ell}_{\gamma_n,\alpha_l} = \hat{\varphi}_{1,n} + \varphi_{2,\hat{\tau}_n,n}$. The same holds for (32); the only difference is that we replace $\zeta^{\alpha}_{l,k,k,n}$ by $\zeta^{\sigma}_{l,k,k,n}$

Next, we consider components (33). Let $a_{n,k,l} := -A_{n,k\bullet}D_{b,l}$ and write

$$\begin{aligned}\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\gamma_n,b,l} - \tilde{\ell}_{\theta_n,b,l}\right] &= \sum_{k=1}^{K} a_{n,k,l}\sqrt{n}\mathbb{P}_n\left[(X_i - \mathbb{E}X_i)[\hat{\phi}_k(V_{i,k,n}) - \phi_k(V_{i,k,n})] + (\mathbb{E}X_i - \bar{X}_n)\phi_k(V_{i,k,n})\right] \\ &\quad + \sum_{k=1}^{K} a_{n,k,l}\sqrt{n}\mathbb{P}_n\left[(\mathbb{E}X_i - \bar{X}_n)[\hat{\varsigma}_{k,n,1}V_{i,k,n} + \hat{\varsigma}_{k,n,2}\kappa(V_{i,k,n})]\right] \\ &\quad - \sum_{k=1}^{K} a_{n,k,l}\sqrt{n}\mathbb{P}_n\left[\mathbb{E}X_i[(\hat{\varsigma}_{k,n,1} - \varsigma_{k,1})V_{i,k,n} + (\hat{\varsigma}_{k,n,2} - \varsigma_{k,2})\kappa(V_{i,k,n})]\right]\end{aligned}$$

Taking the right hand side terms (inside the outer summation) in order, we have that $\sqrt{n}\mathbb{P}_n(X_i - \mathbb{E}X_i)[\hat{\phi}_k(V_{i,k,n}) - \phi_k(V_{i,k,n})] = o_{P_{\theta_n}}(1)$ by Lemma B.3-part (29) applied with $W_{i,n} = X_i - \mathbb{E}X_i$. For the second, $\sqrt{n}\mathbb{P}_n(\mathbb{E}X_i - \bar{X}_n)\phi_k(V_{i,k,n}) = (\mathbb{E}X_i - \bar{X}_n)\sqrt{n}\mathbb{P}_n\phi_k(V_{i,k,n}) = o_{P_{\theta_n}}(1) \times O_{P_{\theta_n}}(1) = o_{P_{\theta_n}}(1)$ by the WLLN & CLT, noting for the latter that $V_{i,k,n} \simeq \epsilon_{i,k}$. We know from Lemma B.7 that $\varsigma_{k,n} \xrightarrow{P_{\theta_n}} \varsigma_k$ and hence adding & subtracting and using the WLLN & CLT again yields that $\sqrt{n}\mathbb{P}_n(\mathbb{E}X_i - \bar{X}_n)[\hat{\varsigma}_{k,n,1}V_{i,k,n} + \hat{\varsigma}_{k,n,2}\kappa(V_{i,k,n})] = o_{P_{\theta_n}}(1)$. The CLT & $\varsigma_{k,n} \xrightarrow{P_{\theta_n}} \varsigma_k$ ensure that $\sqrt{n}\mathbb{P}_n[(\hat{\varsigma}_{k,n,1} - \varsigma_{k,1})V_{i,k,n} + (\hat{\varsigma}_{k,n,2} - \varsigma_{k,2})\kappa(V_{i,k,n})] = o_{P_{\theta_n}}(1)$. Together these observations and that $a_{n,k,l} \to a_{\infty,n,l} := A_{k\bullet}D_{b,l}$ imply that the required condition, $\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\gamma_n,b,l} - \tilde{\ell}_{\theta_n,b,l}\right] = o_{P_{\theta_n}}(1)$, is satisfied.

To verify part 3 we will show that

$$\left\|\hat{I}_{\gamma_n} - \tilde{I}_{\theta_0}\right\|_2 \leq \left\|\hat{I}_{\gamma_n} - \tilde{I}_{\theta_n}\right\|_2 + \left\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\right\|_2 = o_{P_{\theta_n}}(\nu_n^{1/2}). \tag{36}$$

where $\tilde{I}_{\theta_n} := \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}_{\theta_n}(Y_i)\tilde{\ell}_{\theta_n}(Y_i)'$. To obtain the rates we start with $\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2$, for which we show that each component satisfies the required rate. To set this up, let $Q^{r,s}_{l,m,i,n} = \tilde{\ell}_{\theta_n,r_l}(Y_i)\tilde{\ell}_{\theta_n,s_m}(Y_i) - \tilde{\ell}_{0,r_l}(Y_i)\tilde{\ell}_{\theta_0,s_m}(Y_i)$, where $r,s \in \{\alpha,\sigma,b\}$ and l,m denote the indices of the components of the efficient scores. Let $\breve{Q}^{r,s}_{l,m,i,n}$ be defined analogously with $V_{i,k,n}$ replaced by $\epsilon_{i,k}$. Under P_{θ_n} we have that $Q^{r,s}_{l,m,i,n} \simeq \breve{Q}^{r,s}_{l,m,i,n}$. Therefore to show $[\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}]_{l,m} = o_{P_{\theta_n}}(\nu_n^{1/2})$ it suffices to show that for any r,s and l,m

$$\frac{1}{n}\sum_{i=1}^{n}\breve{Q}^{r,s}_{l,m,i,n} - G\breve{Q}^{r,s}_{l,m,i,n} + \frac{1}{n}\sum_{i=1}^{n}G[\breve{Q}^{r,s}_{l,m,i,n} - \breve{Q}^{r,s}_{l,m,i,\infty}] = o_G(\nu_n^{1/2}),$$

where G is the product measure $\prod_{k=0}^{K}G_k$ and each $\breve{Q}^{r,s}_{l,m,i,n}$ is shown to satisfy $\|\breve{Q}^{r,s}_{l,m,i,n}\|_{G,p} < \infty$ in Lemma B.6 given below. The convergence of the second term follows from the assumed Lipschitz continuity of the map defining the ζ's and the $\sqrt{n}$-consistency of β_n for β, since $n^{-1/2} = o(\nu_n^{1/2})$.[25] For the first term, if $p = 2$ in lemma B.6, by Theorem

[25] Note that for large enough $n \in \mathbb{N}$ β_n is in a ball of radius, say, $\delta > 0$ around β. The (continuous) differentiability of $(\alpha,\beta_1) \mapsto A(\alpha,\beta_1)$ and the fact that $D_{b,l}$ is a constant matrix implies that the map $(\alpha,\beta_1) \mapsto [-A(\alpha,\beta_1)_{k\bullet}D_{b,l}]$ is Lipschitz on this set.

2.5.11 in Durrett (2019), we have that for all $\iota > 0$

$$\frac{1}{n}\sum_{i=1}^{n} \breve{Q}^{r,s}_{l,m,i,n} - G\breve{Q}^{r,s}_{l,m,i,n} = o_G\left(n^{-1/2}\log(n)^{1/2+\iota}\right).$$

It follows that

$$\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2 \leq \|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_F = o_{P_{\theta_n}}\left(n^{-1/2}\log(n)^{1/2+\iota}\right).$$

If, instead, $p = 1 + \nu/4 < 2$ in Lemma B.6, then by the Marcinkiewicz & Zygmund SLLN (e.g. Theorem 2.5.12 in Durrett, 2019)

$$\frac{1}{n}\sum_{i=1}^{n} \breve{Q}^{r,s}_{l,m,i,n} - G\breve{Q}^{r,s}_{l,m,i,n} = o_G\left(n^{\frac{1-p}{p}}\right),$$

and similarly

$$\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2 \leq \|\tilde{I}_{\theta_n,n} - \tilde{I}_{\theta_0}\|_F = o_{P_{\theta_n}}\left(n^{\frac{1-p}{p}}\right).$$

That is, for any $p \in (1, 2]$ we have $\|\tilde{I}_{\theta_n} - \tilde{I}_{\theta_0}\|_2 = o_{P_{\theta_n}}(\nu_{n,p}) = o_{P_{\theta_n}}(\nu_n^{1/2})$.

For the other component of the sum, let $r \in \{\alpha, \sigma, b\}$ and let l denote an index, we write $\hat{U}_{n,i,r_l} := \hat{\ell}_{\gamma_n,r_l}(Y_i)$, $\tilde{U}_{i,r_l} := \tilde{\ell}_{\theta_n,r_l}(Y_i)$ and $D_{n,i,r_l} := \hat{\ell}_{\gamma_n,r_l}(Y_i) - \tilde{\ell}_{\theta_n,r_l}(Y_i)$.

Since it is the absolute value of the $(r, l) - (s, m)$ component of $\hat{I}_{\gamma_n,n} - \tilde{I}_{\theta_0,n}$, it is sufficient to show that $\left|\frac{1}{n}\sum_{i=1}^n \hat{U}_{n,i,r,l}D_{n,i,s,m} + \frac{1}{n}\sum_{i=1}^n D_{n,i,r,l}\tilde{U}_{i,s,m}\right| = o_{P_{\theta_n}}(\nu_n^{1/2})$ as $n \to \infty$ for any $r, s \in \{(\alpha, \sigma), b\}$ and l, m. By Cauchy-Schwarz and lemma B.8

$$\left|\frac{1}{n}\sum_{i=1}^{n} D_{n,i,r,l}\tilde{U}_{i,s,m}\right| \leq \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{U}^2_{i,s,m}\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}D^2_{n,i,r,l}\right)^{1/2} = O_{P_{\theta_n}}(1)\times o_{P_{\theta_n}}(\nu_n^{1/2}) = o_{P_{\theta_n}}(\nu_n^{1/2}),$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} \hat{U}_{n,i,r,l}D_{n,i,s,m}\right| \leq \left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}^2_{n,i,r,l}\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}D^2_{n,i,s,m}\right)^{1/2} = O_{P_{\theta_n}}(1)\times o_{P_{\theta_n}}(\nu_n^{1/2}) = o_{P_{\theta_n}}(\nu_n^{1/2}),$$

for any $(r, l) - (s, m)$. It follows that

$$\left[\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{n,i,r,l}D_{n,i,s,m} + D_{n,i,r,l}\tilde{U}_{i,s,m}\right]^2 \leq 2\left[\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{n,i,r,l}D_{n,i,s,m}\right]^2 + 2\left[\frac{1}{n}\sum_{i=1}^{n}D_{n,i,r,l}\tilde{U}_{i,s,m}\right]^2 = o_{P_{\theta_n}}(\nu_n)$$

and hence $\|\hat{I}_{\gamma_n,n} - \tilde{I}_{\theta_0,n}\|_2 \leq \|\hat{I}_{\gamma_n,n} - \tilde{I}_{\theta_0,n}\|_F = o_{P_{\theta_n}}(\nu_n^{1/2})$. We can combine these results to obtain:

$$\|\hat{I}_{\gamma_n,n} - \tilde{I}_{\theta_0}\|_2 \leq \|\hat{I}_{\gamma_n,n} - \tilde{I}_{\theta_n,n}\|_2 + \|\tilde{I}_{\theta_n,n} - \tilde{I}_{\theta_0}\|_2 = o_{P_{\theta_n}}(\nu_n^{1/2}) + o_{P_{\theta_n}}(\nu_n^{1/2}) = o_{P_{\theta_n}}(\nu_n^{1/2}).$$

It remains to show that part 4 of Assumption A.2 holds. Recall that the dominating measure here is λ and re-write the integral in question as

$$\int \left\| \tilde{\ell}_{\theta_n} p_{\theta_n}^{1/2} - \tilde{\ell}_{\theta_0} p_{\theta_0}^{1/2} \right\|^2 \mathrm{d}\lambda = \sum_{l=1}^{L} \int \left[\tilde{\ell}_{\theta_n,l} p_{\theta_n}^{1/2} - \tilde{\ell}_{\theta_0,l} p_{\theta_0}^{1/2} \right]^2 \mathrm{d}\lambda. \qquad (37)$$

It is evidently sufficient to show that each of the integrals in the sum on the rhs converges to zero. To this end, let $f_{r,n} := \tilde{\ell}_{\theta_n,r,l} p_{\theta_n}^{1/2}$ and $f_r := \tilde{\ell}_{\theta_0,r_l} p_{\theta_0}^{1/2}$ for $r \in \{\alpha, \sigma, b\}$ corresponding to (31)-(33) for some arbitrary l. By the expressions for $\tilde{\ell}_{\theta_n}$ and p_{θ_n} given in lemma B.2 and equation (28) respectively along with the continuity of A, D_l and each η_k and ϕ_k (each of which follows from our assumptions), we have that $f_{r,n} \to f_r$ λ-a.e. for all r. Moreover, using the representation in (31) we have

$$\begin{aligned}
\int f_{\alpha,n}^2 \,\mathrm{d}\lambda &= \int \left(\sum_{k=1}^{K} \left[\zeta_{l,k,k,n}^{\alpha} \left[\tau_{k,1} \epsilon_{k,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \right] + \sum_{j=1, j \neq k}^{K} \zeta_{l,k,j,n}^{\alpha} \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \right] \right)^2 \mathrm{d}G \\
&= \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \sum_{b=1}^{K} \sum_{m=1, m \neq b}^{K} \zeta_{l,k,j,n}^{\alpha} \zeta_{l,b,m,n}^{\alpha} \int \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \phi_b(\epsilon_{b,i}) \epsilon_{m,i} \,\mathrm{d}G \\
&\quad + 2 \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \sum_{b=1}^{K} \zeta_{l,k,j,n}^{\alpha} \zeta_{l,b,b,n}^{\alpha} \int \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \left[\tau_{b,1} \epsilon_{b,i} + \tau_{b,2} \kappa(\epsilon_{b,i}) \right] \mathrm{d}G \\
&\quad + \sum_{k=1}^{K} \sum_{b=1}^{K} \zeta_{l,k,k,n}^{\alpha} \zeta_{l,b,b,n}^{\alpha} \int \left[\tau_{b,1} \epsilon_{b,i} + \tau_{b,2} \kappa(\epsilon_{b,i}) \right] \left[\tau_{k,1} \epsilon_{k,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \right] \mathrm{d}G
\end{aligned}$$

where G is the law of ϵ and each of the integrals are finite by assumption 14. By the continuity of A and D_l, this converges to

$$\begin{aligned}
\int f_{\alpha}^2 \,\mathrm{d}\lambda &= \int \left(\sum_{k=1}^{K} \left[\zeta_{l,k,k,\infty}^{\alpha} \left[\tau_{k,1} \epsilon_{k,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \right] + \sum_{j=1, j \neq k}^{K} \zeta_{l,k,j,\infty}^{\alpha} \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \right] \right)^2 \mathrm{d}G \\
&= \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \sum_{b=1}^{K} \sum_{m=1, m \neq b}^{K} \zeta_{l,k,j,\infty}^{\alpha} \zeta_{l,b,m,\infty}^{\alpha} \int \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \phi_b(\epsilon_{b,i}) \epsilon_{m,i} \,\mathrm{d}G \\
&\quad + 2 \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \sum_{b=1}^{K} \zeta_{l,k,j,\infty}^{\alpha} \zeta_{l,b,b,\infty}^{\alpha} \int \phi_k(\epsilon_{k,i}) \epsilon_{j,i} \left[\tau_{b,1} \epsilon_{b,i} + \tau_{b,2} \kappa(\epsilon_{b,i}) \right] \mathrm{d}G \\
&\quad + \sum_{k=1}^{K} \sum_{b=1}^{K} \zeta_{l,k,k,\infty}^{\alpha} \zeta_{l,b,b,\infty}^{\alpha} \int \left[\tau_{b,1} \epsilon_{b,i} + \tau_{b,2} \kappa(\epsilon_{b,i}) \right] \left[\tau_{k,1} \epsilon_{k,i} + \tau_{k,2} \kappa(\epsilon_{k,i}) \right] \mathrm{d}G,
\end{aligned}$$

which is finite by assumption 2.4.1. By Proposition 2.29 in van der Vaart (1998) we conclude that $\int (f_{\alpha,n} - f_\alpha)^2 \,\mathrm{d}\lambda \to 0$. Analogous arguments hold for $r = \sigma, b$; we omit the details. The convergence of each $\int (f_{r,n} - f_r)^2 \,\mathrm{d}\lambda \to 0$ in conjunction with equation (37) is sufficient for part 4. □

B.1. Supporting Lemmas

Lemma B.4. *Suppose that assumption 2.4.1 holds and let $k, j, s, b \in [K]$ with $j \neq k$ and $s \neq b$. Then, for G the law of ϵ and any $p \in [1, 2]$ we have that*

(I) $\|\phi_k(\epsilon_k)\epsilon_j\phi_s(\epsilon_s)\epsilon_b\|_{G,p} < \infty$,

(II) $\|\phi_k(\epsilon_k)\epsilon_j\epsilon_s\|_{G,p} < \infty$,

(III) $\|\epsilon_k\epsilon_s\|_{G,p} < \infty$.

Proof. By Cauchy-Schwarz, independence and our moment conditions we have

$$\begin{aligned}\|\phi_k(\epsilon_k)\epsilon_j\phi_s(\epsilon_s)\epsilon_b\|_{G,p} &\leq \left[G[\phi_k(\epsilon_k)]^{2p}G[\epsilon_j]^{2p}G[\phi_s(\epsilon_s)]^{2p}G[\epsilon_b]^{2p}\right]^{\frac{1}{2p}} < \infty,\\ \|\phi_k(\epsilon_k)\epsilon_j\epsilon_s\|_{G,p} &\leq \left[G[\phi_k(\epsilon_k)]^{2p}G[\epsilon_j]^{2p}G[\epsilon_s]^{2p}\right]^{1/(2p)} < \infty,\\ \|\epsilon_k\epsilon_s\|_{G,p} = \|(\epsilon_k)^p(\epsilon_s)^p\|_{G,1}^{1/p} &\leq \|(\epsilon_k)^p\|_{G,2}^{1/p}\|(\epsilon_s)^p\|_{G,2}^{1/p} < \infty.\end{aligned}$$

□

Lemma B.5. *Suppose that assumption 2.4.1 holds and let $k, j, s \in [K]$ with $j \neq k$. Then, for G the law of ϵ and $1 \leq p \leq \min(1 + \delta/4, 2)$, we have*

(I) $\|\phi_k(\epsilon_k)\epsilon_j\kappa(\epsilon_s)\|_{G,p} < \infty$,

(II) $\|\epsilon_k\kappa(\epsilon_s)\|_{G,p} < \infty$,

(III) $\|\kappa(\epsilon_k)\kappa(\epsilon_s)\|_{G,p} < \infty$.

Proof. By Cauchy-Schwarz, independence and our assumed moment conditions we have

$$\begin{aligned}\|\phi_k(\epsilon_k)\epsilon_j\kappa(\epsilon_s)\|_{G,p} &\leq \left[\left[G[\phi_k(\epsilon_k)]^{2p}G[\epsilon_s]^{4p}\right]^{1/(2p)} + \|\phi_k(\epsilon_k)\|_{G,p}\right]\|\epsilon_j\|_{G,p} < \infty,\\ \|\epsilon_k\kappa(\epsilon_s)\|_{G,p} &\leq \|(\epsilon_k)^p\|_{G,2}^{1/p}\|(\epsilon_s)^{2p}\|_{G,2}^{1/p} + \|\epsilon_k\|_{G,p} < \infty,\\ \|\kappa(\epsilon_k)\kappa(\epsilon_s)\|_{G,p} &\leq \|(\epsilon_k)^{2p}\|_{G,2}^{1/p}\|(\epsilon_s)^{2p}\|_{G,2}^{1/p} + 2\|(\epsilon_k)^2\|_{G,p} + 2\|(\epsilon_s)^2\|_{G,p} + 1 < \infty.\end{aligned}$$

□

Lemma B.6. *Define*

$$\begin{aligned}
q^{\alpha}_{l,i,n} &:= \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i} + \sum_{k=1}^{K}\zeta^{\alpha}_{l,k,k,n}\left[\tau_{k,1}\epsilon_{k,i} + \tau_{k,2}\kappa(\epsilon_{k,i})\right]\\
q^{\sigma}_{l,i,n} &:= \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\sigma}_{l,k,j,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i} + \sum_{k=1}^{K}\zeta^{\sigma}_{l,k,k,n}\left[\tau_{k,1}\epsilon_{k,i} + \tau_{k,2}\kappa(\epsilon_{k,i})\right]\\
q^{b}_{l,i,n} &:= -\sum_{k=1}^{K}[A_{n,k\bullet}D_{b,l}]\left[(X_i - \mathbb{E}X_i)\phi_k(\epsilon_{k,i}) - \mathbb{E}X_i(\varsigma_{k,1}\epsilon_{k,i} + \varsigma_{k,2}\kappa(\epsilon_{k,i}))\right]
\end{aligned}$$

where the dependence of e.g. $\zeta^{\alpha}_{l,k,j,n}$ *on* n *is as in the proof of Theorem 2.4.1.*[26] *Let* $\breve{Q}^{r,s}_{l,m,i,n} := q^{r}_{l,i,n}q^{s}_{m,i,n}$*. Suppose that assumption 2.4.1 holds. Then, for* $1 \leq p \leq \min(1+\delta/4, 2)$ *we have* $\|\breve{Q}^{r,s}_{l,m,i,n}\|_{G,p} < \infty$ *for G the law of* $(\tilde{X}, \epsilon)$*.*

Proof. By definition we have

$$\begin{aligned}
\breve{Q}^{\alpha,\alpha}_{l,m,i,n} =& \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\sum_{s=1}^{K}\sum_{b=1,b\neq s}^{K} \zeta^{\alpha}_{l,k,j,n}\zeta^{\alpha}_{m,s,b,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i}\phi_s(\epsilon_{s,i})\epsilon_{b,i}\\
&+ 2\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\sum_{s=1}^{K}\zeta^{\alpha}_{l,k,j,n}\zeta^{\alpha}_{m,s,s,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i}[\tau_{s,1}\epsilon_{s,i} + \tau_{s,2}\kappa(\epsilon_{s,i})]\\
&+ \sum_{k=1}^{K}\sum_{s=1}^{K}\zeta^{\alpha}_{l,k,k,n}\zeta^{\alpha}_{m,s,s,n}[\tau_{k,1}\epsilon_{k,i} + \tau_{k,2}\kappa(\epsilon_{k,i})][\tau_{s,1}\epsilon_{s,i} + \tau_{s,2}\kappa(\epsilon_{s,i})].\\
\breve{Q}^{\alpha,b}_{l,m,i,n} =& -\sum_{s=1}^{K}\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i}[A_{n,s\bullet}D_{b,l}](X_i - \mathbb{E}X_i)\phi_s(\epsilon_{s,i})\\
&+ \sum_{s=1}^{K}\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j,n}\phi_k(\epsilon_{k,i})\epsilon_{j,i}[A_{n,s\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{s,1}\epsilon_{s,i} + \varsigma_{s,2}\kappa(\epsilon_{s,i}))\\
&- \sum_{s=1}^{K}\sum_{k=1}^{K}\zeta^{\alpha}_{l,k,k,n}[\tau_{k,1}\epsilon_{k,i} + \tau_{k,2}\kappa(\epsilon_{k,i})][A_{n,s\bullet}D_{b,l}](X_i - \mathbb{E}X_i)\phi_s(\epsilon_{s,i})\\
&+ \sum_{s=1}^{K}\sum_{k=1}^{K}\zeta^{\alpha}_{l,k,k,n}[\tau_{k,1}\epsilon_{k,i} + \tau_{k,2}\kappa(\epsilon_{k,i})][A_{n,s\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{s,1}\epsilon_{s,i} + \varsigma_{s,2}\kappa(\epsilon_{s,i}))\\
\breve{Q}^{b,b}_{l,m,i,n} =& \sum_{s=1}^{K}\sum_{k=1}^{K}[A_{n,s\bullet}D_{b,l}](X_i - \mathbb{E}X_i)\phi_s(\epsilon_{s,i})[A_{n,k\bullet}D_{b,l}](X_i - \mathbb{E}X_i)\phi_k(\epsilon_{k,i})\\
&+ 2\sum_{s=1}^{K}\sum_{k=1}^{K}[A_{n,s\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{s,1}\epsilon_{s,i} + \varsigma_{s,2}\kappa(\epsilon_{s,i}))[A_{n,k\bullet}D_{b,l}](X_i - \mathbb{E}X_i)\phi_k(\epsilon_{k,i})\\
&+ \sum_{s=1}^{K}\sum_{k=1}^{K}[A_{n,s\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{s,1}\epsilon_{s,i} + \varsigma_{s,2}\kappa(\epsilon_{s,i}))[A_{n,k\bullet}D_{b,l}]\mathbb{E}X_i(\varsigma_{k,1}\epsilon_{k,i} + \varsigma_{k,2}\kappa(\epsilon_{k,i}))
\end{aligned}$$

[26] See footnote 24.

Hence, by Minkowski's inequality, the independence of ϵ from $\tilde{X}$ (with finite second moments) and lemmas B.4 & B.5, $\|\breve{Q}^{r,s}_{l,m,i,n}\|_{G,p} < \infty$, noting that for σ instead of α we have the same expressions. □

Lemma B.7. *Suppose assumption 2.4.1 holds and $\nu_{n,p}$ and ν_n are as in assumption 2.4.2. Then $\|\hat{\varkappa}_{k,n} - \varkappa_{k,n}\|_2 = o_{P_{\theta_n}}(\nu_{n,p}) = o_{P_{\theta_n}}(\nu_n^{1/2})$ for $\varkappa \in \{\tau, \varsigma\}$.*

Proof. Under P_{θ_n}, $A_{n,k\bullet}(Z_i - B_n X_i) \simeq \epsilon_{k,i} \sim \eta_k$, hence the claim will follow if we show that $\check{\varkappa}_{k,n} - \breve{\varkappa}_k = o_{G_k}(\nu_n^{1/2})$, where

$$\check{\varkappa}_{k,n} := \check{M}_{k,n}^{-1} w, \quad \text{where } \check{M}_{k,n} := \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^n (\epsilon_{k,i})^3 \\ \frac{1}{n}\sum_{i=1}^n (\epsilon_{k,i})^3 & \frac{1}{n}\sum_{i=1}^n (\epsilon_{k,i})^4 - 1 \end{pmatrix},$$

$$\breve{\varkappa}_{k,n} := \breve{M}_{k,n}^{-1} w, \quad \text{where } \breve{M}_{k,n} := \begin{pmatrix} 1 & G_k(\epsilon_{k,i})^3 \\ G_k(\epsilon_{k,i})^3 & G_k(\epsilon_{k,i})^4 - 1 \end{pmatrix},$$

and $w \in \mathbb{R}^2$. By the preceding definitions and the fact that the map $M \mapsto M^{-1}$ is Lipschitz at a positive definite matrix M_0 we have that for a positive constant C then for large enough n, with probability approaching one

$$\|\check{\varkappa}_{k,n} - \breve{\varkappa}_{k,n}\|_2 = \|(\check{M}_{k,n}^{-1} - \breve{M}_k^{-1})w\|_2 \leq \|w\|_2 \|\check{M}_{k,n}^{-1} - \breve{M}_k^{-1}\|_2 \lesssim C\|\check{M}_{k,n} - \breve{M}_k\|_2. \quad (38)$$

If $\upsilon := \delta/4 \geq 1$, we have that by Theorem 2.5.11 in Durrett (2019)

$$\frac{1}{n}\sum_{i=1}^n [(\epsilon_{k,i})^3 - G_k(\epsilon_{k,i})^3] = o_{G_k}\left(n^{-1/2}\log(n)^{1/2+\iota}\right)$$
$$\frac{1}{n}\sum_{i=1}^n [(\epsilon_{k,i})^4 - G_k(\epsilon_{k,i})^4] = o_{G_k}\left(n^{-1/2}\log(n)^{1/2+\iota}\right)$$

for $\iota > 0$, which implies that

$$\|\check{M}_{k,n} - \breve{M}_k\|_2 \leq \|\check{M}_{k,n} - \breve{M}_k\|_F = o_{G_k}\left(n^{-1/2}\log(n)^{1/2+\iota}\right).$$

If $0 < \upsilon < 1$, we have by Theorems 2.5.11 & 2.5.12 in Durrett (2019) that for $\iota > 0$,

$$\frac{1}{n}\sum_{i=1}^n [(\epsilon_{k,i})^3 - G_k(\epsilon_{k,i})^3] = \begin{cases} o_{G_k}\left(n^{-1/2}\log(n)^{1/2+\iota}\right) & \text{if } \upsilon \in [1/2, 1) \\ o_{G_k}\left(n^{\frac{1-p}{p}}\right) & \text{if } \upsilon \in (0, 1/2) \end{cases},$$
$$\frac{1}{n}\sum_{i=1}^n [(\epsilon_{k,i})^4 - G_k(\epsilon_{k,i})^4] = o_{G_k}\left(n^{\frac{1-p}{p}}\right).$$

which together imply that

$$\|\check{M}_{k,n} - \breve{M}_k\|_2 \leq \|\check{M}_{k,n} - \breve{M}_k\|_F = o_{G_k}\left(n^{\frac{1-p}{p}}\right).$$

Combining these convergence rates with equation (38) yields the result in light of the observations made at the beginning of the proof. □

Lemma B.8. *Suppose assumptions 2.4.1 and 2.4.2 hold and $\theta_n = (\alpha_0, \beta_n, \eta)$ where $\sqrt{n}(\beta_n - \beta) = O(1)$ is a deterministic sequence. Then for each $r \in \{\alpha, \sigma, b\}$ and l*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\gamma_n, r_l}(Y_i) - \tilde{\ell}_{\theta_n, r_l}(Y_i)\right)^2 = o_{P_{\theta_n}}(\nu_n).$$

Proof. In this proof we let $M_k := M_{k\bullet}$ for any matrix M. We start by considering elements in $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\gamma_n,\alpha_l}(Y_i) - \tilde{\ell}_{\theta_n,\alpha_l}(Y_i)\right)^2$ (noting that the result for σ will be the same). We define $\tilde{\tau}_{k,n,q} := \hat{\tau}_{k,n,q} - \tau_{k,q}$ and $V_{i,n} = Z_i - B_n X_i$. Since each $|\zeta^{\alpha}_{l,k,j,n}| < \infty$ and the sums over k, j are finite, it is sufficient to demonstrate that for every $k, j, m, s \in [K]$, with $k \neq j$ and $s \neq m$,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right]\left[\hat{\phi}_{s,n}(A_{n,s}V_{i,n}) - \phi_s(A_{n,s}V_{i,n})\right] A_{n,j}V_{i,n}A_{n,m}V_{i,n} = o_{P_{\theta_n}}(\nu_n), \tag{39}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right] A_{n,j}V_{i,n}\left[\tilde{\tau}_{s,n,1}A_{n,s}V_{i,n} + \tilde{\tau}_{s,n,2}\kappa(A_{n,s}V_{i,n})\right] = o_{P_{\theta_n}}(\nu_n), \tag{40}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\tau}_{s,n,1}A_{n,s}V_{i,n} + \tilde{\tau}_{s,n,2}\kappa(A_{n,s}V_{i,n})\right]\left[\tilde{\tau}_{k,n,1}A_{n,k}V_{i,n} + \tilde{\tau}_{k,n,2}\kappa(A_{n,k}V_{i,n})\right] = o_{P_{\theta_n}}(\nu_n). \tag{41}$$

For (41), let $\xi_1(x) = x$ and $\xi_2(x) = \kappa(x)$. Then, we can split the sum into 4 parts, each of which has the following form for some $q, w \in \{1, 2\}$

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w}\xi_q(A_{n,s}V_{i,n})\xi_w(A_{n,k}V_{i,n}) = \tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w}\frac{1}{n}\sum_{i=1}^{n}\xi_q(A_{n,s}V_{i,n})\xi_w(A_{n,k}V_{i,n}) = o_{P_{\theta_n}}(\nu_n),$$

since we have that each $\tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w} = o_{P_{\theta_n}}(\nu_n)$ by lemma B.7.[27] For (40) we can argue similarly. Again let $\xi_1(x) = x$ and $\xi_2(x) = \kappa(x)$. Then, we can split the sum into 2 parts, each of which has the following form for some $q \in \{1, 2\}$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right] A_{n,j}V_{i,n}\tilde{\tau}_{s,n,q}\xi_q(A_{n,s}V_{i,n}) \\
&\quad \leq \tilde{\tau}_{s,n,q}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right]^2 (A_{n,j}V_{i,n})^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_q(A_{n,s}V_{i,n})^2\right)^{1/2} \\
&\quad = o_{P_{\theta_n}}(\nu_n).
\end{aligned}$$

[27] The fact that $\frac{1}{n}\sum_{i=1}^{n}\xi_q(A_{n,s}V_{i,n})\xi_w(A_{n,k}V_{i,n}) = O_{P_{\theta_n}}(1)$ can be seem to hold using the moment and i.i.d. assumptions from assumption 2.4.1 and Markov's inequality, noting once more that $A_{n,k}V_{i,n} \simeq \epsilon_{k,i}$ under P_{θ_n}.

by Lemma B.3 applied with $W_{i,n} = A_{n,j}V_{i,n}$ and $\tilde{\tau}_{s,n,q} = o_{P_{\theta_n}}(\nu_n^{1/2})$.[28] For (39) use Cauchy-Schwarz with lemma B.3:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right]\left[\hat{\phi}_{s,n}(A_{n,s}V_{i,n}) - \phi_s(A_{n,s}V_{i,n})\right]A_{n,j}V_{i,n}A_{n,m}V_{i,n} \\
&\quad\leq \left(\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k}V_{i,n}) - \phi_k(A_{n,k}V_{i,n})\right]^2 (A_{n,j}V_{i,n})^2\right)^{1/2} \\
&\qquad\qquad \times \left(\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_{s,n}(A_{n,s}V_{i,n}) - \phi_s(A_{n,s}V_{i,n})\right]^2 (A_{n,m}V_{i,n})^2\right)^{1/2} \\
&\quad= o_{P_{\theta_n}}(\nu_n).
\end{aligned}$$

Finally, we consider the elements in $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\gamma_n,b_l}(Y_i) - \tilde{\ell}_{\theta_n,b_l}(Y_i)\right)^2$, where we let $a_{n,k,l} := -A_{n,k}D_{b,l}$ and note that

$$\begin{aligned}
&\hat{\ell}_{\gamma_n,b_l}(Y_i) - \tilde{\ell}_{\theta_n,b_l}(Y_i) \\
&\quad= \sum_{k=1}^{K} a_{n,k,l}\left[(X_i - \mathbb{E}X_i)[\hat{\phi}_k(V_{i,k,n}) - \phi_k(V_{i,k,n})] + (\mathbb{E}X_i - \bar{X}_n)\phi_k(V_{i,k,n})\right] \\
&\qquad+ \sum_{k=1}^{K} a_{n,k,l}\left[(\mathbb{E}X_i - \bar{X}_n)[\hat{\varsigma}_{k,n,1}V_{i,k,n} + \hat{\varsigma}_{k,n,2}\kappa(V_{i,k,n})]\right] \\
&\qquad- \sum_{k=1}^{K} a_{n,k,l}\left[\mathbb{E}X_i[(\hat{\varsigma}_{k,n,1} - \varsigma_{k,1})V_{i,k,n} + (\hat{\varsigma}_{k,n,2} - \varsigma_{k,2})\kappa(V_{i,k,n})]\right]
\end{aligned}$$

We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\ell}_{\gamma_n,b_l}(Y_i) - \tilde{\ell}_{\theta_n,b_l}(Y_i)\right)^2 \\
&\quad\lesssim \sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}[a_{n,k,l}(X_i - \mathbb{E}X_i)]^2[\hat{\phi}_k(V_{i,k,n}) - \phi_k(V_{i,k,n})]^2 + [a_{n,k,l}(\mathbb{E}X_i - \bar{X}_n)]^2\phi_k(V_{i,k,n})^2 \\
&\qquad+ \sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}[a_{n,k,l}(\mathbb{E}X_i - \bar{X}_n)]^2[\hat{\varsigma}_{k,n,1}V_{i,k,n} + \hat{\varsigma}_{k,n,2}\kappa(V_{i,k,n})]^2 \\
&\qquad+ \sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}[a_{n,k,l}\mathbb{E}X_i]^2[(\hat{\varsigma}_{k,n,1} - \varsigma_{k,1})V_{i,k,n} + (\hat{\varsigma}_{k,n,2} - \varsigma_{k,2})\kappa(V_{i,k,n})]^2
\end{aligned}$$

The first term is $o_{P_{\theta_n}}(\nu_n)$ by Cauchy-Schwarz and applying lemma B.3, the second and third terms follows from $(a_{n,k,l}(\bar{X}_n - \mathbb{E}X_i))^2 = O_{P_{\theta_n}}(n^{-1}) = o_{P_{\theta_n}}(\nu_n)$ and the fourth term follows from Lemma B.7. □

[28] See footnote 27.

B.2. Proof of Theorem A.1

Proof of Theorem A.1. Let $P_0 := P_{\theta_0}$, where θ_0 is defined in Assumption A.2. The first step is to show that assumption A.2 implies that

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\gamma_n} - \tilde{\ell}_{\theta_n}\right] \xrightarrow{P_0} 0, \quad \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\theta_n} - \tilde{\ell}_{\theta_0}\right] + \sqrt{n}\tilde{I}_{\theta_0}(0,(\beta_n-\beta)')' \xrightarrow{P_0} 0 \tag{42}$$

and

$$\nu_n^{-1}\left\|\hat{I}_{\gamma_n} - \tilde{I}_{\theta_0}\right\| = o_{P_0}(1). \tag{43}$$

To do so, define $b_n := \sqrt{n}(\beta_n - \beta)$ and let $(n_m)_{m\geq 1}$ be an arbitrary subsequence of $(n)_{n\geq 1}$. It is sufficient for (42)-(43) that we can demonstrate that there is a further subsequence $(n_{m(k)})_{k\geq 1}$ along which the claimed convergence holds. There exists a sub-subsequence such that $b_{n_{m(k)}} \to b$ for some $b \in \mathbb{R}^{L_\beta}$.[29] Taking such a subsequence will suffice as we will now demonstrate that the claimed convergence holds for an arbitrary convergent sequence $b_n \to b$.

Let Q_n^n denote the law of $(Y_i)_{i=1}^n$ corresponding to θ_n and P_0^n that corresponding to θ_0. Let $\Lambda_n(Q_n, P_0) = n\mathbb{P}_n \log q_n - \log p_0$ be the corresponding log-likelihood ratio. In view of the differentiability in quadratic mean of the model (e.g. Definition 1) we have by van der Vaart and Wellner, 1996, lemma 3.10.11:

$$\Lambda_n(Q_n, P) = \sqrt{n}\mathbb{P}_n b'\dot{\ell}_{\theta_0,\beta} - \frac{1}{2}b'\dot{I}_{\theta_0,\beta\beta}b + R_n,$$

where $R_n \to 0$ in probability under both P_0^n and Q_n^n and $\dot{I}_{\theta_0} = \mathbb{V}(\dot{\ell}_{\theta_0})$. Noting that $\dot{\ell}_{\theta_0}$ is a score by assumption A.1 and hence in $L_2(P_0)$ (e.g. van der Vaart, 2002, Lemma 1.7) it follows by the CLT that

$$\Lambda_n(Q_n, P) \rightsquigarrow \mathcal{N}\left(-\frac{1}{2}b'\dot{I}_{\theta_0,\beta\beta}b, b'\dot{I}_{\theta_0,\beta\beta}b\right),$$

under P_0, from which we can conclude that $P_0^n \lhd\rhd Q_n^n$ (e.g. van der Vaart and Wellner, 1996, example 3.10.6). This mutual contiguity and Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) ensure that left claim in (42) and (43) hold given parts 2 & 3 of assumption A.2. Noting that $P_0[\tilde{\ell}_{\theta_0}\dot{\ell}'_{\theta_0,\beta}]b = \tilde{I}_{\theta_0}(0,b')'$, the right claim of equation (42) follows by proposition A.10 in van der Vaart (1988b), which requires Assumption A.2-part 4.[30]

Next we show that (42) and (43) continue to hold if γ_n (and $\theta_n = (\gamma_n, \eta)$) is replaced by $\bar{\gamma}_n$ (and $\bar{\theta}_n = (\bar{\gamma}_n, \eta)$) as defined in the theorem.[31] Since $\bar{\beta}_n$ remains $\sqrt{n}$-consistent there is an $M > 0$ such that $P_0\left(\sqrt{n}\|\bar{\beta}_n - \beta\| > M\right) < \varepsilon$. If $\sqrt{n}\|\bar{\beta}_n - \beta\| \leq M$ then

[29] Such a subsequence and b exist by the Bolzano-Weierstrass theorem.

[30] Cf. lemma 7.3 in van der Vaart (2002); the proof of theorem 25.57 in van der Vaart (1998).

[31] The proof is adapted from the proof of Theorem 5.48 in van der Vaart (1998).

the discretized estimator $\bar{\beta}_n$ is equal to one of the values in the finite set $B_n = \{\beta' \in n^{-1/2}C\mathbb{Z}^{L_\beta} : \|\beta' - \beta\| \leq n^{-1/2}M\}$. For each M this set has finite number of elements bounded independently of n, call this upper bound $\overline{B}$. Let

$$R_n'(\beta') := \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\gamma'} - \tilde{\ell}_{\theta'}\right], \; R_n''(\beta') := \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\theta'} - \tilde{\ell}_{\theta_0}\right] + \sqrt{n}\tilde{I}_{\theta_0}(0,(\beta'-\beta)')', \; R_n'''(\beta') := \nu_n^{-1}[\hat{I}_{\gamma'} - \tilde{I}_{\theta_0}],$$

where $\gamma' = (\alpha_0, \beta')$ and $\theta' = (\gamma', \eta)$. Letting R_n denote either R_n', R_n'' or R_n''' we have that for any $\upsilon > 0$

$$\begin{aligned}
P_0\left(\|R_n(\bar{\beta}_n)\| > \upsilon\right) &\leq \varepsilon + \sum_{\beta_n \in B_n} P_0\left(\{\|R_n(\beta_n)\| > \upsilon\} \cap \{\bar{\beta}_n = \beta_n\}\right) \\
&\leq \varepsilon + \sum_{\beta_n \in B_n} P_0\left(\|R_n(\beta_n)\| > \upsilon\right) \\
&\leq \varepsilon + \overline{B}P_0(\|R_n(\beta_n^*)\| > \upsilon),
\end{aligned}$$

where $\beta_n^* \in B_n$ maximises $\beta \mapsto P_0\left(\|R_n(\beta_n)\| > \upsilon\right)$. As $(\beta_n^*)_{n\in\mathbb{N}}$ is a deterministic $\sqrt{n}$-consistent sequence for β we have that $P_0(\|R_n(\beta_n^*)\| > \upsilon) \to 0$ by equations (42) and (43).

By the version of (42) with γ_n, θ_n replaced by $\bar{\gamma}_n, \bar{\theta}_n$ we have

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\gamma}_n} - \tilde{\ell}_{\theta_0}\right] = \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\gamma}_n} - \tilde{\ell}_{\bar{\theta}_n}\right] + \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_0}\right] = -\tilde{I}_{\theta_0}(0, \sqrt{n}(\bar{\beta}_n - \beta)')' + o_{P_0}(1).$$

and by the version of (43) with γ_n, θ_n replaced by $\bar{\gamma}_n, \bar{\theta}_n$, $\hat{I}_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{I}_{\theta_0}$ and so $\hat{\mathcal{K}}_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{K}}_{\theta_0}$ for

$$\tilde{\mathcal{K}}_\theta := \begin{bmatrix} I & -\tilde{I}_{\theta,\alpha\beta}\tilde{I}_{\theta,\beta\beta}^{-1} \end{bmatrix}, \quad \hat{\mathcal{K}}_\gamma := \begin{bmatrix} I & -\hat{I}_{\gamma,\alpha\beta}\hat{I}_{\gamma,\beta\beta}^{-1} \end{bmatrix}.$$

We combine these to obtain

$$\begin{aligned}
&\sqrt{n}\mathbb{P}_n\left[\hat{\kappa}_{\bar{\gamma}_n} - \tilde{\kappa}_{\theta_0}\right] \\
&\quad = \left(\hat{\mathcal{K}}_{\bar{\gamma}_n} - \tilde{\mathcal{K}}_{\theta_0}\right)\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\gamma}_n} - \tilde{\ell}_{\theta_0}\right] + \tilde{\mathcal{K}}_{\theta_0}\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\gamma}_n} - \tilde{\ell}_{\theta_0}\right] + \left(\hat{\mathcal{K}}_{\bar{\gamma}_n} - \tilde{\mathcal{K}}_{\theta_0}\right)\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\theta_0} \\
&\quad = -\tilde{\mathcal{K}}_{\theta_0}\tilde{I}_{\theta_0}(0, \sqrt{n}(\bar{\beta}_n - \beta)')' + o_{P_0}(1) \\
&\quad = -\begin{bmatrix} I & -\tilde{I}_{\theta_0,\alpha\beta}\tilde{I}_{\theta_0,\beta\beta}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{I}_{\theta_0,\alpha\alpha} & \tilde{I}_{\theta_0,\alpha\beta} \\ \tilde{I}_{\theta_0,\beta\alpha} & \tilde{I}_{\theta_0,\beta\beta} \end{bmatrix} \begin{bmatrix} 0 \\ \sqrt{n}(\bar{\beta}_n - \beta) \end{bmatrix} + o_{P_0}(1) \\
&\quad = o_{P_0}(1).
\end{aligned}$$

Then, by assumption A.2-part 1, under P_0,

$$Z_n := \sqrt{n}\mathbb{P}_n\hat{\kappa}_{\bar{\gamma}_n} = \sqrt{n}\mathbb{P}_n\left[\hat{\kappa}_{\bar{\gamma}_n} - \tilde{\kappa}_{\theta_0}\right] + \sqrt{n}\mathbb{P}_n\tilde{\kappa}_{\theta_0} \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{\mathcal{I}}_{\theta_0}).$$

For the next step, observe that

$$\left\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\right\|_2 \leq \left\|\hat{I}_{\bar{\gamma}_n,\alpha\alpha} - \tilde{I}_{\theta_0,\alpha\alpha}\right\|_2 + \left\|\hat{I}_{\bar{\gamma}_n,\alpha\beta}\hat{I}^{-1}_{\bar{\gamma}_n,\beta\beta}\hat{I}_{\bar{\gamma}_n,\beta\alpha} - \tilde{I}_{\theta_0,\alpha\beta}\tilde{I}^{-1}_{\theta_0,\beta\beta}\tilde{I}_{\theta_0,\beta\alpha}\right\|_2 .$$

By repeated addition and subtraction along with the observations that any submatrix has a smaller operator norm than the original matrix and the matrix inverse is Lipschitz continuous at a non-singular matrix we obtain

$$\left\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\right\|_2 \lesssim \left\|\hat{I}_{\bar{\gamma}_n} - \tilde{I}_{\theta_0}\right\|_2 .$$

Hence by equation (43) with $\bar{\gamma}_n$ replacing γ_n we have $P_0\left(\left\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\right\|_2 < \nu_n\right) \to 1$.

The remainder of the proof is split into two cases. First consider the case where $\operatorname{rank}(\tilde{\mathcal{I}}_{\theta_0}) = r > 0$. We first show that $\hat{\mathcal{I}}^t_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$ and the rank estimate $r_n = \operatorname{rank}(\hat{\mathcal{I}}^t_{\bar{\gamma}_n})$ satisfies $P_0(\{r_n = r\}) \to 1$.

Let λ_l denote the lth largest eigenvalue of $\tilde{\mathcal{I}}_{\theta_0}$, similarly define $\hat{\lambda}_{l,n}$ for $\hat{\mathcal{I}}_{\bar{\gamma}_n}$ and $\hat{\lambda}^t_{l,n}$ for $\hat{\mathcal{I}}^t_{\bar{\gamma}_n}$. Define the set $R_n := \{r_n = r\}$, let $\underline{\nu} := \lambda_r/2 > 0$ and note that $\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(\nu_n)$ implies that $\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(1)$.

By Weyl's perturbation theorem[32] we have $\max_{l=1,\ldots,L_\alpha}|\hat{\lambda}_{l,n} - \lambda_l| \leq \|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 = o_{P_0}(1)$. Hence, if we define $E_n := \{\hat{\lambda}_{r,n} \geq \nu_n\}$, for n large enough such that $\nu_n < \underline{\nu}$, we have

$$P_0(E_n) = P_0\left(\hat{\lambda}_{r,n} \geq \nu_n\right) \geq P_0\left(\hat{\lambda}_{r,n} \geq \underline{\nu}\right) \geq P_0\left(|\hat{\lambda}_{r,n} - \lambda_r| < \underline{\nu}\right) \to 1.$$

If $r = L_\alpha$ we have that $R_n \supset E_n$ and therefore $P_0(R_n) \to 1$. Additionally, if $\hat{\lambda}_{L_\alpha,n} \geq \nu_n$ then $\hat{\lambda}^t_{l,n} = \hat{\lambda}_{l,n}$ for each $l \in [L_\alpha]$ and hence $\hat{\mathcal{I}}^t_{\bar{\gamma}_n} = \hat{\mathcal{I}}_{\bar{\gamma}_n}$. Thus, $E_n \cap \{\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\| \leq \upsilon\} \subset \{\|\hat{\mathcal{I}}^t_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\| \leq \upsilon\}$, from which it follows that $\hat{\mathcal{I}}^t_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$.

Now suppose instead that $r < L_\alpha$ and define $F_n := \{\hat{\lambda}_{r+1,n} < \nu_n\}$. It follows by Weyl's perturbation theorem and the fact that $\lambda_l = 0$ for $l > r$ that as $n \to \infty$

$$P(F_n) = P(\hat{\lambda}_{r+1,n} < \nu_n) \geq P(\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 < \nu_n) \to 1.$$

Since $R_n \supset E_n \cap F_n$, this implies that $P(R_n) \to 1$ as $n \to \infty$. Additionally, if $\hat{\lambda}_{r,n} \geq \nu_n$, $\hat{\lambda}_{r+1,n} < \nu_n$ and $\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \upsilon$, we have that $\hat{\lambda}^t_{k,n} = \hat{\lambda}_{k,n}$ for $k \leq r$ and $\hat{\lambda}^t_{l,n} = 0 = \lambda_l$ for $l > r$ and so

$$\|\hat{\Lambda}_n(\nu_n) - \Lambda\|_2 = \max_{l=1,\ldots,r}|\hat{\lambda}^t_{l,n} - \lambda_l| = \max_{l=1,\ldots,r}|\hat{\lambda}_{l,n} - \lambda_l| \leq \|\hat{\Lambda}_n - \Lambda\|_2 \leq \|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \upsilon,$$

and hence $\{\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \upsilon\} \cap E_n \cap F_n \subset \{\|\hat{\Lambda}_n(\nu_n) - \Lambda\|_2 \leq \upsilon\}$, from which it follows

[32] E.g. Corollary III.2.6 in Bhatia (1997).

that $\hat{\Lambda}_n(\nu_n) \xrightarrow{P_0} \Lambda$.

To complete this part of the proof, suppose that $(\lambda_1, \dots, \lambda_r)$ consists of s distinct eigenvalues with values $\lambda^1 > \lambda^2 > \cdots > \lambda^s$ and multiplicities $\mathfrak{m}_1, \dots, \mathfrak{m}_s$ (each at least one), where the superscripts on the λs are indices, not exponents. $\lambda^{s+1} = 0$ is an eigenvalue with multiplicity $\mathfrak{m}_{s+1} = L_\alpha - r$. Let l_i^k for $k = 1, \dots, s+1$ and $i = 1, \dots, \mathfrak{m}_k$ denote the column indices of the eigenvectors in U corresponding to each λ^k. For each λ^k, the total eigenprojection is $\Pi_k := \sum_{i=1}^{\mathfrak{m}_k} u_{l_i^k} u'_{l_i^k}$.[33] Total eigenprojections are continuous.[34] Therefore, if we construct $\hat{\Pi}_{k,n}$ in in an analogous fashion to Π_k but replace columns of U with columns of $\hat{U}_n$, we have $\hat{\Pi}_{k,n} \xrightarrow{P_0} \Pi_k$ for each $k \in [s+1]$ since $\hat{\mathcal{I}}_{\bar{\gamma}_n} \xrightarrow{P_0} \tilde{\mathcal{I}}_{\theta_0}$. Spectrally decompose $\tilde{\mathcal{I}}_{\theta_0}$ as $\tilde{\mathcal{I}}_{\theta_0} = \sum_{k=1}^{s} \lambda^k \Pi_k$, where the sum runs to s rather than $s+1$ since $\lambda^{s+1} = 0$. Then,

$$\hat{\mathcal{I}}_{\bar{\gamma}_n}^t = \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} \hat{\lambda}_{l_i^k,n}^t \hat{u}_{l_i^k,n} \hat{u}'_{l_i^k,n} = \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} (\hat{\lambda}_{l_i^k,n}^t - \lambda^k) \hat{u}_{l_i^k,n} \hat{u}'_{l_i^k,n} + \sum_{k=1}^{s} \lambda^k \hat{\Pi}_{k,n},$$

and so

$$\|\hat{\mathcal{I}}_{\bar{\gamma}_n}^t - \tilde{\mathcal{I}}_{\theta_0}\|_2 \leq \sum_{k=1}^{s+1} \sum_{i=1}^{\mathfrak{m}_k} |\hat{\lambda}_{l_i^k,n}^t - \lambda^k| \|\hat{u}_{l_i^k,n} \hat{u}'_{l_i^k,n}\|_2 + \sum_{k=1}^{s} |\lambda^k| \|\hat{\Pi}_{k,n} - \Pi_k\|_2 \xrightarrow{P_0} 0,$$

by $\hat{\Pi}_{k,n} \xrightarrow{P} \Pi_k$, $\hat{\Lambda}_n(\nu_n) \xrightarrow{P_0} \Lambda$ and since we have $\|u_{l_i^k,n} u'_{l_i^k,n}\|_2 = 1$ for any i, k, n.

Hence, we have that $\hat{\mathcal{I}}_{\bar{\gamma}_n}^t \overset{P_0}{\to} \tilde{\mathcal{I}}_{\theta_0}$ and $P_0(\{r_n = r\}) \to 1$. This implies that $\hat{\mathcal{I}}_{\bar{\gamma}_n}^{t,\dagger} \overset{P_0}{\to} \tilde{\mathcal{I}}_{\theta_0}^{\dagger}$ where $\tilde{\mathcal{I}}_{\theta_0}^{\dagger}$ is the Moore-Penrose inverse of $\tilde{\mathcal{I}}_{\theta_0}$.[35]

Now consider the score statistic $\hat{S}_{\bar{\gamma}_n}$, by Slutsky's lemma and the continuous mapping theorem we have that

$$\hat{S}_{\bar{\gamma}_n} = Z_n' \hat{\mathcal{I}}_{\bar{\gamma}_n}^{t,\dagger} Z_n \rightsquigarrow Z' \tilde{\mathcal{I}}_{\theta_0}^{\dagger} Z \sim \chi_r^2$$

where the distributional result $X := Z' \tilde{\mathcal{I}}_{\theta_0}^{\dagger} Z \sim \chi_r^2$, follows from e.g. Theorem 9.2.2 in Rao and Mitra (1971).

Finally, recall that $R_n = \{r_n = r\}$. On these sets c_n is the $1 - a$ quantile of the χ_r^2 distribution, which we will call c. Hence, we have $c_n \xrightarrow{P_0} c$ as $P_0(R_n) \to 1$. As a result, we obtain $\hat{S}_{\bar{\gamma}_n} - c_n \rightsquigarrow X - c$ where $X \sim \chi_r^2$. Since the χ_r^2 distribution is continuous, we have by the Portmanteau theorem

$$P_0\left(\hat{S}_{\bar{\gamma}_n} > c_n\right) = 1 - P_0\left(\hat{S}_{\bar{\gamma}_n} - c_n \leq 0\right) \to 1 - P_0\left(X - c \leq 0\right) = 1 - P_0\left(X \leq c\right) = a\,,$$

which completes the proof in the case that $r > 0$.

[33] See e.g Chapter 8.8 of Magnus and Neudecker (2019).

[34] E.g. Theorem 8.7 of Magnus and Neudecker (2019).

[35] See e.g. Theorem 2 of Andrews (1987).

It remains to handle the case with $r = 0$. We first note that $Z_n \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{\mathcal{I}}_{\theta_0})$ continues to hold by our assumptions, though in this case $\tilde{\mathcal{I}}_{\theta_0}$ is the zero matrix and hence the limiting distribution is degenerate: $Z = 0$ a.s.. Let $E_n = \{r_n = 0\}$. Part 3 of assumption A.2 and Weyl's perturbation theorem imply that

$$P_0(E_n) = P_0\left(r_n = 0\right) = P_0\left(\max_{l=1,\ldots,L_\alpha} |\hat{\lambda}_{n,l}| < \nu_n\right) \geq P_0\left(\|\hat{\mathcal{I}}_{\bar{\gamma}_n} - \tilde{\mathcal{I}}_{\theta_0}\|_2 < \nu_n\right) \to 1.$$

On the sets E_n we have that $\hat{\mathcal{I}}^t_{\bar{\gamma}_n}$ is the zero matrix, whose Moore-Penrose inverse is also the zero matrix. Hence on the sets E_n we have $\hat{S}_{\bar{\gamma}_n} = 0$ and $c_n = 0$ and therefore do not reject, implying

$$P_0(\hat{S}_{\bar{\gamma}_n} > c_n) \leq 1 - P_0(E_n) \to 0.$$

It follows that $P_0(\hat{S}_{\bar{\gamma}_n} > c_n) \to 0$. □

C. Proofs for Lemmas 1-3

In this section we provide the proofs for lemmas B.1-B.3. The proofs of these lemmas depend on a number of supporting results which can be found in section C.1. Many of these results are standard but are nevertheless included for convenience.

Proof of Lemma B.1. The log density for the semiparametric LSEM is given by

$$\ell_\theta(y, \tilde{x}) \coloneqq \log p_\theta(y, \tilde{x}) = \log |A| + \sum_{k=1}^{K} \log \eta_k(A_{k\bullet}(y - Bx)) + \log \eta_0(\tilde{x})\ .$$

For convenience let $v = v_\theta \coloneqq y - Bx$ with $x = (1, \tilde{x})$. We define $\dot{\ell}_\theta(y, \tilde{x}) \coloneqq \nabla_\gamma \ell_\theta(y, \tilde{x})$, where we recall that γ partitions as $\gamma = (\alpha, \beta)$, with $\beta = (\sigma, b)$, and some derivations show that the components of $\dot{\ell}_\theta(y, \tilde{x})$ can be written as

$$\begin{aligned}
\dot{\ell}_{\theta,\alpha_l}(y, \tilde{x}) &= \operatorname{tr}(A^{-1} D_{\alpha,l}(\alpha, \sigma)) + \sum_{k=1}^{K} \phi_k(A_{k\bullet} v) \times [D_{\alpha,l}(\alpha, \sigma)]_{k\bullet} v \\
&= \operatorname{tr}(D_{\alpha,l}(\alpha, \sigma) A^{-1}) + \sum_{k=1}^{K}\sum_{j=1}^{K} \phi_k(A_{k\bullet} v) \times \left([D_{\alpha,l}(\alpha, \sigma)]_{k\bullet} A^{-1}_{\bullet j}\right) A_{j\bullet} v \\
&= \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta^{\alpha}_{l,k,j} \phi_k(A_{k\bullet} v) A_{j\bullet} v + \sum_{k=1}^{K} \zeta^{\alpha}_{l,k,k} \left(\phi_k(A_{k\bullet} v) A_{k\bullet} v + 1\right), \qquad (44) \\
\dot{\ell}_{\theta,\sigma_l}(y, \tilde{x}) &= \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta^{\sigma}_{l,k,j} \phi_k(A_{k\bullet} v) A_{j\bullet} v + \sum_{k=1}^{K} \zeta^{\sigma}_{l,k,k} \left(\phi_k(A_{k\bullet} v) A_{k\bullet} v + 1\right), \\
\dot{\ell}_{\theta,b_l}(y, \tilde{x}) &= \sum_{k=1}^{K} \phi_k(A_{k\bullet} v) \times [-A_{k\bullet} D_{b,l} x]\ ,
\end{aligned}$$

where $\zeta^{\alpha}_{l,k,j} := [D_{\alpha,l}]_{k\bullet} A^{-1}_{\bullet j}$, $\zeta^{\sigma}_{l,k,j} := [D_{\sigma,l}]_{k\bullet} A^{-1}_{\bullet j}$, $D_{\alpha,l} = \partial A(\alpha,\sigma)/\partial\alpha_l$, $D_{\sigma,l} = \partial A(\alpha,\sigma)/\partial\sigma_l$ and $D_{b,l} = \partial B/\partial b_l$. Paths of the form $t \to P_{\gamma+tg,\eta}$ have an associated tangent space given by

$$\mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L} = \{g'\dot{\ell}_\theta(y,\tilde{x}) : g \in \mathbb{R}^L\} \; . \tag{45}$$

To constructing the tangent space of the non-parametric part we consider submodels of the following form. Let

$$\eta^{h_k}_{k,t}(\cdot) = \eta_k(\cdot)(1 + th_k(\cdot)) \qquad k = 0,\dots,K \; ,$$

which for $t = 0$ recover η_k. For $k = 1,\dots,K$, h_k is some function such that $h_k \in H_k$ with

$$H_k := \left\{h_k \in \mathcal{C}^1_b(\lambda) : \mathbb{E}h_k(\epsilon_k) = 0, \mathbb{E}\epsilon_k h_k(\epsilon_k) = 0, \mathbb{E}\kappa(\epsilon_k)h_k(\epsilon_k) = 0\right\} \; , \tag{46}$$

where $\mathcal{C}^1_b(\lambda)$ denotes the space of functions from $\mathbb{R} \to \mathbb{R}$ which are bounded and continuously differentiable with bounded derivatives λ-a.e.. Letting G_k be the law on $\mathbb{R}$ corresponding to η_k for $k = 1,\dots,K$, it is clear that H_k is a linear subspace of $L_2(G_k)$. The additional restrictions on h_k ensure that for t small enough $\eta_{k,t} \in \mathscr{H}$. For $k = 0$, define

$$H_0 := \left\{h_0 \in \mathcal{C}_b(\lambda, \mathbb{R}^{d-1}) : \mathbb{E}h_0(\tilde{X}) = 0\right\} \; , \tag{47}$$

where $\mathcal{C}_b(\lambda,\mathbb{R}^{d-1})$ denotes the space of bounded λ-a.e. continuous functions from $\mathbb{R}^{d-1} \to \mathbb{R}$.[36] Letting G_0 be the law on $\mathbb{R}^{d-1}$ corresponding to η_0, it is clear that H_0 is a linear subspace of $L_2(G_0)$. The additional restrictions on h_0 ensure that for t small enough $\eta_{0,t} \in \mathscr{Z}$. Now let $H := \prod_{k=0}^K H_k$. For any $h = (h_0, h_1, \dots, h_K) \in H$ and any $\theta \in \Theta$ we can define a path $\eta_t(\theta,h) := (\eta^{h_0}_{0,t}, \eta^{h_1}_{1,t}, \dots, \eta^{h_K}_{K,t})$. Given the preceding discussion, for each $h \in H$ there is a $\delta > 0$ small enough such that $\eta^{h_0}_{0,t} \in \mathscr{Z}$ and $\eta^{h_k}_{k,t} \in \mathscr{H}$ for each $k = 1,\dots,K$ when $t \in (-\delta,\delta)$. Now, we use this to define a path $\theta_t(\theta,h) := (\gamma, \eta_t(\theta,h))$. Then, $p_{\theta_t(\gamma,h)}$ defines a path towards p_θ according to:

$$p_{\theta_t(\theta,h)}(y,\tilde{x}) = |\det A| \times \prod_{k=1}^K \eta^{h_k}_{k,t}(A_{k\bullet}v) \times \eta^{h_0}_{0,t}(\tilde{x}) \; . \tag{48}$$

Given the discussion above, for $t \in (-\delta,\delta)$, the submodel $\{P_{\theta_t(\theta,h)} : t \in (-\delta,\delta)\} \subset \mathcal{P}_\Theta$. Let $s : \mathbb{R}^K \to \mathbb{R}$ be given by

$$\begin{aligned} s(y,\tilde{x}) := \left.\frac{\partial \log p_{\theta_t(\theta,h)}(y,\tilde{x})}{\partial t}\right|_{t=0} &= \left.\frac{h_0(\tilde{x})}{1 + th_0(\tilde{x})}\right|_{t=0} + \sum_{k=1}^K \left.\frac{h_k(A_{k\bullet}v)}{1 + th_k(A_{k\bullet}v)}\right|_{t=0} \\ &= h_0(\tilde{x}) + \sum_{k=1}^K h_k(A_{k\bullet}v) \; . \end{aligned} \tag{49}$$

[36] We make no notational distinction between the Lebesgue measure on $\mathbb{R}$ and that on $\mathbb{R}^{d-1}$; which is meant can be inferred from context.

s is a score function associated to the differentiable path $t \mapsto P_{\theta_t(\theta,h)}$ from $[0, \delta) \to \mathcal{P}_\Theta$ and the associated tangent space for η is given by

$$\mathcal{T}^{\eta|\gamma}_{P_\theta,H} := \left\{ y \mapsto h_0(\tilde{x}) + \sum_{k=1}^{K} h_k(A_{k\bullet}v) : h = (h_0, h_1, \dots, h_K) \in H \right\}. \tag{50}$$

These calculations establish the form of the score functions for the parameteric part and non-parametric part of the model separately. To verify assumption A.1 we rather need to consider the (joint) paths given by $\theta_t(\theta, g, h) = (\gamma + tg, \eta_t(\theta, h))$.

By the definitions of $\mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L}$ and $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ given in (45) and (50) respectively and the fact that both $\mathbb{R}^L$ and H are linear spaces, it follows that $\mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L}$ and $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ are linear spaces, implying that the same is true of their sum. Therefore, provided we show that $\mathcal{T}_{P_\theta,\mathcal{J}}$ is a tangent *set* to the model at P_θ and that it is the sum of $\mathcal{T}^{\gamma|\eta}_{P_\theta,\mathbb{R}^L}$ and $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$, we immediately obtain that it is a tangent *space*. That the second equality in the display in the statement of the lemma holds is clear by the definition of a sum of linear subspaces and the form of the elements on the right hand side given in equations (45) and (50). So it remains to prove the first equality. That is, for any $g \in \mathbb{R}^L$ and $h \in H$ there is a small enough $\delta > 0$ such that the path $t \mapsto P_{\theta_t(\theta,g,h)}$ from $[0, \delta)$ to (a subset of) $\mathcal{P}_\Theta$ is a differentiable path with score function $y \mapsto g'\dot{\ell}_\theta(y) + h_0(\tilde{x}) + \sum_{k=1}^{K} h_k(A_{k\bullet}v)$. Fix $g \in \mathbb{R}^L, h \in H$ and $\theta \in \Theta$ and let θ_t abbreviate $\theta_t(\theta, g, h)$. Recall that γ partitions as $\gamma = ((\alpha, \sigma), b)$ and let $g = (g_1, g_2)$ be the conforming partition for any $g \in \mathbb{R}^L$. Further, let G_2 be such that $g_2 = \operatorname{vec}(G_2)$. Additionally throughout the proof we will let $M_k = M_{k\bullet}$ for any matrix M and to save on notation, we define $\tilde{A}(t) := A((\alpha', \sigma')' + tg_1)$, $\tilde{B}(t) = B + G_2 t$, $\tilde{v}(t) := y - \tilde{B}(t)x$ and $\tilde{D}_k(t) := \frac{\mathrm{d}[\tilde{A}(a)]_k \tilde{v}(a)}{\mathrm{d}a}(t)$.

We will now compute the (pointwise) derivative of $t \mapsto \ell_{\theta_t}(y, \tilde{x}) := \log p_{\theta_t}(y, \tilde{x})$ on $(-\delta, \delta)$. We have that

$$\begin{aligned} \ell_{\theta_t}(y, \tilde{x}) = & \log|\det \tilde{A}(t)| + \log \eta_0(\tilde{x}) + \sum_{k=1}^{K} \log \eta_k \left([\tilde{A}(t)]_k \tilde{v}(t) \right) \\ & + \log\left(1 + th_0(\tilde{x})\right) + \sum_{k=1}^{K} \log\left(1 + th_k\left([\tilde{A}(t)]_k \tilde{v}(t)\right)\right). \end{aligned}$$

For sufficiently small t (i.e. there is some neighbourhood $(-\delta, \delta)$ on which) the arguments of the logarithms on the second line are positive. We proceed by repeatedly applying the chain rule to conclude that

$$\begin{aligned} \acute{\ell}_{\theta_t}(y, \tilde{x}) := \frac{\partial \ell_{\theta_t}(y, \tilde{x})}{\partial t} = & \operatorname{tr}\left(\left[\tilde{A}(t)\right]^{-1} \frac{\mathrm{d}\tilde{A}(t)}{\mathrm{d}t} \right) + \frac{h_0(\tilde{x})}{1 + th_0(\tilde{x})} + \sum_{k=1}^{K} \left[\phi_k\left([\tilde{A}(t)]_k \tilde{v}(t) \right) \times \tilde{D}_k(t) \right] \\ & + \sum_{k=1}^{K} \frac{h_k([\tilde{A}(t)]_k \tilde{v}(t)) + th_k'([\tilde{A}(t)]_k \tilde{v}(t)) \times \tilde{D}_k(t)}{1 + th_k([\tilde{A}(t)]_k \tilde{v}(t))}, \end{aligned}$$

for all $y, \tilde{x}$ such that $p_{\theta_t}(y, \tilde{x}) > 0$ and define it as 0 elsewhere. Use (44) to evaluate the preceding display at $t = 0$ and obtain (for y such that $p_{\theta_t}(y, \tilde{x}) > 0$ and set it to 0 otherwise):

$$\begin{aligned} s(y, \tilde{x}) := \frac{\partial \ell_{\theta_t}(y, \tilde{x})}{\partial t}_{|t=0} &= \operatorname{tr}\left(\left[\tilde{A}(t)\right]^{-1} \frac{\mathrm{d}\tilde{A}(t)}{\mathrm{d}t}_{|t=0}\right) + h_0(\tilde{x}) + \sum_{k=1}^{K} \left[\phi_k\left(A_k v\right) \times \tilde{D}_k(0) v\right] + h_k\left(A_k v\right) \\ &= g' \dot{\ell}_\theta + h_0(\tilde{x}) + \sum_{k=1}^{K} h_k\left(A_k v\right). \end{aligned}$$

We will demonstrate that the conditions in Lemma 7.6 of van der Vaart (1998) (alternatively Lemma 1.8 of van der Vaart (2002)) are satisfied for the map $t \mapsto P_{\theta_t}$ from $(-\delta, \delta)$ to $\mathcal{P}_\Theta$, from which we will be able to conclude that this is a differentiable path with score function as in the preceding display.[37]

Firstly, by the imposed continuous differentiability conditions we have that $t \mapsto \sqrt{p_{\theta_t}}$ is continuously differentiable λ-a.e..

It remains to show that $\int \left(\frac{\dot{p}_{\theta_t}}{p_{\theta_t}}\right)^2 \mathrm{d}P_{\theta_t}$ is finite and continuous in t. For this, note that when it exists we have $\acute{\ell}_{\theta_t} = \frac{\dot{p}_{\theta_t}}{p_{\theta_t}}$. Therefore, we can bound our integral by

$$\begin{aligned} \int \left(\acute{\ell}_{\theta_t}(y, \tilde{x})\right)^2 \mathrm{d}P_{\theta_t} \lesssim \operatorname{tr}\left(\left[\tilde{A}(t)\right]^{-1} \frac{\mathrm{d}\tilde{A}(t)}{\mathrm{d}t}\right)^2 &+ \int \left(\frac{h_0(\tilde{x})}{1 + t h_0(\tilde{x})}\right)^2 \mathrm{d}P_{\theta_t} \\ &+ \sum_{k=1}^{K} \int \left[\phi_k\left([\tilde{A}(t)]_k \tilde{v}(t)\right) \times \tilde{D}_k(t)\right]^2 \mathrm{d}P_{\theta_t} \\ &+ \sum_{k=1}^{K} \int \left(\frac{h_k([\tilde{A}(t)]_k \tilde{v}(t)) + t h_k'([\tilde{A}(t)]_k \tilde{v}(t)) \times \tilde{D}_k(t)}{1 + t h_k([\tilde{A}(t)]_k \tilde{v}(t))}\right)^2 \mathrm{d}P_{\theta_t}. \end{aligned}$$

The first rhs term can be ensured finite by choosing δ small enough since $[\tilde{A}(t)]^{-1} \frac{\mathrm{d}\tilde{A}(t)}{\mathrm{d}t}$ is continuous in t.[38] The same is true of the second term, since h_0 is bounded λ-a.s., hence G_0-a.s., and

$$\int \left(\frac{h_0(\tilde{x})}{1 + t h_0(\tilde{x})}\right)^2 \mathrm{d}P_{\theta_t} = \int \left(\frac{h_0(\tilde{x})}{1 + t h_0(\tilde{x})}\right)^2 \eta_0(\tilde{x})(1 + t h_0(\tilde{x}))\, \mathrm{d}\lambda = \int \frac{h_0(\tilde{x})^2}{1 + t h_0(\tilde{x})} \mathrm{d}G_0(\tilde{x}).$$

For the third term it suffices to consider the integral for an arbitrary $k \in [K]$, which by

[37] Strictly speaking, applying lemma 7.6 as stated in van der Vaart (1998) would require continuous differentiability for every y. Nevertheless, with appropriate modifications, the same proof demonstrates the claim remains valid with continuous differentiability holding “only” λ-a.e.. See also proposition 2.1.1 of Bickel et al. (1998).

[38] By our assumptions that $(\alpha, \beta_1) \mapsto A(\alpha, \beta_1)$ is continuously differentiable and $A(\alpha, \beta_1)$ is invertible.

Cauchy-Schwarz is bounded by

$$\int \left[\phi_k\left([\tilde{A}(t)]_k\tilde{v}(t)\right) \times \tilde{D}_k(t)\right]^2 \mathrm{d}P_{\theta_t} \leq \left\|\phi_k\left([\tilde{A}(t)]_k\tilde{v}(t)\right)^2\right\|_{P_{\theta_t},2}\left\|\left[\tilde{D}_k(t)\right]^2\right\|_{P_{\theta_t},2}$$
$$< \infty$$

For the first term observe that if $Y, \tilde{X}$ has law P_{θ_t}, then $[\tilde{A}(t)]_k\tilde{v}(t)$ is distributed according to the density $\eta_k(1 + th_k) \in \mathscr{H}$ (for small enough δ), and thus the integral is finite by the definition of $\mathscr{H}$, i.e. assumption 2.4.1-part 1. For the second term write

$$\tilde{D}_k(t) = \frac{\mathrm{d}[\tilde{A}(a)]_k}{\mathrm{d}a}(t)\left(z - \tilde{B}(t)x\right) - [\tilde{A}(t)]_k\left(\frac{\mathrm{d}\tilde{B}(a)}{\mathrm{d}a}(t)x\right),$$

and note that for small enough δ, $P_{\theta_t} \in \mathcal{P}_\Theta$ and so for some small enough $\nu > 0$, each $P_{\theta_t}|Y_k|^{4+\nu} < \infty$ and $P_{\theta_t}|X_l|^{4+\nu} < \infty$ (by assumption 2.4.1), hence $\left\|\left[\tilde{D}_k(t)\right]^2\right\|_{P_{\theta_t},2} = \sqrt{\int[\tilde{D}_k(t)]^4\,\mathrm{d}P_{\theta_t}} < \infty$ since $\int\|\tilde{D}_k(t)\|_2^{4+\nu}\,\mathrm{d}P_{\theta_t} < \infty$.

For the final term on the rhs, it is again sufficient to consider the integral for any arbitrary $k \in [K]$. Here, let $c > 0$ be a bound away from zero for $1 + th_k$ on $(-\delta, \delta)$ and let $M > 0$ bound both h_k and h_k' on the same interval, which we know to be possible by their definition. Then this integral can be bounded by

$$\int\left(\frac{h_k([\tilde{A}(t)]_k\tilde{v}(t)) + th_k'([\tilde{A}(t)]_k\tilde{v}(t)) \times \tilde{D}_k(t)}{1 + th_k([\tilde{A}(t)]_k\tilde{v}(t))}\right)^2 \mathrm{d}P_{\theta_t} \leq \int\left(\frac{M + tM\tilde{D}_k(t)}{c}\right)^2 \mathrm{d}P_{\theta_t},$$

where the right hand side can be seen to be finite by the fact that $\int[\tilde{D}_k(t)]^2\,\mathrm{d}P_{\theta_t} < \infty$ as implied by the corresponding finite 4th moment obtained above.

To show continuity, let $t_n \to t$ be an arbitrary convergent sequence in $[0, \delta)$ with δ chosen such that if $0 \leq t \leq \delta$ then each $h_k, h_k', h_0 \leq M$ and $1 + th_k, 1 + th_0 \geq c > 0$. Suppose that $Z_n = (Y_n, \tilde{X}_n)$ and $Z = (Y, \tilde{X})$ have laws $P_{\theta_{t_n}}$ and P_{θ_t} respectively and let $\tilde{v}(t, Z) := Y - \tilde{B}(t)X$. We have

$$b_n := \operatorname{tr}\left(\left[\tilde{A}(t_n)\right]^{-1}\frac{\mathrm{d}\tilde{A}(t)}{\mathrm{d}t}(t_n)\right) \to b := \operatorname{tr}\left(\left[\tilde{A}(t)\right]^{-1}\frac{\mathrm{d}\tilde{A}(t)}{\mathrm{d}t}(t)\right),$$

which converges by the continuity of all its constituent functions. Define for $k = 1, \ldots, K$

$$\begin{aligned}
U_{k,n} &\coloneqq \phi_k \left([\tilde{A}(t_n)]_k \tilde{v}(t_n, Z_n)\right) \\
W_{k,n} &\coloneqq \tilde{D}_k(t_n) \\
V_{k,n} &\coloneqq \frac{h_k([\tilde{A}(t_n)]_k \tilde{v}(t_n, Z_n))}{1 + t_n h_k([\tilde{A}(t_n)]_k \tilde{v}(t_n, Z_n))} \\
Q_{k,n} &\coloneqq \frac{t_n h'_k([\tilde{A}(t_n)]_k \tilde{v}(t_n, Z_n))}{1 + t_n h_k([\tilde{A}(t_n)]_k \tilde{v}(t_n, Z_n))} \\
E_n &\coloneqq \frac{h_0(\tilde{X}_n)}{1 + t_n h_0(\tilde{X}_n)},
\end{aligned}$$

and analogously U_k, V_k, W_k, Q_k, E where the t_n are replaced by t and the Z_n by Z respectively. Since $p_{\theta_{t_n}} \to p_{\theta_t}$ we have that $\tilde{Z}_n \rightsquigarrow \tilde{Z}$ by Scheffé's theorem. Hence, by the continuous mapping theorem

$$\begin{aligned}
&(U_{1,n}, V_{1,n}, W_{1,n}, Q_{1,n}, \ldots, U_{K,n}, V_{K,n}, W_{K,n}, Q_{K,n}, E_n) \\
&\qquad\qquad \rightsquigarrow (U_1, V_1, W_1, Q_1, \ldots, U_K, V_K, W_K, Q_K, E).
\end{aligned}$$

Moreover, $V_{k,n}$, $Q_{k,n}$ and E_n are bounded above. We have that $(U_{k,n}^4)_{n\geq 1}$ and $(W_{k,n}^4)_{n\geq 1}$ are uniformly integrable for each $k \in [K]$. For the former, note that each $[\tilde{A}(t_n)]_k \tilde{v}(t_n, Z_n)$ is distributed according to the density $\eta_k(1 + t_n h_k)$. Hence we have for small enough but positive ν

$$\sup_{n\in\mathbb{N}} \mathbb{E}|U_{k,n}|^{4+\nu} = \sup_{n\in\mathbb{N}} \int |\phi_k(z)|^{4+\nu} \eta_k(z)(1+t_n h_k(z)) \,\mathrm{d}z \lesssim \int |\phi_k(z)|^{4+\nu} \eta_k(z) \,\mathrm{d}z < \infty.$$

Similarly, using Cauchy-Schwarz, for small enough but positive ν

$$\begin{aligned}
\sup_{n\in\mathbb{N}} \mathbb{E}|W_{k,n}|^{4+\nu} &= \sup_{n\in\mathbb{N}} \int |\tilde{D}_k(t_n)|^{4+\nu} \,\mathrm{d}P_{\theta_{t_n}} \\
&\lesssim \sup_{n\in\mathbb{N}} \int \|\epsilon_n\|_2^{4+\nu} \,\mathrm{d}P_{\theta_{t_n}} + \sup_{n\in\mathbb{N}} \int \|X_n\|_2^{4+\nu} \,\mathrm{d}P_{\theta_{t_n}} \\
&\lesssim \sup_{n\in\mathbb{N}} \sum_{k=1}^{K} \int |e_k|^{4+\nu} \eta_k(e_k)(1 + t_n h_k(e_k)) \,\mathrm{d}e_k \\
&\qquad + \sup_{n\in\mathbb{N}} \int \|(1, \tilde{x}')'\|_2^{4+v} \eta_0(\tilde{x})(1 + t_n(h_0(\tilde{x}))) \,\mathrm{d}\tilde{x} \\
&\lesssim \sum_{k=1}^{K} \int |\epsilon_k|^{4+\nu} \,\mathrm{d}G_k + \int \|(1, \tilde{X}')'\|_2^{4+v} \,\mathrm{d}G_0 \\
&< \infty.
\end{aligned}$$

With this in hand, using continuous mapping theorem and noting that each of the relevant sequences is $P_{\theta_{t_n}}$-UI given the preceding discussion we have, as $n \to \infty$

$$P_{\theta_{t_n}}\left[b_n + E_n + \sum_{k=1}^{K} U_{k,n}W_{k,n} + \sum_{k=1}^{K} V_{k,n} + Q_{k,n}\right]^2 \to P_{\theta_t}\left[b + E + \sum_{k=1}^{K} U_k W_k + \sum_{k=1}^{K} V_k + Q_k\right]^2,$$

yielding the required continuity.

□

Proof of Lemma B.2. To construct the efficient score function for γ, we need to project the elements of $\dot{\ell}_\theta(y)$, as given in (44), onto the orthogonal complement of $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ (equation (50)), that is: $\breve{\ell}_{\theta,l} = \Pi\left(\dot{\ell}_{\theta,l}\middle| \left[\mathcal{T}^{\eta|\gamma}_{P_\theta,H}\right]^\perp\right)$.[39] The efficient score then follows as $\tilde{\ell}_{\theta,l} := \dot{\ell}_{\theta,l} - \Pi_\theta \dot{\ell}_{\theta,l} = \dot{\ell}_{\theta,l} - \Pi\left(\dot{\ell}_{\theta,l}\middle| \left[\mathcal{T}^{\eta|\gamma}_{P_\theta,H}\right]^\perp\right)$.

We first provide some results that simplify the exposition. Lemma C.2 proves that the closure of H_k is given by

$$\operatorname{cl} H_k = \{h_k \in L_2(G_k) : \mathbb{E}h_k(\epsilon_k) = 0, \mathbb{E}\epsilon_k h_k = 0, \mathbb{E}\kappa(\epsilon_k)h_k(\epsilon_k) = 0\},$$

and similarly

$$\operatorname{cl} H_0 = \{h_0 \in L_2(G_0) : \mathbb{E}h_0(\tilde{X}) = 0\}.$$

Now, let $\tilde{H}^\gamma_k := \{y \mapsto h_k(A_{k\bullet}v) : h_k \in H_k\}$ for $k = 1, \ldots, K$, $\tilde{H}^\gamma_0 := \{y \mapsto h_0(\tilde{x}) : h_0 \in H_0\}$ and note that $\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ can be written as

$$\mathcal{T}^{\eta|\gamma}_{P_\theta,H} = \tilde{H}^\gamma_0 + \tilde{H}^\gamma_1 + \cdots + \tilde{H}^\gamma_K. \tag{51}$$

It follows that for $k = 1, \ldots, K$

$$\operatorname{cl} \tilde{H}^\gamma_0 = \{y \mapsto h_0(\tilde{x}) : h_0 \in \operatorname{cl} H_0\}, \quad \operatorname{cl} \tilde{H}^\gamma_k = \{y \mapsto h_k(A_{k\bullet}v) : h_k \in \operatorname{cl} H_k\}, \tag{52}$$

which are (closed) subspaces of $L_2(P_\theta)$.[40]

[39] See e.g. Section 2.2 of van der Vaart (2002).

[40] To see this let $y \mapsto h_k(A_k v) \in \{y \mapsto h_k(A_k v) : h_k \in \operatorname{cl} H_k\}$. There are $h_{n,k} \in H_k$ such that $h_{n,k} \to h_k$ in $L_2(G_k)$. Hence, recalling that $A_k v$ is distributed according to η_k under P_θ, it follows immediately that $\int [h_{n,k}(A_k v) - h_k(A_k v)]^2 \, \mathrm{d}P_\theta \to 0$ as $n \to \infty$. Hence $y \mapsto h_k(A_k v) \in \operatorname{cl} \tilde{H}^\gamma_k$. For the reverse inclusion, let $y \mapsto h_k(A_k v) \in \operatorname{cl} \tilde{H}^\gamma_k$. So there are $y \mapsto h_{n,k}(A_k v)$ in $\tilde{H}^\gamma_k$ such that $\int [h_{n,k}(A_k v) - h_k(A_k v)]^2 \, \mathrm{d}P_\theta \to 0$ as $n \to \infty$. Again noting that $A_k v$ is distributed according to η_k under P_θ, this immediately implies that $h_{n,k} \to h_k$ in $L_2(G_k)$. That $\operatorname{cl} \tilde{H}^\gamma_k$ is a subspace of $L_2(P_\theta)$ follows directly from the fact that $\operatorname{cl} H_k$ is a subspace of $L_2(G_k)$ once more noting $A_k v$ is distributed according to η_k under P_θ. The argument for $\tilde{H}^\gamma_0$ is analogous.

Define $\mathscr{T} := \operatorname{cl}\tilde{H}_1^\gamma + \cdots + \operatorname{cl}\tilde{H}_K^\gamma$ and the following finite dimensional subset of $L_2(P_\theta)$

$$\mathscr{L}_0 := \mathscr{L}_1 \cup \mathscr{L}_2 := \{y \mapsto A_{k\bullet}v, y \mapsto \kappa(A_{k\bullet}v) : k \in [K]\} \cup \{y \mapsto \phi_k(A_{k\bullet}v)A_{j\bullet}v : j,k \in [K],\ j \neq k\}, \tag{53}$$

where $\kappa(w) := w^2 - 1$ and $\mathscr{L} := \operatorname{lin}\mathscr{L}_0$. Lemma C.4 proves that $\mathscr{L} \subset \mathscr{T}^\perp$.

Since orthogonal projections are linear we have that for $\kappa = \alpha, \sigma$

$$\begin{aligned}
\Pi\left(\dot{\ell}_{\theta,\kappa_l}|\mathscr{T}^\perp\right) &= \sum_{k=1}^K \sum_{j=1, j\neq k}^K \zeta_{l,k,j}^\kappa \Pi\left(\phi_k(A_{k\bullet}v)A_{j\bullet}v|\mathscr{T}^\perp\right) \\
&\quad + \sum_{k=1}^K \zeta_{l,k,k}^\kappa \Pi\left(\phi_k(A_{k\bullet}v)A_{k\bullet}v + 1|\mathscr{T}^\perp\right) \\
&= \sum_{k=1}^K \sum_{j=1, j\neq k}^K \zeta_{l,k,j}^\kappa \phi_k(A_{k\bullet}v)A_{j\bullet}v + \sum_{k=1}^K \zeta_{l,k,k}^\kappa \Pi\left(\phi_k(A_{k\bullet}v)A_{k\bullet}v + 1|\mathscr{T}^\perp\right)
\end{aligned}$$

where the second equality follows from $y \mapsto \phi_k(A_{k\bullet}v)A_{j\bullet}v \in \mathscr{L} \subset \mathscr{T}^\perp$, for $j \neq k$.

What remains is $\Pi\left(\phi_k(A_{k\bullet}v)A_{k\bullet}v + 1|\mathscr{T}^\perp\right)$. For this we specialise to the case for $\theta = (\gamma, \eta)$ such that $\eta \in \mathcal{H}_0$, for which we can establish an explicit expression.

In particular, we will show that for each $k \in [K]$, there are τ_i for $i = 1, 2$ such that $y \mapsto w(A_k v) \in \operatorname{cl}\tilde{H}_k^\gamma$ where $w(A_k v) := \phi_k(A_k v)A_k v + 1 - r(A_k v)$ and $r(A_k v) := \tau_1 A_k v + \tau_2 \kappa(A_k v)$. This would imply that we can write $\phi_k(A_k v)A_k v + 1 = w(A_k v) + r(A_k v)$ where the first summand on the right hand side is in $\mathscr{T}$ and the latter is in $\mathscr{L} \subset \mathscr{T}^\perp$.[41] Since orthogonal decompositions are unique this would further imply that $\Pi\left(\phi_k(A_k v)A_k v + 1|\mathscr{T}^\perp\right) = \Pi\left(\phi_k(A_k v)A_k v + 1|\mathscr{L}\right) = r(A_k v)$.[42]

To show that $y \mapsto w(A_k v) \in \operatorname{cl}\tilde{H}_k^\gamma$ let $h_k(z) := \phi_k(z)z + 1 - \tau_{k,1}z - \tau_{k,2}\kappa(z)$. We first note that $h_k \in L_2(G_k)$, which can be easily seen by the triangle inequality along with the fact that all of ϵ_k, $\kappa(\epsilon_k)$, 1 and $\phi_k(\epsilon_k)\epsilon_k$ are in $L_2(G_k)$. Next, $\int \phi_k(z)z \,\mathrm{d}G_k + 1 - \tau_{k,1}\int z \,\mathrm{d}G_k - \tau_{k,2}\int \kappa(z)\,\mathrm{d}G_k = 1 + \int \phi_k(z)z\,\mathrm{d}G_k$, and so as $\eta \in \mathcal{H}_0$,

$$\int h_k(z)\,\mathrm{d}G_k = 1 - 1 = 0.$$

Next, we will demonstrate that $\tau_{k,1}$ and $\tau_{k,2}$ can be chosen such that $\int h_k(z)z\,\mathrm{d}G_k =$

[41] Take $h_k = w$ and $h_j = 0$ for all $j \neq k$ to see that $y \mapsto w(A_k v) \in \operatorname{cl}\tilde{H}_k^\gamma$ implies $y \mapsto w(A_k v) \in \mathscr{T}$.

[42] See e.g. Theorem 4.11 in Rudin (1987).

$\int h_k(z)\kappa(z)\,\mathrm{d}G_k = 0$. As $\eta \in \mathcal{H}_0$ we have that

$$\begin{aligned}\int h_k(z)z\,\mathrm{d}G_k &= \int \phi_k(z)z^2\,\mathrm{d}G_k + \int z\,\mathrm{d}G_k - \tau_{k,1}\int z^2\,\mathrm{d}G_k - \tau_{k,2}\int \kappa(z)z\,\mathrm{d}G_k \\ &= -\tau_{k,1}\int z^2\,\mathrm{d}G_k - \tau_{k,2}\int z^3\,\mathrm{d}G_k + \tau_{k,2}\int z\,\mathrm{d}G_k \\ &= -\tau_{k,1}\mathbb{E}\epsilon_k^2 - \tau_{k,2}\mathbb{E}\epsilon_k^3 \\ &= -\tau_{k,1}1 - \tau_{k,2}\mathbb{E}\epsilon_k^3,\end{aligned}$$

where we note that $\mathbb{E}\epsilon_k^2 = 1$. Similarly,

$$\begin{aligned}\int h_k(z)\kappa(z)\,\mathrm{d}G_k &= \int \phi_k(z)(z^3 - z)\,\mathrm{d}G_k + \int \kappa(z)\,\mathrm{d}G_k - \tau_{k,1}\int z(z^2-1)\,\mathrm{d}G_k - \tau_{k,2}\int (z^2-1)^2\,\mathrm{d}G_k \\ &= -2 - \tau_{k,1}\left[\int z^3\,\mathrm{d}G_k - \int z\,\mathrm{d}G_k\right] - \tau_{k,2}\left[\int z^4\,\mathrm{d}G_k - 2\int z^2\,\mathrm{d}G_k + 1\right] \\ &= -2 - \tau_{k,1}\int z^3\,\mathrm{d}G_k - \tau_{k,2}\left[\int z^4\,\mathrm{d}G_k - 2\int z^2\,\mathrm{d}G_k + 1\right] \\ &= -2 - \tau_{k,1}\mathbb{E}\epsilon_k^3 - \tau_{k,2}[\mathbb{E}\epsilon_k^4 - 1].\end{aligned}$$

Hence we need to choose $\tau_{k,1}$ and $\tau_{k,2}$ such that:

$$\begin{bmatrix} 1 & \mathbb{E}\epsilon_k^3 \\ \mathbb{E}\epsilon_k^3 & \mathbb{E}\epsilon_k^4 - 1\end{bmatrix}\begin{bmatrix}\tau_{k,1} \\ \tau_{k,2}\end{bmatrix} = \begin{bmatrix} 0 \\ -2\end{bmatrix}.$$

The matrix $M_k := \left[\begin{smallmatrix} 1 & \mathbb{E}\epsilon_k^3 \\ \mathbb{E}\epsilon_k^3 & \mathbb{E}\epsilon_k^4 - 1\end{smallmatrix}\right] = \left[\begin{smallmatrix} \mathbb{E}\epsilon_k^2 & \mathbb{E}\epsilon_k^3 \\ \mathbb{E}\epsilon_k^3 & \mathbb{E}\epsilon_k^4 - 1\end{smallmatrix}\right]$ is nonsingular by assumption 2.4.1; see footnote 5. Hence we can take $(\tau_{k,1}, \tau_{k,2})' = M_k^{-1}(0,-2)'$, which is non zero by the nonsingularity of M_k^{-1}. We conclude that

$$\begin{aligned}\Pi\left(\dot{\ell}_{\theta,\alpha_l}|\mathscr{T}^{\perp}\right) &= \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\zeta_{l,k,j}^{\alpha}\phi_k(A_{k\bullet}v)A_{j\bullet}v + \sum_{k=1}^{K}\zeta_{l,k,k}^{\alpha}\left[\tau_{k,1}A_{k\bullet}v + \tau_{k,2}\kappa(A_{k\bullet}v)\right]\ , \\ \Pi\left(\dot{\ell}_{\theta,\sigma_l}|\mathscr{T}^{\perp}\right) &= \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\zeta_{l,k,j}^{\sigma}\phi_k(A_{k\bullet}v)A_{j\bullet}v + \sum_{k=1}^{K}\zeta_{l,k,k}^{\sigma}\left[\tau_{k,1}A_{k\bullet}v + \tau_{k,2}\kappa(A_{k\bullet}v)\right]\ ,\end{aligned}$$

Moreover, by independence, for any $h_0 \in \operatorname{cl}\tilde{H}_0^{\gamma}$

$$\begin{aligned}P_\theta\left[\Pi\left(\dot{\ell}_{\theta,\alpha_l}|\mathscr{T}^{\perp}\right)h_0\right] &= P_\theta\left[\Pi\left(\dot{\ell}_{\theta,\alpha_l}|\mathscr{T}^{\perp}\right)\right]P_\theta h_0 = 0, \\ P_\theta\left[\Pi\left(\dot{\ell}_{\theta,\sigma_l}|\mathscr{T}^{\perp}\right)h_0\right] &= P_\theta\left[\Pi\left(\dot{\ell}_{\theta,\sigma_l}|\mathscr{T}^{\perp}\right)\right]P_\theta h_0 = 0,\end{aligned}$$

and so by lemma C.3 we can conclude that (see e.g. Bickel et al., 1998, Proposition A.2.3.B)

$$\Pi\left(\dot{\ell}_{\theta,\alpha_l}|\mathscr{T}^{\perp}\right) = \Pi\left(\dot{\ell}_{\theta,\alpha_l}|\left[\mathcal{T}_{P_\theta,H}^{\eta|\gamma}\right]^{\perp}\right) \qquad \text{and} \qquad \Pi\left(\dot{\ell}_{\theta,\sigma_l}|\mathscr{T}^{\perp}\right) = \Pi\left(\dot{\ell}_{\theta,\sigma_l}|\left[\mathcal{T}_{P_\theta,H}^{\eta|\gamma}\right]^{\perp}\right).$$

For the remaining part corresponding to b, let $\varsigma_k := M_k^{-1}(1,0)'$ and define $q(y,\tilde{x}) := \phi_k(A_{k\bullet}v) + \varsigma_{k,1}A_{k\bullet}v + \varsigma_{k,2}\kappa(A_{k\bullet}v)$. Then we have that for any $a \in \mathbb{R}^d$

$$y \mapsto q(y,\tilde{X}) \times a'\mathbb{E}X \in \operatorname{cl}\tilde{H}_k^{\gamma} \subset \operatorname{cl}\mathcal{T}_{P_\theta,H}^{\eta|\gamma},$$

since letting $\tilde{a} := a'\mathbb{E}X$ we have $P_\theta(\tilde{a}q(Y,\tilde{X}))^2 < \infty$ by the triangle inequality & $P_\theta \tilde{a}q(Y,\tilde{X}) = 0$,

$$\begin{aligned} P_\theta \tilde{a}q(Y,\tilde{X})A_{k\bullet v} &= \tilde{a}\left[\int \phi_k(\epsilon_k)\epsilon_k \, \mathrm{d}G_k + \varsigma_{k,1}\int \epsilon_k^2 \, \mathrm{d}G_k + \varsigma_{k,2}\int \epsilon_k^3 - \epsilon_k \, \mathrm{d}G_k\right] \\ &= \tilde{a}\left[-1 + \varsigma_{k,1} + \varsigma_{k,2}\mathbb{E}\epsilon_k^3\right] \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} P_\theta \tilde{a}q(Y,\tilde{X})\kappa(A_{k\bullet v}) &= \tilde{a}\left[\int \phi_k(\epsilon_k)(\epsilon_k^2 - 1) \, \mathrm{d}G_k + \varsigma_{k,1}\int \epsilon_k^3 - \epsilon_k \, \mathrm{d}G_k + \varsigma_{k,2}\int \epsilon_k^4 - 2\epsilon_k^2 + 1 \, \mathrm{d}G_k\right] \\ &= \tilde{a}\left[\varsigma_{k,1}\mathbb{E}\epsilon_k^3 + \varsigma_{k,2}(\mathbb{E}\epsilon_k^4 - 1)\right] \\ &= 0 \end{aligned}$$

by the choice of ς_k. Moreover, since for any $h \in \mathcal{T}_{P_\theta,H}^{\eta|\gamma}$ we have

$$\begin{aligned} &P_\theta\left(\left[a'X\phi_k(A_{k\bullet}v) - a'\mathbb{E}X\left(\phi_k(A_{k\bullet}v) + \varsigma_{k,1}A_{k\bullet}v + \varsigma_{k,2}\kappa(A_{k\bullet}v)\right)\right]h(Y,\tilde{X})\right) \\ &\quad = P_\theta\left(\left[a'(X - \mathbb{E}X)\phi_k(A_{k\bullet}v) - a'\mathbb{E}X\left(\varsigma_{k,1}A_{k\bullet}v + \varsigma_{k,2}\kappa(A_{k\bullet}v)\right)\right]\left[h_0(\tilde{X}) + \sum_{j=1}^{K} h_j(A_{j\bullet}v)\right]\right) \\ &\quad = 0, \end{aligned}$$

it follows that

$$\Pi\left(\dot{\ell}_{\theta,b,l} \middle| \left[\mathcal{T}_{P_\theta,H}^{\eta|\gamma}\right]^{\perp}\right) = \sum_{k=1}^{K}[-A_{k\bullet}D_{b,l}]\left[(x - \mathbb{E}x)\phi_k(A_{k\bullet}v) - \mathbb{E}x\left(\varsigma_{k,1}A_{k\bullet}v + \varsigma_{k,2}\kappa(A_{k\bullet}v)\right)\right].$$

□

Proof of Lemma B.3. We start by showing that $\hat{\phi}_k$ satisfies equation (29). Under P_{θ_n}, we have that $A_{n,k\bullet}(Y_i - B_nX_i) \simeq \epsilon_{i,k} \sim \eta_k$, where $A_{n,k\bullet}$ denotes the kth row of

$A_n \equiv A(\alpha_0, \sigma_n)$. Additionally, we can write

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\phi}_k(\epsilon_{i,k})W_{i,n} - \phi_k(\epsilon_{i,k})W_{i,n}\right| \leq & \left|\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_k(\epsilon_{i,k}) - \tilde{\phi}_k(\epsilon_{i,k})\right]W_{i,n}\right| \\ & + \left|\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_k(\epsilon_{i,k}) - \phi_{k,n}(\epsilon_{i,k})\right]W_{i,n}\right| \\ & + \left|\frac{1}{n}\sum_{i=1}^{n}\left[\phi_{k,n}(\epsilon_{i,k}) - \phi_k(\epsilon_{i,k})\right]W_{i,n}\right| ,\end{aligned} \tag{54}$$

where $\hat{\phi}_k(z) = \hat{\gamma}_k' b_k(z)$ as defined in equation (2.7), $\tilde{\phi}_k(z) := \gamma_k' b_k(z)$, where

$$\gamma_k = -G_k[b_k b_k']^{-1} G_k c_k ,$$

with G_k being the law corresponding to η_k. Finally, $\phi_{k,n} := \phi_k \mathbf{1}_{[\Xi^L_{k,n}, \Xi^U_{k,n}]}$ as in Assumption (2.4.2) and ϕ_k is the true log density score. We will show that each of these three terms on the right hand side are $o_G(n^{-1/2})$, where G is the product of G_k and G_w, which implies that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\phi}_{k,n}(A_{n,k}Y_i)W_{i,n} - \phi_k(A_{n,k}Y_i)W_{i,n}\right| \xrightarrow{P_{\theta_n}} 0.$$

For the last term in (54), by assumption $G_k\{\epsilon_{i,k} \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]\} \downarrow 0$ and hence by independence and Cauchy-Schwarz

$$\begin{aligned} G\left([\phi_{k,n}(\epsilon_{i,k}) - \phi_k(\epsilon_{i,k})]^2 W_{i,n}^2\right) &= G_k\left[\phi_k(\epsilon_{i,k})^2 \mathbf{1}\{\epsilon_{i,k} \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]\}\right] G_w W_{i,n}^2 \\ &\leq \left[G_k \phi_k(\epsilon_{i,k})^4\right]^{1/2}\left[G_k \mathbf{1}\{\epsilon_{i,k} \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]\}\right]^{1/2} G_w W_{i,n}^2 \\ &\to 0. \end{aligned} \tag{55}$$

By Markov's inequality it follows that for any $\upsilon > 0$,

$$G\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\phi_{k,n}(\epsilon_{i,k}) - \phi_k(\epsilon_{i,k})]W_{i,n}\right| > \upsilon\right) \leq \frac{nG\left([\phi_{k,n}(\epsilon_{i,k}) - \phi_k(\epsilon_{i,k})]^2 W_{i,n}^2\right)}{n\upsilon} \to 0.$$

For the second term, we note that by our hypotheses and lemma C.6 we have

$$\begin{array}{rcl} G\left([\tilde{\phi}_k(\epsilon_{i,k}) - \phi_{k,n}(\epsilon_{i,k})]^2 W_{i,n}^2\right) & = & G_k\left([\tilde{\phi}_k(\epsilon_{i,k}) - \phi_{k,n}(\epsilon_{i,k})]^2\right) G_w W_{i,n}^2 \\ & \leq & C^2 \delta_{k,n}^6 \|\phi_k^{(3)}\|_\infty^2 G_w W_{i,n}^2 \to 0 \end{array} , \tag{56}$$

as $n \to \infty$, and hence again by Markov's inequality for any $\upsilon > 0$,

$$G\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\tilde{\phi}_k(\epsilon_{i,k}) - \phi_{k,n}(\epsilon_{i,k})]W_{i,n}\right| > \upsilon\right) \leq \frac{nG\left([\tilde{\phi}_k(\epsilon_{i,k}) - \phi_{k,n}(\epsilon_{i,k})]^2 W_{i,n}^2\right)}{n\upsilon} \to 0.$$

For the first term, by Cauchy-Schwarz

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_k(\epsilon_{i,k})-\tilde{\phi}_k(\epsilon_{i,k})\right]W_{i,n}\right| \leq \|\hat{\gamma}_k-\gamma_k\|_2\left\|\frac{1}{n}\sum_{i=1}^{n}b_k(\epsilon_{i,k})W_{i,n}\right\|_2 = o_G(n^{-1/2}),$$

by lemmas C.7 and C.8.

Next, we show that $\hat{\phi}_k$ satisfies equation (30). We write:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\left[\hat{\phi}_k(\epsilon_{i,k})-\phi_k(\epsilon_{i,k})\right]W_{i,n}\right)^2 \leq \frac{4}{n}\sum_{i=1}^{n}&\left[\hat{\phi}_k(\epsilon_{i,k})-\tilde{\phi}_k(\epsilon_{i,k})\right]^2W_{i,n}^2\\ &+\frac{4}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_k(\epsilon_{i,k})-\phi_{k,n}(\epsilon_{i,k})\right]^2W_{i,n}^2 \qquad (57)\\ &+\frac{4}{n}\sum_{i=1}^{n}\left[\phi_{k,n}(\epsilon_{i,k})-\phi_k(\epsilon_{i,k})\right]^2W_{i,n}^2.\end{aligned}$$

We will show that (1/4 of) each of the right hand side terms is $o_G(\nu_n)$ under our assumptions, which is sufficient for equation (30) since $A_{k,n}(Y_i - B_nX_i) \simeq \epsilon_{i,k} \sim \eta_k$ under P_{θ_n}. For the last term, for any $\upsilon > 0$, by Markov's inequality, independence and Cauchy-Schwarz we have

$$G\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\phi_{k,n}(\epsilon_{i,k})-\phi_k(\epsilon_{i,k})\right]^2W_{i,n}^2\right| > \upsilon\nu_n\right) \lesssim \frac{G_k\mathbf{1}\{\epsilon_{i,k}\notin[\Xi_{k,n}^L,\Xi_{k,n}^U]\}G_wW_{i,n}^2}{\upsilon\nu_n} = o(1).$$

For the second term, for any $\upsilon > 0$, by Markov's inequality, independence and lemma C.6:

$$\begin{aligned}G\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\phi}_k(\epsilon_{i,k})-\phi_{k,n}(\epsilon_{i,k})\right]^2W_{i,n}^2\right| > \upsilon\nu_n\right) &\leq \frac{G_k\left([\tilde{\phi}_k(\epsilon_{i,k})-\phi_{k,n}(\epsilon_{i,k})]^2\right)G_wW_{i,n}^2}{\upsilon\nu_n}\\ &\leq \frac{C\delta_{k,n}^6\|\phi_k^{(3)}\|_\infty^2G_wW_{i,n}^2}{\upsilon\nu_n}\\ &= o(1).\end{aligned}$$

Finally, for the first term in the decomposition, by lemma C.8 and Assumption 2.4.2-part (ii) we have

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\phi}_k(\epsilon_{i,k})-\tilde{\phi}_k(\epsilon_{i,k})\right]^2W_{i,n}^2 \leq \|\hat{\gamma}_k-\gamma_k\|_2^2\left[\frac{1}{n}\sum_{i=1}^{n}\|b_k(\epsilon_{i,k})\|_2^2W_{i,n}^2\right] = o_G(\nu_n).$$

□

C.1. Supporting results

Definition C.1. *Let $\mathcal{C}^k$ denote the space of real functions which have a continuous derivative of order k. Let $\mathcal{C}^\infty := \bigcap_{k\geq 1} \mathcal{C}^k$. Let $\mathcal{C}_c^\infty$ be the subset of $\mathcal{C}^\infty$ consisting of functions $f \in \mathcal{C}^\infty$ such that* $\operatorname{supp}(f)$ *is compact.*[43]

Lemma C.1. *Let μ be a probability measure on $\mathbb{R}$. Then, $\mathcal{C}_c^\infty$ is dense in $L_2(\mu)$.*

Proof. Let $\mathcal{C}_c$ denote the set of compactly supported real functions on $\mathbb{R}$. By theorem 1.1 of Billingsley (1999) and proposition 7.9 of Folland (1999), we have that $\mathcal{C}_c$ is dense in $L_2(\mu)$ and hence it suffices to show that $\mathcal{C}_c^\infty$ is dense in $\mathcal{C}_c$ with respect to the $L_2(\mu)$ norm.[44] Now, let $g \in \mathcal{C}_c$ and choose $R > 0$ such that $\operatorname{supp}(g) \subset (-R, R) \subset \mathbb{R}$. By the $\mathcal{C}^\infty$ Urysohn lemma (8.18 in Folland, 1999), there is a $h \in \mathcal{C}_c^\infty$ such that $h \in [0, 1]$, $h = 1$ on $\operatorname{supp}(g)$ and $\operatorname{supp}(h) \subset (-R, R)$. By the Weierstrass approximation theorem (see e.g. p. 247 of Royden and Fitzpatrick, 2010) there is a sequence of polynomials $(p_n)_{n\geq 1}$ such that $p_n \to g$ uniformly in $[-R, R]$. Note that the product $p_n h \in \mathcal{C}_c^\infty$. We have that $p_n h \to gh = g$ uniformly on $\operatorname{supp}(h)$. It follows that $\|p_n h - g\|_{\mu,2} \to 0$.[45] □

Lemma C.2. *Let H_k be defined as in* (46). *We have that*

$$\operatorname{cl} H_k = \{h_k \in L_2(G_k) : \mathbb{E}h_k(\epsilon_k) = 0, \mathbb{E}\epsilon_k h_k = 0, \mathbb{E}\kappa(\epsilon_k)h_k(\epsilon_k) = 0\},$$

where G_k is the law on $\mathbb{R}$ corresponding to η_k and ϵ_k is distributed according to G_k.

Let H_0 be defined as in (47). *We have that*

$$\operatorname{cl} H_0 = \{h_0 \in L_2(G_0) : \mathbb{E}h_0(\tilde{X}) = 0\},$$

where G_0 is the law on $\mathbb{R}^{d-1}$ corresponding to η_0 and $\tilde{X}$ is distributed according to G_0.

Proof. Let $h_k \in \operatorname{cl} H_k$. Then, there are $h_{n,k} \in H_k \subset L_2(G_k)$ with $\|h_{n,k} - h_k\|_{G_k,2} \to 0$. Hence, $h_k \in L_2(G_k)$. Since the inner product is continuous we have

$$\mathbb{E}h_k(\epsilon_k)\xi(\epsilon_k) = \langle h_k(\epsilon_k), \xi(\epsilon_k)\rangle_{G_k} = \lim_{n\to\infty} \langle h_{n,k}(\epsilon_k), \xi(\epsilon_k)\rangle_{G_k} = \lim_{n\to\infty} 0 = 0,$$

for each $\xi \in \{\xi_0, \xi_1, \kappa\}$, where $\xi_0(x) := 1$, $\xi_1(x) := x$. Hence, $h_k \in \{h_k \in L_2(G_k) : \mathbb{E}h_k(\epsilon_k) = 0, \mathbb{E}\epsilon_k h_k = 0, \mathbb{E}\kappa(\epsilon_k)h_k(\epsilon_k) = 0\}$ and thus we have that

[43] The *support* of f is $\operatorname{supp}(f) := \operatorname{cl}\{x : f(x) \neq 0\}$.

[44] Suppose we have shown this. Then since for each $g \in \mathcal{C}_c$ we have $g \in \operatorname{cl}\mathcal{C}_c^\infty$ and hence $\mathcal{C}_c \subset \operatorname{cl}\mathcal{C}_c^\infty$. Noting that $\mathcal{C}_c$ is dense in $L_2(\mu)$ we obtain the chain of inclusions $L_2(\mu) \subset \operatorname{cl}\mathcal{C}_c \subset \operatorname{cl}\operatorname{cl}\mathcal{C}_c^\infty = \operatorname{cl}\mathcal{C}_c^\infty \subset L_2(\mu)$ where the last inclusion is evident from the fact that any function in $\mathcal{C}_c^\infty$ is bounded and hence in $L_2(\mu)$, which is itself closed.

[45] Fix $\epsilon > 0$. Take N large enough that for all $n \geq N$ we have $|p_n h - g| < \epsilon$ on $\operatorname{supp}(h)$. Then, $\int (p_n h - g)^2 \,\mathrm{d}\mu = \int_{\operatorname{supp}(h)} (p_n h - g)^2 \,\mathrm{d}\mu + \int_{\mathbb{R}\setminus\operatorname{supp}(h)} (p_n h - g)^2 = \int_{\operatorname{supp}(h)} (p_n h - g)^2 \,\mathrm{d}\mu < \epsilon^2$ since $p_n h - g = 0$ outside of $\operatorname{supp}(h)$.

$$\operatorname{cl} H_k \subset \{h_k \in L_2(G_k) : \mathbb{E}h_k(\epsilon_k) = 0, \mathbb{E}\epsilon_k h_k = 0, \mathbb{E}\kappa(\epsilon_k)h_k(\epsilon_k) = 0\}.$$

For the other inclusion, h_k be in $L_2(G_k)$ and orthogonal to each $\xi \in \{\xi_0, \xi_1, \kappa\}$. We want to approximate (in the $L_2(G_k)$ norm) h_k by functions in H_k. First ignore the orthogonality constraints: the space $\mathcal{C}_c^\infty$ (see definition C.1) (of which $\mathcal{C}_b^1(\lambda) \subset L_2(G_k)$ is a superset) is dense in $L_2(G_k)$ by lemma C.1. Hence there is a sequence $(h_{n,k})_{n\geq 1}$ in $\mathcal{C}_b^1(\lambda)$ such that $\|h_{n,k} - h_k\|_{G_k,2} \to 0$. Introduce the function

$$\tilde{h}_{n,k}(z) := h_{n,k}(z) + \upsilon_n + \nu_n v(z) + \omega_n w(z),$$

where each of υ_n, ν_n and ω_n are in $\mathbb{R}$ and $v, w \in \mathcal{C}_b^1(\lambda)$ are such that

$$\mathbb{E}v(\epsilon_k) = \mathbb{E}w(\epsilon_k) = 0, \quad \mathbb{E}\epsilon_k w(\epsilon_k) = \mathbb{E}\kappa(\epsilon_k)v(\epsilon_k) = 0, \quad \mathbb{E}\epsilon_k v(\epsilon_k) = \mathbb{E}\kappa(\epsilon_k)w(\epsilon_k) = 1,$$

and the existence of such functions is guaranteed by lemma C.5. It is clear from its definition that $\tilde{h}_{n,k} \in \mathcal{C}_b^1(\lambda)$. Now, put

$$\upsilon_n := -\mathbb{E}h_{n,k}(\epsilon_k), \quad \nu_n := -\mathbb{E}[h_{n,k}(\epsilon_k)\epsilon_k], \quad \omega_n := -\mathbb{E}[h_{n,k}\kappa(\epsilon_k)].$$

Then, we clearly have that

$$\langle \tilde{h}_{n,k}, \xi_0 \rangle_{G_k} = \mathbb{E}\left[h_{n,k}(\epsilon_k) + \upsilon_n\right] = \mathbb{E}h_{n,k}(\epsilon_k) - \mathbb{E}h_{n,k}(\epsilon_k) = 0,$$

$$\langle \tilde{h}_{n,k}, \xi_1 \rangle_{G_k} = \mathbb{E}\left[h_{n,k}(\epsilon_k)\epsilon_k + \nu_n \mathbb{E}[v(\epsilon_k)\epsilon_k]\right] = \mathbb{E}\left[h_{n,k}(\epsilon_k)\epsilon_k\right] - \mathbb{E}[h_{n,k}(\epsilon_k)\epsilon_k] = 0$$

$$\langle \tilde{h}_{n,k}, \kappa \rangle_{G_k} = \mathbb{E}\left[h_{n,k}\kappa(\epsilon_k) + \omega_n \mathbb{E}[w(\epsilon_k)\kappa(\epsilon_k)]\right] = \mathbb{E}[h_{n,k}\kappa(\epsilon_k)] - \mathbb{E}[h_{n,k}\kappa(\epsilon_k)] = 0.$$

Moreover, since $h_{n,k} \xrightarrow{L_2(G_k)} h_k$ we have that $(\upsilon_n, \nu_n, \omega_n) \to 0$ as $n \to \infty$. Therefore,

$$\begin{aligned}
\|\tilde{h}_{n,k} - h_k\|_{G_k,2} &\leq \|h_{n,k} - h_k\|_{G_k,2} + \|\upsilon_n + \nu_n v + \omega_n w\|_{G_k,2} \\
&\leq \|h_{n,k} - h_k\|_{G_k,2} + |\upsilon_n| + |\nu_n|\|v\|_{G_k,2} + |\omega_n|\|w\|_{G_k,2} \\
&\to 0,
\end{aligned}$$

as $n \to \infty$ where we note that $\|v\|_{G_k,2} < \infty$ and $\|w\|_{G_k,2} < \infty$ since the functions are bounded λ-a.e. (and hence G_k-a.s.). Thus $(\tilde{h}_{n,k})_{n\geq 1}$ is a sequence in H_k such that $\|\tilde{h}_{n,k} - h_k\|_{G_k,2} \to 0$ and we conclude that $\{h_k \in L_2(G_k) : \mathbb{E}h_k(\epsilon_k) = 0, \mathbb{E}\epsilon_k h_k = 0, \mathbb{E}\kappa(\epsilon_k)h_k(\epsilon_k) = 0\} \subset \operatorname{cl} H_k$.

For H_0 let $h_0 \in \operatorname{cl} H_0$. There are $(h_{n,0} \in H_0 \subset L_2(G_0)$ with $\|h_{n,0} - h_0\|_{G_0,2} \to 0$. Hence $h_0 \in L_2(G_0)$ and $\int h_0 \,\mathrm{d}G_0 = \lim_{n\to\infty} \int h_{n,0} \,\mathrm{d}G_0 = 0$. Conversely, suppose that $h_0 \in L_2^0(G_0)$. Since $C_b(\lambda, \mathbb{R}^{d-1}) \subset L_2(G_0)$ is a superset of the compactly supported continuous functions on $\mathbb{R}^{d-1}$ (when considered as elements of $L_2(G_0)$) it is dense in $L_2(G_0)$ by e.g. Theorem 3.14 in Rudin (1987). Hence there exists a sequence $(h_{n,0})_{n\geq 1} \subset C_b(\lambda, \mathbb{R}^{d-1})$

with $\|h_{n,0} - h_0\|_{G_0,2} \to 0$. This implies that $0 = \int h_0 \,\mathrm{d}G_0 = \lim_{n\to\infty} \int h_{n,0} \,\mathrm{d}G_0$ and so also $\|\tilde{h}_{n,0} - h_0\|_{G_0,2} \to 0\|_{G_0,2} \to 0$ where $\tilde{h}_{n,0} := h_{n,0} - \int h_{n,0} \,\mathrm{d}G_0 \in H_0$, implying that $h_0 \in \operatorname{cl} H_0$. □

Lemma C.3. *Let $\tilde{H}_k^\gamma$ be defined as in the proof of Lemma B.2. We have that*

$$\mathscr{T} = \operatorname{cl}\left(\tilde{H}_1^\gamma + \cdots + \tilde{H}_K^\gamma\right),$$

and

$$\operatorname{cl} \mathcal{T}_{P_\theta,H}^{\eta|\gamma} = \operatorname{cl} \tilde{H}_0^\gamma + \operatorname{cl} \tilde{H}_1^\gamma + \cdots + \operatorname{cl} \tilde{H}_K^\gamma = \operatorname{cl} \tilde{H}_0^\gamma + \mathscr{T}.$$

Proof. For the first display, the sets in the sum on the right hand side are pairwise orthogonal. Note that we have for any $k, j \in [K]$ and any $(h_j, h_k) \in H_j \times H_k$,

$$\langle h_j(A_j v), h_k(A_k v)\rangle_{P_\theta} = P_\theta h_j(A_j v) h_k(A_k v) = \mathbb{E} h_j(\epsilon_j) h_k(\epsilon_k) = \mathbb{E} h_j(\epsilon_j) \mathbb{E} h_k(\epsilon_k) = 0,$$

due to the independence of the elements of ϵ. So $y \mapsto h_j(A_j v) \in [\tilde{H}_k^\gamma]^\perp = [\operatorname{cl} \tilde{H}_k^\gamma]^\perp$.[46] Recalling that the sum of closed pairwise orthogonal subspaces is closed,[47] we conclude that $\operatorname{cl}\left(\tilde{H}_1^\gamma + \cdots + \tilde{H}_K^\gamma\right) \subset \operatorname{cl} \tilde{H}_1^\gamma + \cdots + \operatorname{cl} \tilde{H}_K^\gamma = \mathscr{T}$ since the closure of a set is the smallest closed set containing that set. For the opposite inclusion, let $g = \sum_{k=1}^K g_k \in \mathscr{T}$ and note there are $g_{i,n}(y) = h_{i,n}(A_i v) \in \tilde{H}_i^\gamma$ such that each $g_{i,n} \to g_i$ in $L_2(P_\theta)$. Let $g_n = \sum_{k=1}^K g_{k,n}$. Clearly this is in $\tilde{H}_1^\gamma + \cdots + \tilde{H}_K^\gamma$ and hence its limit g is in $\operatorname{cl}\left(\tilde{H}_1^\gamma + \cdots + \tilde{H}_K^\gamma\right)$. Thus $\mathscr{T} \subset \operatorname{cl}\left(\tilde{H}_1^\gamma + \cdots + \tilde{H}_K^\gamma\right)$. The second display is analogous, noting the independence between $\tilde{X}$ and ϵ. □

Lemma C.4. *We have*

$$\mathscr{L} \subset \mathscr{T}^\perp,$$

where both are as defined in the proof of Lemma B.2.

Proof. Suppose that $y \mapsto f(y)$ is in $\mathscr{L}_0$ and let $y \mapsto \sum_{k=1}^K h_k(A_k v) \in \tilde{H}_1^\gamma + \cdots + \tilde{H}_K^\gamma$.[48] We have

$$\left\langle f(Y), \sum_{k=1}^K h_k(A_k V)\right\rangle_{P_\theta} = \sum_{k=1}^K \langle f(Y), h_k(A_k V)\rangle_{P_\theta},$$

where $V = Z - BX$ so it suffices to show that $\langle f(Y), h_k(A_k V)\rangle_{P_\theta} = 0$ for any $k \in [K]$ and any $h_k \in H_k$. First suppose that $f(Y) \in \{A_k V, \kappa(A_k V)\}$. Then, by the definition of H_k we have

$$\langle f(Y), h_k(A_k V)\rangle_{P_\theta} = \int f(y) h_k(A_k v) \,\mathrm{d}P_\theta = P_\theta[f(Y) h_k(A_k V)] = 0.$$

[46] Note that for any Hilbert space V and a linear subspace U of V, $U^\perp = [\operatorname{cl} U]^\perp$.

[47] See e.g. II.3.4 in Conway (1985).

[48] See Lemma C.3.

Second suppose that $f(Y) \in \{A_l V, \kappa(A_l V)\}$ for some $l \neq k$. Then we have that

$$\langle f(Y), h_k(A_k V)\rangle_{P_\theta} = \int f(y) h_k(A_k v)\, \mathrm{d}P_\theta = P_\theta[f(Y)h_k(A_k V)] = P_\theta f(Y) P_\theta h_k(A_k V) = 0,$$

by the independence of $A_k V = \epsilon_k$ and $A_l V = \epsilon_l$ and the fact that by the definition of H_k we have $P_\theta[h_k(A_k V)] = 0$. Now, let $i \neq j$ with both in $[K]$ and suppose that $f(Y) = \phi_i(A_i V) A_j V$. If $k = i \neq j$ we have

$$\langle f(Y), h_k(A_k V)\rangle_{P_\theta} = P_\theta[\phi_i(A_i V) A_j V h_k(A_k V)] = P_\theta[\phi_k(A_k V) h_k(A_k V)] P_\theta[A_j V] = 0,$$

by independence of $A_k V$ and $A_j V$ and that $P_\theta[A_j V] = 0$. If $k = j \neq i$,

$$\langle f(Y), h_k(A_k V)\rangle_{P_\theta} = P_\theta[\phi_i(A_i V) A_j V h_k(A_k V)] = P_\theta[h_k(A_k V) A_k V] P_\theta[\phi_i(A_i V)] = 0,$$

by independence of $A_k V$ and $A_i V$ and the definition of h_k. Lastly, if $k \neq j \neq i$ then

$$\langle f(Y), h_k(A_k V)\rangle_{P_\theta} = P_\theta[\phi_i(A_i V) A_j V h_k(A_k V)] = P_\theta[h_k(A_k V)] P_\theta[A_j V] P_\theta[\phi_i(A_i V)] = 0,$$

by independence of $A_k V, A_j V, A_i V$ and $P_\theta[A_j V] = 0$. □

Lemma C.5. *Let $\kappa(x) := x^2 - 1$ and let $L_2(G_k)$ denote the space of functions from $\mathbb{R} \to \mathbb{R}$ square-integrable with respect to the probability measure G_k, which is absolutely continuous with respect to Lebesgue measure, λ. Let $\mathcal{C}_b^1(\lambda) \subset L_2(G_k)$ denote the subspace of functions which are bounded and continuously differentiable with bounded derivatives λ-a.e. Suppose that $\kappa \in L_2(G_k)$, $\int z\, \mathrm{d}G_k = \int \kappa(z)\, \mathrm{d}G_k = 0$ and $\int \kappa(z)^2\, \mathrm{d}G_k > 0$. Then, there are functions $v, w \in \mathcal{C}_b^1(\lambda)$ such that*

$$\int v(z)\, \mathrm{d}G_k = \int w(z)\, \mathrm{d}G_k = 0,$$

$$\int z w(z)\, \mathrm{d}G_k = \int \kappa(z) v(z)\, \mathrm{d}G_k = 0$$

and

$$\int z v(z) = \int \kappa(z) w(z)\, \mathrm{d}G_k = 1.$$

Proof. We first note that the requirement that v, w be mean zero is easily met, once we have $\tilde{v}, \tilde{w}$ satisfying the other required properties. Suppose that is the case, then put $v := \tilde{v} - \int \tilde{v}(z)\, \mathrm{d}G_k$ and likewise for w. Clearly these are zero mean. Moreover, they are bounded and continuously differentiable with bounded derivative λ-a.e. and the inner product conditions also continue to hold in view of the assumption that $\int z\, \mathrm{d}G_k = \int \kappa(z)\, \mathrm{d}G_k = 0$. Therefore, we now construct $\tilde{v}, \tilde{w}$ ignoring the zero-mean requirement.

We start with $\tilde{v}$. Let $a < b < c$ and define

$$M := \begin{pmatrix} \int_a^b z \, \mathrm{d}G_k & \int_b^c z \, \mathrm{d}G_k \\ \int_a^b (z^2 - 1) \, \mathrm{d}G_k & \int_b^c (z^2 - 1) \, \mathrm{d}G_k \end{pmatrix}.$$

Provided M^{-1} exists there must exist a $v^* = (v_1^*, v_2^*)'$ such that $Mv^* = (1, 0)'$. Then, we can define

$$\tilde{v}(z) := \begin{cases} v_1^* & \text{if } z \in [a, b) \\ v_2^* & \text{if } z \in [b, c) \\ 0 & \text{otherwise} \end{cases},$$

to yield

$$\begin{pmatrix} \int z\tilde{v}(z) \, \mathrm{d}G_k \\ \int (z^2 - 1)\tilde{v}(z) \, \mathrm{d}G_k \end{pmatrix} = \begin{pmatrix} v_1^* \int_a^b z \, \mathrm{d}G_k + v_2^* \int_b^c z \, \mathrm{d}G_k \\ v_1^* \int_a^b (z^2 - 1) \, \mathrm{d}G_k + v_2^* \int_b^c (z^2 - 1) \, \mathrm{d}G_k \end{pmatrix} = Mv^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

as required. It remains to demonstrate that there are a, b, c such that M^{-1} exists. To see that this is always possible, note first that since $\int z \, \mathrm{d}G_k = 0$ and $\int z^2 \, \mathrm{d}G_k = 1$, G_k must place mass both on the negative and positive parts of the real line. Since also $\int (z^2 - 1) \, \mathrm{d}G_k = 0$ and $\int (z^2 - 1)^2 \, \mathrm{d}G_k > 0$ at least one of $G_k([-1, 0)) > 0$ or $G_k([0, 1)) > 0$ must hold. Without loss of generality assume the latter.[49] Take $a < 0$ such that $G_k((a, 0)) > 0$. Take $b = 0$ and $c < 1$ such that $G_k([0, c)) > 0$ and $G_k([c, 1)) > 0$. Note that this ensures that $\int_a^b z \, \mathrm{d}G_k < 0$ and $\int_b^c (z^2 - 1) \, \mathrm{d}G_k < 0$, so neither of the rows are 0. Now, either M is non-singular and we are done or there is a $\tau \neq 0$ such that $\int_a^b z \, \mathrm{d}G_k = \tau \int_a^b (z^2 - 1) \, \mathrm{d}G_k$ and $\int_b^c z \, \mathrm{d}G_k = \tau \int_b^c (z^2 - 1) \, \mathrm{d}G_k$. If $\tau > 0$, adjust c upwards to $c^* \in (c, 1)$ such that $G_k([c, c^*)) > 0$. We have

$$\int_b^{c^*} z \, \mathrm{d}G_k > \int_b^c z \, \mathrm{d}G_k = \tau \int_b^c (z^2 - 1) \, \mathrm{d}G_k > \tau \int_b^{c^*} (z^2 - 1) \, \mathrm{d}G_k.$$

If $\tau < 0$, adjust c downwards to $c' > 0$ with $c' < c$ such that $G_k([c', c)) > 0$. We have

$$\int_b^{c'} z \, \mathrm{d}G_k < \int_b^c z \, \mathrm{d}G_k = \tau \int_b^c (z^2 - 1) \, \mathrm{d}G_k < \tau \int_b^{c'} (z^2 - 1) \, \mathrm{d}G_k.$$

Since $\int_a^b z \, \mathrm{d}G_k = \tau \int_a^b (z^2 - 1) \, \mathrm{d}G_k$ continues to hold, the two rows are now linearly independent and hence M is invertible.

We have constructed a $\tilde{v} \in \mathcal{C}_b^1(\lambda)$ satisfying the required conditions. The construction for $\tilde{w}$ can be perfomed analogously, taking $w^* := M^{-1}(0, 1)'$. □

Lemma C.6 (Cf. Lemma A.5, Chen and Bickel, 2006). *Let* $\tilde{\phi}_k(z) = \gamma_k' b_k$*, with* $\gamma_k =$

[49] If instead $G_k([0, 1)) = 0$, an analogous argument can be made, interchanging the roles of a and c.

$-G_k[b_k b_k']^{-1} G_k c_k$ and $\phi_{k,n}$ is defined as in Assumption 2.4.2. If part (iv) of Assumption 2.4.2 holds, we have

$$G_k \left(\tilde{\phi}_k(\epsilon_{i,k}) - \phi_{k,n}(\epsilon_{i,k}) \right)^2 \leq C^2 \delta_{k,n}^6 \|\phi_{k,n}^{(3)}\|_\infty^2.$$

Proof. By the definition of $\tilde{\phi}_k$ and lemma C.10 we have

$$G_k \left(\tilde{\phi}_k(\epsilon_{i,k}) - \phi_{k,n}(\epsilon_{i,k}) \right)^2 = \inf_{g \in \mathscr{G}_k(\xi_{k,n})} G_k \left(g(\epsilon_{i,k}) - \phi_{k,n}(\epsilon_{i,k}) \right)^2 \leq C^2 \delta_{k,n}^6 \|\phi_{k,n}^{(3)}\|_\infty^2.$$

The first inequality comes from the fact that we can equivalently see $\gamma_k = -G_k[b_k b_k']^{-1} G_k c_k$ as the solution to minimizing

$$\begin{aligned} \int (\phi_k(z) - \gamma_k' b_k(z))^2 \eta_k(z) \, \mathrm{d}z &= \int \phi_k^2 \, \mathrm{d}G_k + \int (\gamma_k' b_k)^2 \, \mathrm{d}G_k + 2 \int \gamma_k' c_k(z) \eta_k(z) \, \mathrm{d}z \\ &= G_k \phi_k^2 + \gamma_k' G_k [b_k b_k'] \gamma_k + 2 \gamma_k' G_k c_k. \end{aligned} \tag{58}$$

where we only integrate over the support of $\phi_{k,n}$ since this is also the support of b_k and c_k. □

Lemma C.7 (Cf. Lemma A.3, Chen and Bickel, 2006)**.** *Under assumptions 2.4.1 and 2.4.2, and that $W_{i,n}$ is independent of $\epsilon_{i,k}$ we have*

$$\left\| \frac{1}{n} \sum_{i=1}^n b_k(\epsilon_{i,k}) W_{i,n} \right\|_2 = O_G(n^{-1/2}).$$

Proof. By the fact that $\sum_{m=1}^{B_k} b_{m,k}(x)^2 \leq 1$ (see e.g. (36) on p. 96 of de Boor, 2001) and the given assumptions we have that

$$G \left(\left\| \frac{1}{n} \sum_{i=1}^n b_k(\epsilon_{i,k}) W_{i,n} \right\|_2^2 \right) = \frac{1}{n} G_k \left(\sum_{m=1}^{B_k} b_{m,k}(\epsilon_{i,k})^2 \right) G_w W_{i,n}^2 \leq \frac{G_w W_{i,n}^2}{n}$$

Fix $\epsilon > 0$ and take $M > 0$ large enough such that $G_w W_{i,n}^2 / M^2 < \epsilon$. Markov's inequality yields

$$G \left(\sqrt{n} \left\| \frac{1}{n} \sum_{i=1}^n b_k(\epsilon_{i,k}) W_{i,n} \right\|_2 > M \right) \leq \frac{G \left(n \left\| \frac{1}{n} \sum_{i=1}^n b_k(\epsilon_{i,k}) W_{i,n} \right\|_2^2 \right)}{M^2} \leq \frac{G_w W_{i,n}^2}{M^2} < \epsilon.$$

□

Lemma C.8 (Cf. Lemma A.2, Chen and Bickel, 2006)**.** *Let $\hat{\gamma}_k$ be as defined in equation* (2.7) *and $\gamma_k = -G_k[b_k b_k']^{-1} G_k c_k$. Suppose that Assumptions 2.4.1 and 2.4.2 hold. Then,* *if we define*

$$\hat{\Gamma}_{k,n} := \frac{1}{n}\sum_{i=1}^{n} b_k(\epsilon_{i,k})b_k(\epsilon_{i,k})', \quad \Gamma_{k,n} := G_k b_k b_k',$$

and

$$\hat{C}_{k,n} := \frac{1}{n}\sum_{i=1}^{n} c_{k,n}(\epsilon_{i,k}), \quad C_{k,n} := G_k c_k,$$

we have that

(I) $\|C_{k,n}\|_2 = O(\delta_{k,n}B_k^{1/2})$,

(II) $\|\hat{C}_{k,n} - C_{k,n}\|_2 = O_G\left(\sqrt{\frac{B_k \log B_k}{n\delta_{k,n}^2}}\right)$,

(III) $\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 = O_G\left(\sqrt{\frac{B_k \log B_k}{n}}\right)$,

(IV) $\|\Gamma_{k,n}\|_2 = O(\delta_{n,k})$

(V) $\|\Gamma_{k,n}^{-1}\|_2 = O(\delta_{k,n}^{-2})$.

In particular, $\|\hat{\gamma}_k - \gamma_k\|_2 = O_G(n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-4}(\Delta_{k,n}\delta_{k,n}^{-1})^{\iota}) = o_G(1)$ *and* $\|\hat{\Gamma}_{k,n}\|_2 = o_G(1)$.

Proof. The proof follows the relevant parts of the proof of lemma A.2 in Chen and Bickel (2006). Firstly, from the representation of the derivative of the cubic spline (e.g. de Boor, 2001) we can write $c_{k,i} = \left(b_{k,i}^{(3)} - b_{k,i+1}^{(3)}\right)/\delta_{k,n}$. We have, for large enough $n \in \mathbb{N}$,

$$\begin{aligned}
|C_{k,n,i}| = |G_k c_{k,i}| &= \delta_{k,n}^{-1}\left|\int b_{k,i}^{(3)}(t)\eta_k(t)\,\mathrm{d}t - \int b_{k,i+1}^{(3)}(t)\eta_k(t)\,\mathrm{d}t\right| \\
&= \delta_{k,n}^{-1}\left|\int b_{k,i}^{(3)}(t)\eta_k(t)\,\mathrm{d}t - \int b_{k,i}^{(3)}(t)\eta_k(t+\delta_{k,n})\,\mathrm{d}t\right| \\
&\leq \left|\int b_{k,i}^{(3)}(t)\frac{\eta_k(t+\delta_{k,n}) - \eta_k(t)}{\delta_{k,n}}\,\mathrm{d}t\right| \\
&\leq 2\|\eta_k'\|_\infty \int b_{k,i}^{(3)}(t)\,\mathrm{d}t \\
&\leq 6\|\eta_k'\|_\infty \delta_{k,n},
\end{aligned}$$

where the last inequality is due to (20) on p. 91 in de Boor (2001) and the fact that splines (of any order) take values in $[0, 1]$.[50] It follows immediately that for large enough $n \in \mathbb{N}$,

$$\sum_{i=1}^{B_k} C_{k,n,i}^2 \leq \sum_{i=1}^{B_k} 6^2\|\eta_k'\|_\infty^2 \delta_{k,n}^2 = B_k 6^2 \|\eta_k'\|_\infty^2 \delta_{k,n}^2,$$

from which (I) follows.

[50] This is evident from their definition. See also property (36) (p. 96) of de Boor (2001).

We have that $c_{k,i} = \left(b^{(3)}_{k,i} - b^{(3)}_{k,i+1}\right)/\delta_{k,n}$ and since splines (of any order) take values in $[0,1]$ (both as noted above), we have that $c_{k,i} \in [-\delta^{-1}_{k,n}, \delta^{-1}_{k,n}]$. Hence, by Hoeffdings's inequality for $t \geq 0$ we have

$$G\left(\left|\frac{1}{n}\sum_{i=1}^{n} c_{k,m}(\epsilon_{i,k}) - G_k c_{k,m}\right| \geq t\right) \leq 2\exp\left(\frac{-n^2t^2}{2n\delta^{-2}_{k,n}}\right) = 2\exp(-nt^2\delta^2_{k,n}/2).$$

Therefore,

$$\begin{aligned} G\left(\|\hat{C}_{k,n} - C_{k,n}\|_2 \geq t\right) &\leq \sum_{m=1}^{B_k} G\left(\left|\frac{1}{n}\sum_{i=1}^{n} c_{k,m}(\epsilon_{i,k}) - G_k c_{k,m}\right| \geq \frac{t}{\sqrt{B_k}}\right) \\ &\leq 2B_k \exp(-nt^2 B_k^{-1}\delta^2_{k,n}/2), \end{aligned}$$

and so for any fixed $\epsilon > 0$ we can take $t = \sqrt{\frac{4B_k \log B_k}{n\delta^2_{k,n}}}$ to obtain

$$G\left(\|\hat{C}_{k,n} - C_{k,n}\|_2 \geq t\right) \leq 2B_k^{-1} \to 0,$$

yielding (II).

Since for any $m, s \in [B_k]$ we have $b_{k,m}b_{k,s} \in [0,1]$ by Hoeffding's inequality it follows that for any $t \geq 0$

$$G\left(\left|\frac{1}{n}\sum_{i=1}^{n} b_{k,m}(\epsilon_{i,k})b_{k,s}(\epsilon_{i,k}) - G_k b_{k,m}b_{k,s}\right| \geq t\right) \leq 2\exp\left(\frac{-2n^2t^2}{n}\right) = 2\exp(-2nt^2).$$

Therefore, since $\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \leq \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_F$ and both $\hat{\Gamma}_{k,n}$ and $\Gamma_{k,n}$ are zero for all (m,s) entries where $|m-s| > 3$ (de Boor, 2001, (20), p. 91) we have that

$$\begin{aligned} &G\left(\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \geq t\right) \\ &\quad\leq G\left(\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_F \geq t\right) \\ &\quad\leq \sum_{m=1}^{B_k} \sum_{s=\max(m-3,1)}^{\min(B_k,m+3)} G\left(\left|\frac{1}{n}\sum_{i=1}^{n} b_{k,m}(\epsilon_{i,k})b_{k,s}(\epsilon_{i,k}) - G_k b_{k,m}b_{k,s}\right| \geq \frac{t}{\sqrt{7B_{k,n}}}\right) \\ &\quad\leq 14B_k \exp\left(\frac{-2nt^2}{7B_k}\right). \end{aligned}$$

Putting $t = \sqrt{\frac{7B_k \log B_k}{n}}$ we obtain

$$G\left(\|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \geq t\right) \leq 14B_k^{-1} \to 0,$$

yielding (III).

Since $\Gamma_{k,n}$ is symmetric and positive (semi-)definite we have that $\|\Gamma_{k,n}\|_2 \leq \|\Gamma_{k,n}\|_\infty = \max_{m=1,\dots,B_k} \sum_{s=1}^{B_k} G_k b_{k,m} b_{k,s}$.[51] Then, since for any $z \in \mathbb{R}$, each row of $b_k(z)b_k(z)'$ has at most 7 non-zero entries,[52] all of which are bounded above by 1 we have

$$\begin{aligned}
\|\Gamma_{k,n}\|_2 &\leq \max_{m=1,\dots,B_k} \sum_{s=1}^{B_k} G_k b_{k,m} b_{k,s} \\
&= \max_{m=1,\dots,B_k} \sum_{s=1}^{B_k} \int_{\xi_{k,n,m}}^{\xi_{k,n,m+4}} b_{k,m}(z) b_{k,s}(z) \eta_k(z)\, \mathrm{d}z \\
&\leq \max_{m=1,\dots,B_{k,n}} 7\|\eta_k\|_\infty 4\delta_{k,n} \\
&= 28\|\eta_k\|_\infty \delta_{k,n},
\end{aligned}$$

which yields (IV) in conjunction with requirement (iii) of Assumption 2.4.2.

By Assumption 2.4.2 part (v), on $[\Xi^L_{k,n}, \Xi^U_{k,n}]$ we have $\eta(x) \geq c\delta_{k,n}$. Hence $\eta(x) - c\delta_{k,n} \geq 0$ and so $\int b_k b_k'(\eta - c\delta_{k,n})\lambda = \int (b_k\sqrt{\eta - c\delta_{k,n}})(b_k\sqrt{\eta - c\delta_{k,n}})'\lambda$. Note that the functions $b_{k,i}\sqrt{\eta - c\delta_{k,n}}$ satisfy $\int (b_{k,i}\sqrt{\eta - c\delta_{k,n}})^2\, \mathrm{d}\lambda < \infty$ and hence belong to $L_2(\lambda)$. It follows that the matrix $\int b_k b_k'(\eta - c\delta_k)\lambda$ is a Gram matrix and hence positive semi-definite. This implies that $\Gamma_{k,n} \succeq c\delta_{k,n}\tilde{\Gamma}_{k,n}$ where $\tilde{\Gamma}_{k,n}$ is defined as in lemma C.9. Hence, by the Rayleigh quotient theorem (see e.g. Theorem 4.2.2 in Horn and Johnson, 2013) and lemma C.9

$$\lambda_{\min}(\Gamma_{k,n}) \geq \lambda_{\min}(c\delta_{k,n}\tilde{\Gamma}_{k,n}) = c\delta_{k,n}\lambda_{\min}(\tilde{\Gamma}_{k,n}) \geq cv\delta_{k,n}^2,$$

for a $v > 0$, from which we may conclude that

$$\|\Gamma_{k,n}^{-1}\|_2 = \frac{1}{\lambda_{\min}(\Gamma_{k,n})} \leq (cv)^{-1}\delta_{k,n}^{-2},$$

which yields (V).

To demonstrate the last claim, note that with the results just derived, under our assumptions we have,

$$\|\hat{C}_{k,n}\|_2 \leq \|\hat{C}_{k,n} - C_{k,n}\|_2 + \|C_{k,n}\|_2 = O_G\left(\sqrt{\frac{B_k \log B_k}{n\delta_{k,n}^2}}\right) + O\left(\delta_{k,n}\sqrt{B_k}\right) = O_G\left(\delta_{k,n}\sqrt{B_k}\right),$$

[51] See e.g. Theorem 5.6.9 in Horn and Johnson (2013).

[52] $b_{k,m}(z) = 0$ outside $[\xi_{k,n,m}, \xi_{k,n,m+4})$. See (20) on p. 91 in de Boor (2001).

and, using inequality (5.8.2) from Horn and Johnson (2013),

$$\begin{aligned}
\|\hat{\Gamma}_{k,n}^{-1}\|_2 &\leq \|\Gamma_{k,n}^{-1}(I + [\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1})^{-1}\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \|(I + [\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1})^{-1}\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \left(1 - \|[\hat{\Gamma}_{k,n} - \Gamma_{k,n}]\Gamma_{k,n}^{-1}\|_2\right)^{-1} \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \left(1 - \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 \|\Gamma_{k,n}^{-1}\|_2\right)^{-1} \\
&= O_G(\delta_{k,n}^{-2}).
\end{aligned} \tag{59}$$

Using these intermediate results along with (II) - (V) and our hypotheses we obtain that

$$\begin{aligned}
\|\hat{\gamma}_k - \gamma_k\|_2 &= \|\hat{\Gamma}_{k,n}^{-1}\hat{C}_{k,n} - \Gamma_{k,n}^{-1}C_{k,n}\|_2 \\
&\leq \|(\hat{\Gamma}_{k,n}^{-1} - \Gamma_{k,n}^{-1})\hat{C}_{k,n}\|_2 + \|\Gamma_{k,n}^{-1}(\hat{C}_{k,n} - C_{k,n})\|_2 \\
&\leq \|\Gamma_{k,n}^{-1}\|_2 \|\Gamma_{k,n} - \hat{\Gamma}_{k,n}\|_2 \|\hat{\Gamma}_{k,n}^{-1}\|_2 \|\hat{C}_{k,n}\|_2 + \|\Gamma_{k,n}^{-1}\|_2 \|\hat{C}_{k,n} - C_{k,n}\|_2 \\
&= O_G\left(\sqrt{\frac{B_k^2 \log B_k}{\delta_{k,n}^6 n}}\right) + O_G\left(\sqrt{\frac{B_k \log B_k}{\delta_{k,n}^6 n}}\right) \\
&= o_G(1),
\end{aligned}$$

by Assumption 2.4.2 part (ii), since we have $B_k \leq \Delta_{k,n}\delta_{k,n}^{-1}$ and hence the dominant term above vanishes since for all large enough n,

$$\sqrt{\frac{B_k^2 \log B_k}{\delta_{k,n}^6 n}} \leq n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-4}\log(\Delta_{k,n}\delta_{k,n}^{-1}) \leq n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-4}(\Delta_{k,n}\delta_{k,n}^{-1})^{\iota} = o(1).$$

Finally, by (III) and (IV) and Assumption 2.4.2 part (ii) we have

$$\|\hat{\Gamma}_{k,n}\|_2 \leq \|\hat{\Gamma}_{k,n} - \Gamma_{k,n}\|_2 + \|\Gamma_{k,n}\|_2 = O_G\left(\sqrt{\frac{B_{k,n}\log B_k}{n}}\right) + O(\delta_{k,n}) = o_G(1),$$

since $\delta_{k,n} \to 0$ and for large enough n,

$$\sqrt{\frac{B_k \log B_k}{n}} \leq n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-1}\log(\Delta_{k,n}\delta_{k,n}^{-1}) \leq \delta_{k,n}^3 n^{-1/2}\Delta_{k,n}\delta_{k,n}^{-4}(\Delta_{k,n}\delta_{k,n}^{-1})^{\iota} = o(1).$$

□

Lemma C.9. *The smallest eigenvalue of the $B_k \times B_k$ Gram matrix $\tilde{\Gamma}_{k,n} := \int b_k b_k' \,\mathrm{d}\lambda$ satisfies*

$$\lambda_{\min}(\tilde{\Gamma}_{k,n}) \geq \upsilon\delta_{k,n} > 0,$$

for a $\upsilon > 0$.

Proof. Since $b_{k,m}(x)b_{k,s}(x)$ is non-zero only for $|m - s| \leq 3$ and each $b_{k,m}$ is non-zero only on $[\xi_{k,n,m}, \xi_{k,n,m+4})]$ (e.g. (20) p. 91 of de Boor, 2001), $\tilde{\Gamma}_{k,n}$ is a symmetric banded Toeplitz matrix.[53] Its entries can be computed by direct integration:

$$[\tilde{\Gamma}_{k,n}]_{m,s} = \delta_{k,n} \times \begin{cases} \frac{151}{315} & \text{if } m = s \\ \frac{397}{1680} & \text{if } |m-s| = 1 \\ \frac{1}{42} & \text{if } |m-s| = 2 \\ \frac{1}{5040} & \text{if } |m-s| = 3 \\ 0 & \text{if } |m-s| > 3 \end{cases} .$$

For simplicity of notation let $f_0 := \frac{151}{315}$, $f_1 := f_{-1} := \frac{397}{1680}$, $f_2 := f_{-2} := \frac{1}{42}$ and $f_3 := f_{-3} := \frac{1}{5040}$ and let $f_s := 0$ for $|s| > 3$. Now, let $f(\theta) := \sum_{s=-3}^{3} f_s e^{i(s\theta)}$. Then, $\tilde{\Gamma}_{k,n}/\delta_{k,n}$ is then the matrix generated by f in the sense that $\tilde{\Gamma}_{k,n}/\delta_{k,n} = \mathscr{T}_n(f) := \sum_{s=-\min(B_k-1,3)}^{\min(B_k-1,3)} f_k J_n^s$ where each J_n^s is the $B_k \times B_k$ matrix which is zero everywhere except for the (i,j)-th entries where $i - j = s$, where it has a value of 1.[54] Since $f \in L_1([-\pi,\pi])$ and is real on $[-\pi,\pi]$ by Theorem 6.1 in Garoni and Serra-Capizzano (2017) we have that $\lambda_{\min}(\tilde{\Gamma}_{k,n}) = \delta_{k,n}\lambda_{\min}(\tilde{\Gamma}_{k,n}/\delta_{k,n}) \geq \delta_{k,n} \inf_{\theta\in[-\pi,\pi]} f(\theta) = \delta_{k,n}\upsilon$, where $\upsilon := \inf_{\theta\in[-\pi,\pi]} f(\theta) > 0$. □

Lemma C.10. *Suppose $\xi \in \mathbb{R}^{N+1}$ such that $a = \xi_0 < \xi_1 < \cdots < \xi_N = b$, $h := \max_{i\in[N]} \xi_i - \xi_{i-1}$, and let $\mathscr{G}_k(\xi)$ be the linear space formed by degree k splines with knots ξ. Then, if $f \in C^{k-1}[a,b]$ we have that*

$$\inf_{g\in\mathscr{G}_k(\xi)} \|g - f\|_\infty \leq \frac{(k+1)!}{2^k} h^{k-1} \|f^{(k-1)}\|_\infty = c_k h^{k-1} \|f^{(k-1)}\|_\infty,$$

where c_k depends only on k.

Proof. This follows as a special case of Theorem 20.3 in Powell (1981). □

D. Additional auxillary results

We present a few additional results that explicitly prove some claims made in the main text. First, we show that if two errors $\epsilon_{i,k}$ and $\epsilon_{i,j}$ are Gaussian the efficient information matrix becomes singular. Second, we provide an explicit example of a density which satisfies the first part of the Assumption 2.4.1 but not the second. Third we prove that if Assumption 2.4.1 part 1 holds then a sufficient condition for part 2 is that η_k has tails that decay to zero at a polynomial rate.

[53] As can be easily verified, unlike in the case of linear ($\kappa = 2$) or quadratic splines ($\kappa = 3$), this matrix is *not* diagonally dominant. In the case of $\kappa \in \{2,3\}$ this argument could be completed in a simpler fashion by using the Gershgorin circle theorem.

[54] See section 6.1 in Garoni and Serra-Capizzano (2017), noting that it is clear that $f \in L_1([-\pi,\pi])$.

Lemma D.1. *Consider the LSEM model* (2.3) *with $B = 0$ (for ease of exposition only) and Assumption 2.4.1 parts 1-3 hold. Define the vector-valued function $Q : \mathbb{R}^K \to \mathbb{R}^{K^2}$ according to*

$$Q(y) = (Q_1(y)', \dots, Q_K(y)')',$$

where each $Q_k : \mathbb{R}^K \to \mathbb{R}^K$ and the j-th element of Q_k for $j \in [K]$ is given by

$$Q_{k,j}(y) = \begin{cases} \phi_k(A_k y) A_j y & \text{if } k \neq j \\ \tau_{k,1} A_k y + \tau_{k,2} \kappa(A_k y) & \text{if } k = j \end{cases}.$$

Next define the $K^2 \times L$ matrix ζ according to $\zeta = (\mathrm{vec}\,[D_1(\alpha)A^{-1}]', \dots, \mathrm{vec}\,[D_L(\alpha)A^{-1}]')$, where in the definition of both Q and ζ we have $A = A(\gamma)$. Equipped with these definitions, we can write the efficient score function as defined in lemma B.2 as

$$\tilde{\ell}_\theta(y) = \zeta' Q(y). \tag{60}$$

Then,

(I) *$\mathbb{E}_\theta QQ'$ is non-singular if and only if for each pair (k, j) with $k \neq j$ and each $k, j \in [K]$ we have that $[\mathbb{E}_\theta \phi_k^2(A_k Y)][\mathbb{E}_\theta \phi_j^2(A_j Y)] \neq 1$.*

(II) *$\tilde{I}_\theta$ is non-singular if $\mathrm{rank}(\zeta) = L$ and $\mathbb{E}_\theta QQ'$ is non-singular.*

(III) *If $\mathrm{rank}(\zeta) < L$ then $\tilde{I}_\theta$ is singular.*

(IV) *If $L = K^2$ and $\mathbb{E}_\theta QQ'$ is singular then $\tilde{I}_\theta$ is singular.*

(V) *If $\mathbb{E}_\theta QQ'$ is singular, $\tilde{I}_\theta$ may be singular when $\mathrm{rank}(\zeta) = L < K^2$.*

In particular, if both ϵ_k and ϵ_j ($k \neq j$) have a Gaussian distribution and $L = K^2$, $\tilde{I}_\theta$ is singular.

Proof. For (I), let j, k, m, i all be in $[K]$. We will consider the entries of the matrix $\mathbb{E}_\theta QQ'$, which are of the form $\langle Q_{k,j}, Q_{m,i} \rangle_{P_\theta}$. In particular, the s,t-th element of the matrix is given by the form $\langle Q_{k,j}, Q_{m,i} \rangle_{P_\theta}$ where $(k-1)K + j = s$ and $(m-1)K + i = t$. If $k = j = m = i$ we have $s = t$ and $\langle Q_{k,j}, Q_{m,i} \rangle_{P_\theta} = \mathbb{E}_\theta[\tau_{k,1} A_k Y + \tau_{k,2} \kappa(A_k Y)]^2$. The other diagonal entries occur when $k = m \neq j = i$, and have the form $\langle Q_{k,j}, Q_{m,i} \rangle_{P_\theta} = \mathbb{E}_\theta[\phi_k^2(A_k Y)]$. Inspection of the other possible cases reveals that the only other case with non-zero entries is $k = i \neq m = j$ which has value $\langle Q_{k,j}, Q_{m,i} \rangle_{P_\theta} = \mathbb{E}_\theta[\phi_k(A_k Y) A_k Y] \mathbb{E}_\theta[\phi_k(A_m Y) A_m Y] = 1$ by assumption 2.4.1.

Therefore for any $k, j \in [K]$, column $(k-1)K + j$ has non-zero entries in row $(k-1)K + j$ only if $k = j$ and otherwise in rows $(k-1)K + j$ and $(j-1)K + k$, with values $\mathbb{E}_\theta \phi_k^2(A_k Y)$ and 1 respectively. There are therefore no columns that can be linearly related to column $(k-1)K + j$ if $k = j$. If $k \neq j$, then column $(k-1)K + j$ has zeros everywhere except row $(k-1)K+j$ where it has $\mathbb{E}_\theta \phi_k^2(A_k Y)$ and row $(j-1)+k$ where it has 1. Column $(j-1)K+k$ has zeros everywhere except row $(j-1)K+k$ where it has $\mathbb{E}_\theta \phi_j^2(A_j Y)$ and row $(k-1)K+j$ where it has 1. Since no other columns have entries in these rows, it follows that column $(k-1)K+j$ is linearly independent of all the other columns if and only if it is linearly independent of column $(j-1)K+k$, which occurs if and only if $[\mathbb{E}_\theta \phi_k^2(A_k Y)][\mathbb{E}_\theta \phi_j^2(A_j Y)] \neq 1$.

For (II), suppose that $\operatorname{rank}(\zeta) = L$ and $\mathbb{E}_\theta QQ'$ is non-singular. Then there is a (unique) positive definite $[\mathbb{E}_\theta QQ']^{1/2}$ and we have $\tilde{I}_\theta = \left([\mathbb{E}_\theta QQ']^{1/2}\zeta\right)' \left([\mathbb{E}_\theta QQ']^{1/2}\zeta\right)$ which has full rank, since $\left([\mathbb{E}_\theta QQ']^{1/2}\zeta\right)$ has full column rank.

For the remaining parts note first that

$$\tilde{I}_\theta = \mathbb{E}_\theta \tilde{\ell}_\theta \tilde{\ell}_\theta' = \zeta' \left[\mathbb{E}_\theta QQ'\right] \zeta,$$

and so $\operatorname{rank}(\tilde{I}_\theta) \leq \min\{\operatorname{rank}(\zeta' \mathbb{E}_\theta QQ'), \operatorname{rank}(\zeta)\}$. Hence if $\operatorname{rank}(\zeta) < L$, $\operatorname{rank}(\tilde{I}_\theta) < L$ implying (III).

For (IV), suppose that $\operatorname{rank}(\mathbb{E}_\theta QQ') < K^2 = L$. Then, there is a non-zero $x \in \mathbb{R}^L$ such that $\mathbb{E}_\theta QQ'x = 0$ and hence $\zeta' \mathbb{E}_\theta QQ'x = 0$. Hence $\dim(N(\zeta' \mathbb{E}_\theta QQ')) \geq 1$. It follows that $\operatorname{rank}(\zeta' \mathbb{E}_\theta QQ') \leq L - 1 < L$ and hence $\operatorname{rank}(\tilde{I}_\theta) \leq \min\{\operatorname{rank}(\zeta' \mathbb{E}_\theta QQ'), \operatorname{rank}(\zeta)\} < L$.

For (V) suppose that $K = 2$, ϵ_1 and ϵ_2 are both Gaussian and $A(\gamma) = \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) \\ \sin(\gamma) & \cos(\gamma) \end{bmatrix}$. We have for $l \in \{1, 2\}$, $\phi_l(z) = -z$, hence $\phi_l^2(z) = z^2$ and so $\mathbb{E}_\theta \phi_l^2(\epsilon_l) = \mathbb{E}_\theta \phi_l^2(A_l Y) = 1$. $D_1(\gamma) = \begin{bmatrix} -\sin(\gamma) & -\cos(\gamma) \\ \cos(\gamma) & -\sin(\gamma) \end{bmatrix}$ and hence

$$D_1(\gamma) A(\gamma)^{-1} = D_1(\gamma) A(\gamma)' = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

which implies $\zeta = (0, -1, 1, 0)'$ and hence $\operatorname{rank}(\zeta) = 1 = L < K^2 = 4$. Explicit calculation reveals that

$$E_\theta QQ' = \begin{bmatrix} 8/9 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 8/9 \end{bmatrix},$$

which is clearly singular with rank 3. We have

$$\tilde{I}_\theta = \zeta' \left[\mathbb{E}_\theta QQ'\right] \zeta = \zeta' \begin{bmatrix} 8/9 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 8/9 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} = \zeta' \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = 0.$$

For the last part, suppose that $k \neq j$ and ϵ_k and ϵ_j are both Gaussian. Since both have zero mean and unit variance, we have for $l \in \{k, j\}$, $\phi_l(z) = -z$, hence $\phi_l^2(z) = z^2$ and so $\mathbb{E}_\theta \phi_l^2(\epsilon_l) = \mathbb{E}_\theta \phi_l^2(A_l Y) = 1$. $E_\theta QQ'$ is singular by (I) and hence by (IV) $\tilde{I}_\theta$ is singular. □

Example 3 (Necessity of part 2 of assumption 2.4.1)**.** *Suppose that $\tilde{\epsilon}_k \sim \chi_2^2$ and let $\epsilon_k = (\tilde{\epsilon}_k - 2)/2$. Then ϵ_k has mean zero, variance one and density function $\eta_k(z) = \exp(-z-1)$ on its support $[-1, \infty)$ on which we also have that $\phi_k(z) = -1$. Explicit calculation reveals that part 1 of assumption 2.4.1 is satisfied. However, $\mathbb{E}\phi_k(z) = -1 \neq 0$ as would be required by part 2 of assumption 2.4.1.*

Note also that this example does not satisfy the requirements of lemma D.2: we have $a_k = -1, b_k = \infty$ and

$$\lim_{z \downarrow a_k} \eta_k(x) = \lim_{z \downarrow -1} \exp(-z-1) = 1 \neq 0,$$

and hence the required condition is violated for $r = 0$.

Lemma D.2. *Let $a_k = \inf\{x \in \mathbb{R} \cup \{-\infty\} : \eta_k(x) > 0\}$ and $b_k = \sup\{x \in \mathbb{R} \cup \{\infty\} : \eta_k(x) > 0\}$. Suppose that, for $r = 0, 1, 2, 3$: (i) if $a_k = -\infty$ then $\eta_k(x) = o(x^{-3})$ as $x \to -\infty$, else $a_k^r \lim_{x \downarrow a_k} \eta_k(x) = 0$, and (ii) if $b_k = \infty$ then $\eta_k(x) = o(x^{-3})$ as $x \to \infty$, else $b_k^r \lim_{x \uparrow b_k} \eta_k(x) = 0$. Then, if part 1 of assumption 2.4.1 holds, part 2 is also satisfied.*

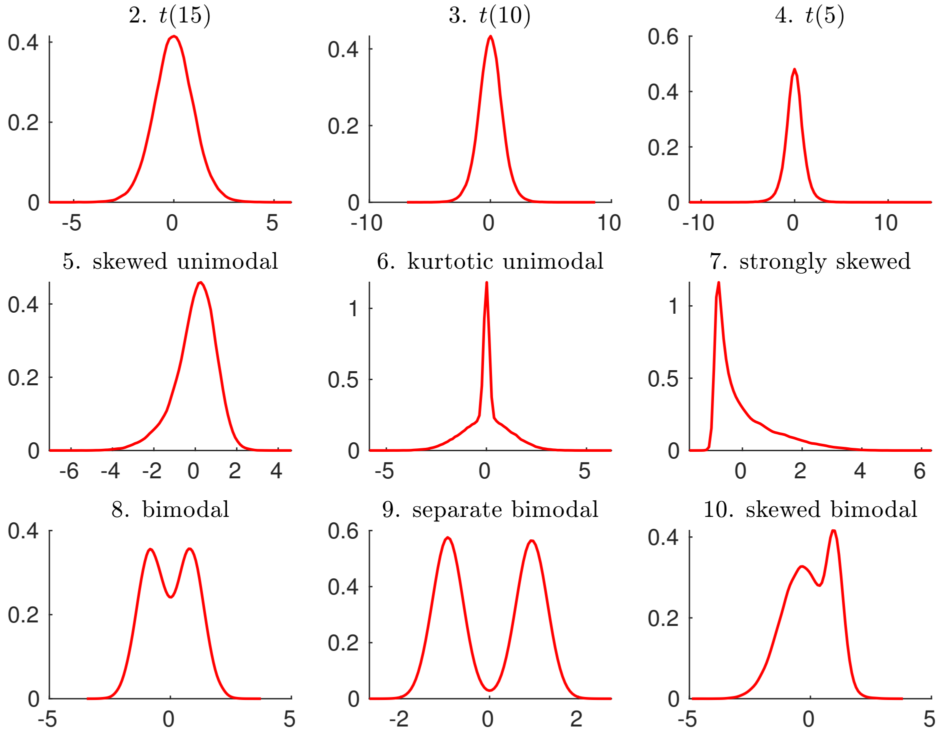
Proof. Let $r \in \{0, 1, 2, 3\}$, $b_k = \sup\{x \in \mathbb{R} : \eta_k(x) > 0\}$ and $a_k = \inf\{x \in \mathbb{R} : \eta_k(x) > 0\}$. We have, by integration by parts, with G_k denoting the measure on $\mathbb{R}$ corresponding to η_k,

$$\int \phi_k(z) z^r \, \mathrm{d}G_k = \int \frac{\eta_k'(z)}{\eta_k(z)} \eta_k(z) z^r \, \mathrm{d}z = \int \eta_k'(z) z^r \, \mathrm{d}z = \eta_k(z) z^r \Big|_{a_k}^{b_k} - \int \eta_k(z) \frac{\mathrm{d}z^r}{\mathrm{d}z} \, \mathrm{d}z.$$

Our hypothesis ensures that $z^r \eta_k(z)\big|_{a_k}^{b_k} = 0$. Therefore we have $G_k \phi_k(z) z^r = -G_k \frac{\mathrm{d}}{\mathrm{d}z} z^r$. For $r = 0$ this equals zero as $\frac{\mathrm{d}}{\mathrm{d}z} z^0 = \frac{\mathrm{d}}{\mathrm{d}z} 1 = 0$. For $r \in \{1, 2, 3\}$ we have $\frac{\mathrm{d}z^r}{\mathrm{d}z} = r z^{r-1}$ and hence $G_k \phi_k(z) z^r = -r G_k z^{r-1}$. Since $G_k 1 = 1$, $G_k z = 0$, and $G_k z^2 = 1$, the result follows. □

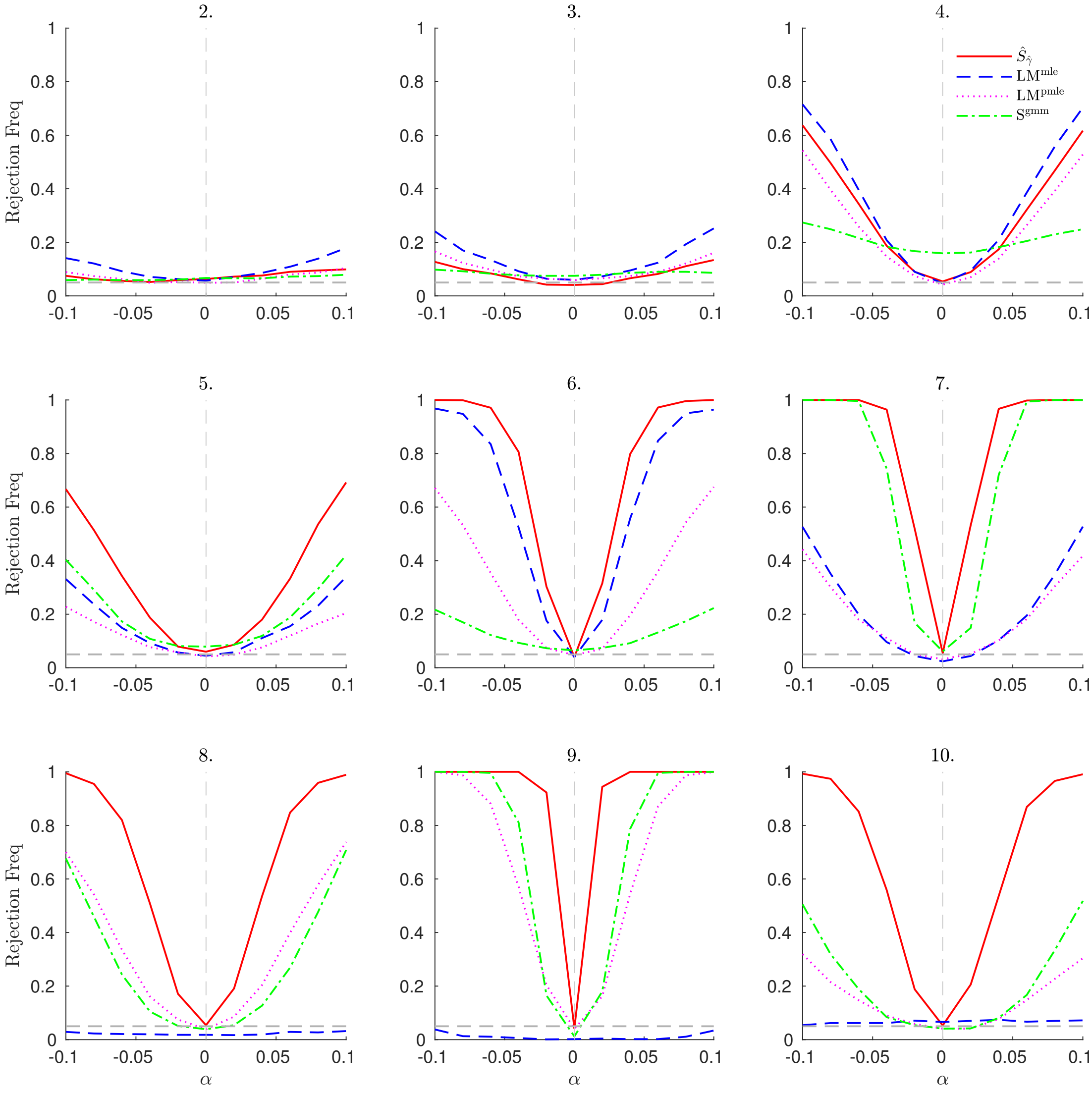
E. Figures and tables

Figure E.3: Structural Shock Densities



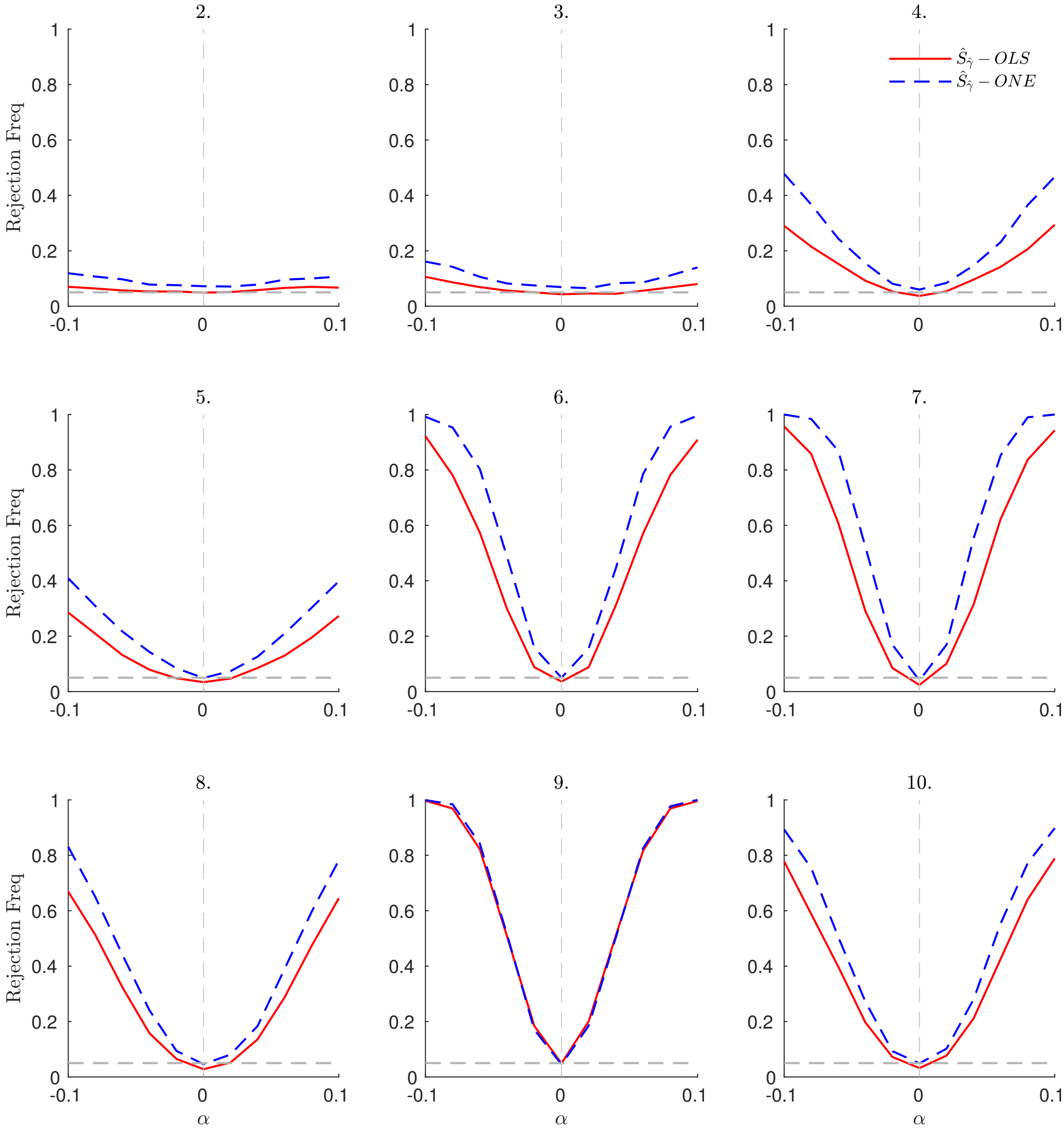
Notes: The plots show the different densities considered for simulating the structural shocks. Densities 2-4 are t-distributions normalised to have unit variance. Densities 5 - 10 (and their names) are mixtures of normals taken from Marron and Wand (1992); see their table 1 for the definitions. Density 1 is the standard Gaussian and omitted from the figure.

Figure E.4: Power Comparison Baseline model



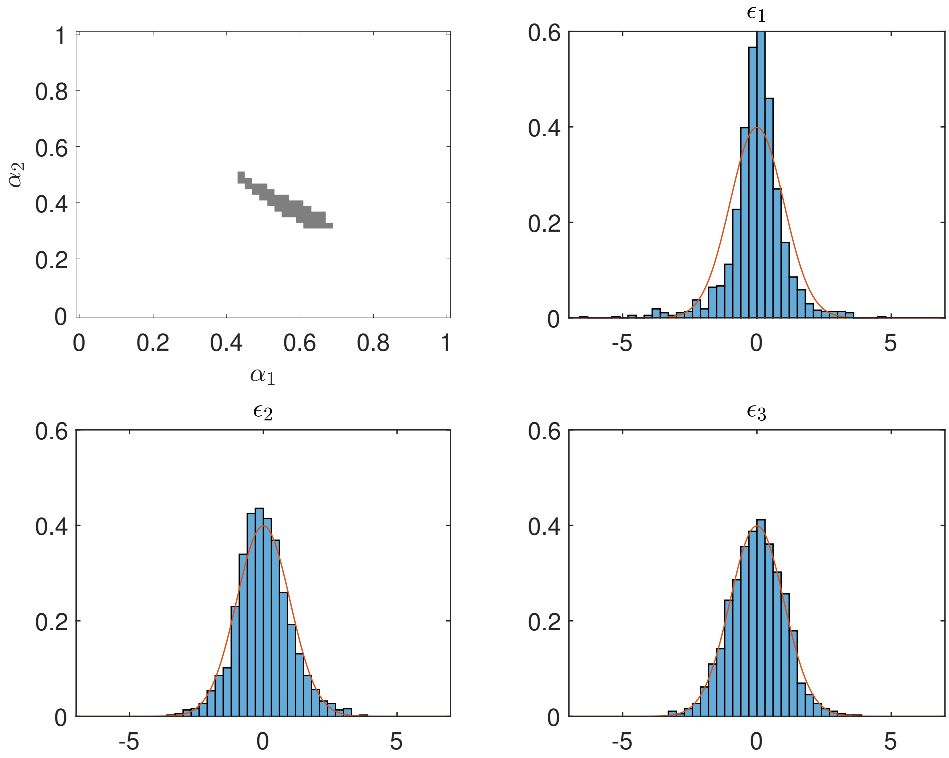
Notes: Empirical power curves for the baseline model with $k = 2$ and $n = 1000$. Each plot corresponds to the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure E.3. The solid red line corresponds to $S_{\hat{\gamma}}$, the dashed blue line to LM^{mle}, the dotted pink line to LM^{pmle} and the dot-dashed green line to S^{gmm}.

Figure E.5: Power LSEM



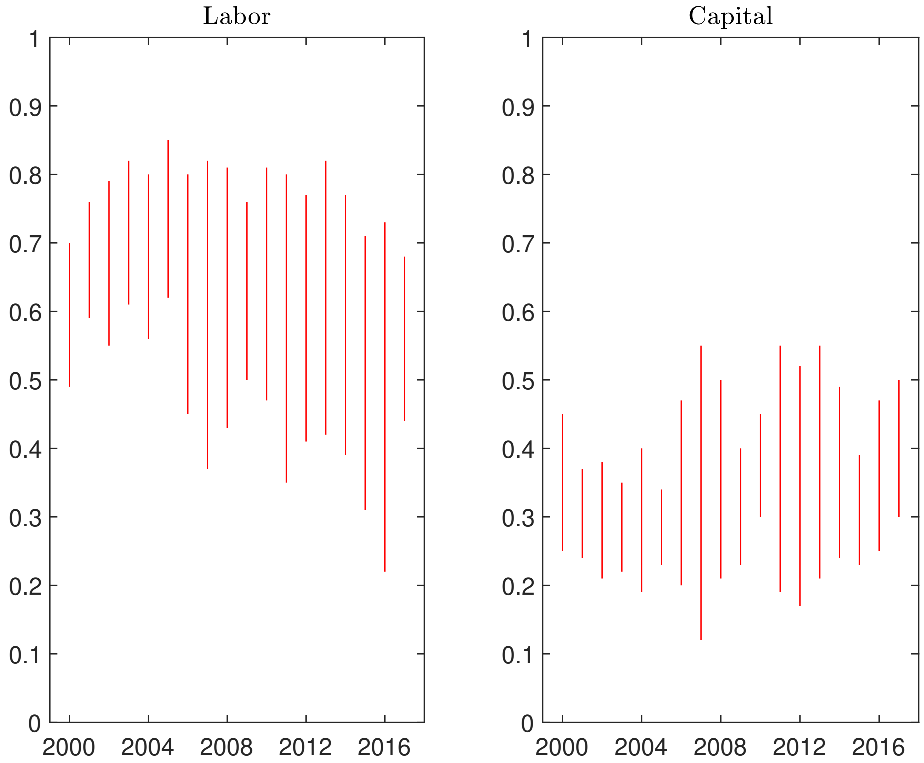
Notes: Empirical power curves for the LSEM model with $k = 2$, $d = 2$ and $n = 1000$. Each plot corresponds to the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure E.3. The solid red line corresponds to the empirical rejection frequency of the $\hat{S}_{\hat{\gamma}}$ test where $\hat{\gamma} = (\alpha_0, \hat{\beta})$, with $\hat{\beta}$ the OLS estimator. The dashed blue line corresponds to the empirical rejection frequency of the $\hat{S}_{\hat{\gamma}}$ test where $\hat{\gamma} = (\alpha_0, \hat{\beta})$, with $\hat{\beta}$ the one-step efficient MLE estimator.

Figure E.6: LSEM Production Function Output 2017



Notes: The top left panel shows the confidence region for the labor α_1 and capital α_2. The other three panels show the empirical densities of the residuals together with the standard normal distribution.

Figure E.7: Confidence intervals labor and capital 2000-2017



Notes: The vertical lines describe the confidence bands for labor and capital for each year between 2000 and 2017. Each pair of bands is based on firms observed in the corresponding year and estimated using the LSEM .

F. Additional simulation results

In this section we provide a number of additional simulation results.

F.1. Additional power results for the baseline model

Figure E.4 in the main text compared the power of different tests for the baseline model $Y_i = R'\epsilon_i$ for the case where $n = 1000$. Here we show the results for $n = 200$ and $n = 500$. Specifically, Figures F.8 and F.9 show the results.

Overall, the patterns that we find are similar as in the main text. One thing that stands out is that the $\mathrm{S}^{\mathrm{gmm}}$ test is not correctly sized for these smaller sample sizes, essentially confirming the results in Table E.3. It is possible that a more careful selection of the relevant higher order moments will improve this finding.

Besides this our two main findings from the main text hold. First, the standard LM test is the preferred approach whenever the true density is known, but the semi-parametric score test comes close in terms of power. Second, for all other densities the semi-parametric score test shows the highest power.

F.2. Additional power results for the LSEM

Figure E.5 compared the power of different tests for the LSEM model for the case where $n = 1000$. Here we show the results for $n = 200$ and $n = 500$. Specifically, Figures F.10 and F.11 show the results.

We find that for $n = 200$ the power of tests is generally quite low, indicating that for small sample sizes little can be learned by exploiting deviations from the Gaussian density. This holds most notably for the Student's t densities, the skewed unimodal density and the bimodal density. Intuitively, given a small sample these densities are hard to distinguish from the normal density and little can be learned about the parameter α. A reassuring finding is that the size of the test remains well controlled. These findings persist, to a lesser extent, when we increase to $n = 500$.

Overall, the implementing the test with one-step efficient estimates leads to higher power, but the size of the test is controlled less well. Therefore we recommend using OLS estimates for β when the sample size is small.

Table E.2: Rejection Frequencies $\hat{S}_{\hat{\gamma}}$ test for Baseline model

n	K	B	1	2	3	4	5	6	7	8	9	10
200	2	4	0.049	0.049	0.048	0.040	0.047	0.049	0.034	0.049	0.048	0.048
200	2	6	0.048	0.045	0.049	0.044	0.048	0.053	0.047	0.045	0.058	0.051
200	2	8	0.050	0.049	0.047	0.044	0.048	0.048	0.053	0.050	0.051	0.047
200	3	4	0.043	0.039	0.039	0.039	0.044	0.048	0.026	0.049	0.052	0.050
200	3	6	0.045	0.038	0.040	0.044	0.041	0.048	0.044	0.047	0.052	0.043
200	3	8	0.047	0.046	0.040	0.040	0.044	0.048	0.042	0.049	0.044	0.051
200	5	4	0.032	0.034	0.033	0.034	0.035	0.039	0.015	0.041	0.045	0.043
200	5	6	0.037	0.033	0.036	0.032	0.032	0.040	0.043	0.045	0.043	0.044
200	5	8	0.039	0.038	0.038	0.030	0.035	0.043	0.045	0.040	0.041	0.038
500	2	4	0.053	0.046	0.053	0.045	0.047	0.052	0.031	0.049	0.045	0.046
500	2	6	0.048	0.049	0.048	0.048	0.049	0.052	0.057	0.047	0.047	0.049
500	2	8	0.048	0.048	0.045	0.049	0.047	0.045	0.051	0.052	0.048	0.045
500	3	4	0.042	0.039	0.040	0.046	0.048	0.048	0.021	0.042	0.046	0.047
500	3	6	0.043	0.045	0.042	0.042	0.045	0.047	0.047	0.051	0.044	0.045
500	3	8	0.046	0.045	0.040	0.035	0.042	0.047	0.044	0.045	0.050	0.047
500	5	4	0.040	0.036	0.039	0.036	0.041	0.046	0.016	0.048	0.047	0.046
500	5	6	0.041	0.039	0.039	0.039	0.040	0.049	0.046	0.045	0.044	0.044
500	5	8	0.039	0.040	0.036	0.041	0.043	0.050	0.050	0.044	0.046	0.047
1000	2	4	0.042	0.052	0.040	0.055	0.047	0.052	0.046	0.052	0.046	0.048
1000	2	6	0.054	0.052	0.045	0.050	0.045	0.049	0.049	0.054	0.045	0.057
1000	2	8	0.047	0.048	0.048	0.047	0.048	0.052	0.050	0.048	0.055	0.052
1000	3	4	0.049	0.041	0.043	0.045	0.048	0.050	0.054	0.051	0.051	0.047
1000	3	6	0.048	0.044	0.038	0.040	0.050	0.047	0.046	0.049	0.051	0.045
1000	3	8	0.046	0.047	0.047	0.042	0.049	0.045	0.050	0.052	0.043	0.047
1000	5	4	0.038	0.035	0.038	0.047	0.041	0.044	0.050	0.046	0.047	0.048
1000	5	6	0.041	0.043	0.039	0.042	0.043	0.049	0.044	0.048	0.048	0.049
1000	5	8	0.042	0.042	0.038	0.039	0.048	0.050	0.049	0.047	0.045	0.049

Notes: The table shows the empirical rejection frequencies for the $S_{\hat{\gamma}}$ test based on $S = 5,000$ Monte Carlo replications for the baseline model $Y_i = R'\epsilon_i$. The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the model K, the number of B-splines B and the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure E.3.

Table E.3: Rejection Frequencies Alternative Tests for Baseline model

Cat (i)	n	1	2	3	4	5	6	7	8	9	10
$\mathrm{W}^{\mathrm{mle}}$	200	0.179	0.149	0.139	0.127	0.113	0.059	0.097	0.152	0.125	0.171
	500	0.180	0.133	0.114	0.115	0.095	0.167	0.073	0.114	0.097	0.150
	1000	0.188	0.101	0.079	0.074	0.061	0.405	0.058	0.124	0.103	0.170
$\mathrm{LR}^{\mathrm{mle}}$	200	0.028	0.054	0.060	0.046	0.054	0.026	0.048	0.017	0.018	0.024
	500	0.043	0.056	0.068	0.054	0.065	0.023	0.053	0.016	0.017	0.024
	1000	0.049	0.065	0.063	0.061	0.053	0.031	0.051	0.022	0.018	0.025
$\mathrm{W}^{\mathrm{pmle}}$	200	0.375	0.211	0.198	0.086	0.141	0.058	0.105	0.495	0.998	0.467
	500	0.485	0.264	0.204	0.073	0.163	0.030	0.079	0.973	0.999	0.870
	1000	0.570	0.230	0.180	0.051	0.131	0.023	0.068	0.428	1.000	0.947
$\mathrm{LR}^{\mathrm{pmle}}$	200	0.255	0.163	0.133	0.055	0.103	0.035	0.064	0.745	0.997	0.542
	500	0.411	0.229	0.168	0.059	0.136	0.024	0.066	0.982	0.999	0.865
	1000	0.522	0.254	0.170	0.049	0.119	0.022	0.060	0.999	1.000	0.971
$\mathrm{LR}^{\mathrm{gmm}}$	200	0.413	0.411	0.425	0.441	0.290	0.379	0.120	0.216	0.086	0.232
	500	0.292	0.246	0.246	0.286	0.141	0.171	0.025	0.109	0.066	0.106
	1000	0.232	0.181	0.155	0.176	0.074	0.115	0.014	0.068	0.059	0.049

Cat (ii)	n	1	2	3	4	5	6	7	8	9	10
$\hat{S}_{\hat{\gamma}}$	200	0.051	0.047	0.048	0.040	0.049	0.049	0.047	0.048	0.050	0.044
	500	0.047	0.047	0.054	0.047	0.044	0.043	0.047	0.048	0.051	0.054
	1000	0.047	0.043	0.046	0.049	0.048	0.047	0.050	0.044	0.049	0.043
$\mathrm{LM}^{\mathrm{mle}}$	200	0.052	0.058	0.054	0.043	0.040	0.043	0.023	0.018	0.002	0.059
	500	0.056	0.052	0.052	0.042	0.046	0.047	0.028	0.017	0.001	0.062
	1000	0.062	0.052	0.050	0.049	0.039	0.040	0.029	0.016	0.002	0.052
$\mathrm{LM}^{\mathrm{plme}}$	200	0.049	0.045	0.049	0.035	0.038	0.046	0.030	0.041	0.042	0.042
	500	0.049	0.047	0.050	0.039	0.047	0.046	0.034	0.046	0.044	0.051
	1000	0.046	0.048	0.053	0.044	0.041	0.046	0.034	0.042	0.052	0.047
$\mathrm{S}^{\mathrm{gmm}}$	200	0.188	0.209	0.248	0.326	0.236	0.264	0.195	0.108	0.059	0.130
	500	0.094	0.105	0.123	0.223	0.116	0.133	0.103	0.057	0.028	0.064
	1000	0.061	0.070	0.081	0.162	0.069	0.078	0.054	0.031	0.019	0.035

Notes: The table shows the empirical rejection frequencies based on $S = 5,000$ Monte Carlo replications for the baseline model $Y_i = R'\epsilon_i$, with $n = 500$ and $K = 2$. All tests have nominal size $a = 0.05$. The first column indicates the test. The remaining columns denote the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure E.3.

Table E.4: Rejection Frequencies $\hat{S}_{\hat{\gamma}}$ test for LSEM - OLS $\hat{\beta}$

n	K	d	1	2	3	4	5	6	7	8	9	10
200	2	2	0.050	0.054	0.049	0.049	0.038	0.030	0.038	0.043	0.057	0.046
200	2	3	0.049	0.054	0.054	0.048	0.046	0.059	0.042	0.035	0.029	0.052
200	3	2	0.056	0.058	0.050	0.062	0.059	0.031	0.018	0.038	0.047	0.050
200	3	3	0.063	0.054	0.057	0.065	0.060	0.025	0.023	0.051	0.058	0.049
200	5	2	0.098	0.104	0.109	0.142	0.094	0.051	0.064	0.054	0.023	0.057
200	5	3	0.116	0.116	0.131	0.155	0.103	0.039	0.029	0.061	0.026	0.072
500	2	2	0.049	0.050	0.039	0.042	0.041	0.027	0.029	0.036	0.026	0.029
500	2	3	0.048	0.041	0.047	0.047	0.037	0.029	0.024	0.034	0.050	0.051
500	3	2	0.051	0.051	0.048	0.040	0.037	0.028	0.029	0.038	0.022	0.039
500	3	3	0.048	0.050	0.047	0.051	0.053	0.028	0.048	0.041	0.037	0.036
500	5	2	0.071	0.078	0.068	0.081	0.049	0.023	0.060	0.042	0.039	0.038
500	5	3	0.067	0.068	0.080	0.085	0.063	0.022	0.045	0.049	0.027	0.051
1000	2	2	0.040	0.051	0.049	0.029	0.043	0.032	0.033	0.045	0.049	0.041
1000	2	3	0.048	0.044	0.040	0.040	0.040	0.030	0.038	0.046	0.030	0.044
1000	3	2	0.045	0.038	0.043	0.034	0.033	0.032	0.034	0.040	0.039	0.042
1000	3	3	0.044	0.045	0.043	0.036	0.030	0.032	0.035	0.040	0.024	0.034
1000	5	2	0.059	0.051	0.057	0.051	0.039	0.024	0.063	0.030	0.028	0.036
1000	5	3	0.057	0.058	0.056	0.050	0.035	0.018	0.046	0.036	0.029	0.040

Notes: The table shows the empirical rejection frequencies for the $S_{\hat{\gamma}}$ test based on $S = 5,000$ Monte Carlo replications for the linear simultaneous equations model. The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the model K, the number of covariates d and the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure E.3. The $S_{\hat{\gamma}}$ test was implemented using $B = 6$ B-splines.

Table E.5: Rejection Frequencies $\hat{S}_{\hat{\gamma}}$ test for LSEM - One-step $\hat{\beta}$

n	K	d	1	2	3	4	5	6	7	8	9	10
200	2	2	0.067	0.080	0.068	0.081	0.070	0.031	0.054	0.056	0.061	0.051
200	2	3	0.068	0.074	0.076	0.072	0.066	0.071	0.057	0.047	0.026	0.061
200	3	2	0.095	0.106	0.104	0.120	0.090	0.041	0.026	0.059	0.036	0.061
200	3	3	0.099	0.103	0.105	0.114	0.098	0.037	0.028	0.071	0.035	0.064
200	5	2	0.187	0.226	0.247	0.264	0.178	0.063	0.040	0.072	0.020	0.068
200	5	3	0.212	0.238	0.262	0.289	0.193	0.064	0.049	0.089	0.036	0.088
500	2	2	0.062	0.062	0.068	0.067	0.057	0.034	0.049	0.041	0.021	0.037
500	2	3	0.059	0.064	0.071	0.069	0.056	0.031	0.019	0.046	0.031	0.051
500	3	2	0.078	0.078	0.081	0.079	0.066	0.026	0.024	0.047	0.021	0.045
500	3	3	0.076	0.081	0.091	0.088	0.068	0.025	0.029	0.050	0.042	0.042
500	5	2	0.112	0.149	0.158	0.181	0.097	0.036	0.035	0.060	0.030	0.044
500	5	3	0.129	0.151	0.168	0.180	0.101	0.033	0.023	0.069	0.031	0.058
1000	2	2	0.059	0.059	0.065	0.048	0.049	0.025	0.021	0.055	0.050	0.038
1000	2	3	0.060	0.060	0.060	0.068	0.057	0.038	0.052	0.050	0.027	0.051
1000	3	2	0.061	0.067	0.068	0.065	0.053	0.023	0.048	0.047	0.023	0.045
1000	3	3	0.064	0.066	0.072	0.070	0.054	0.040	0.016	0.047	0.022	0.041
1000	5	2	0.091	0.105	0.108	0.111	0.069	0.032	0.026	0.042	0.029	0.043
1000	5	3	0.085	0.102	0.120	0.103	0.065	0.026	0.020	0.047	0.026	0.050

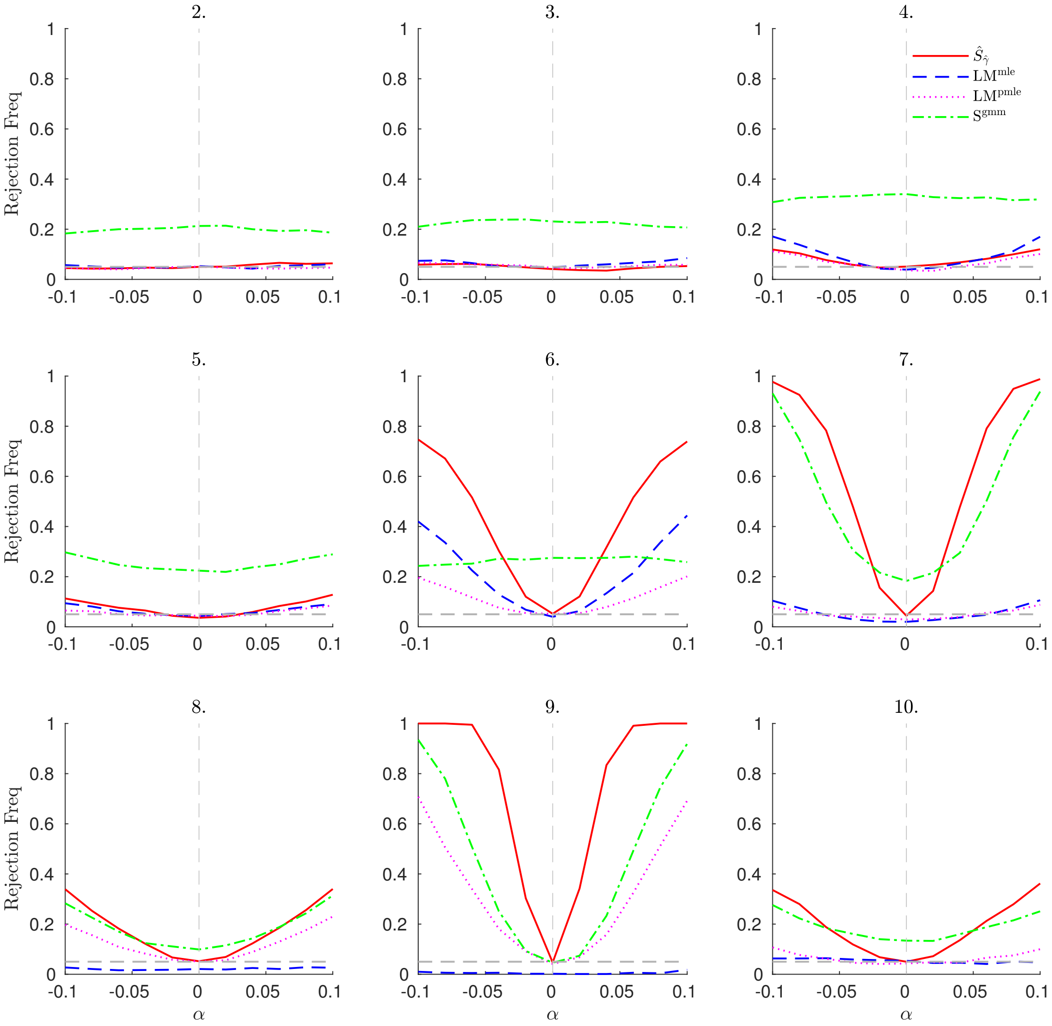
Notes: The table shows the empirical rejection frequencies for the $\hat{S}_{\hat{\gamma}}$ test based on $S = 5,000$ Monte Carlo replications for the linear simultaneous equations model (2.3). The test has nominal size $a = 0.05$. The columns denote the sample size n, the dimension of the observations K, the number of covariates d and the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure E.3. The $S_{\hat{\gamma}}$ test was implemented using $B = 6$ B-splines and using OLS estimates for β.

Table E.6: Production function estimates 2017

	LSEM		OLS
Labor	[0.41, 0.64]	[0.44,0.68]	[0.89, 0.99]
Capital	[0.27, 0.50]	[0.32,0.50]	[0.18, 0.26]
Age		✓	✓
n	1247	1247	1247
p^{ind}	0.12	0.16	

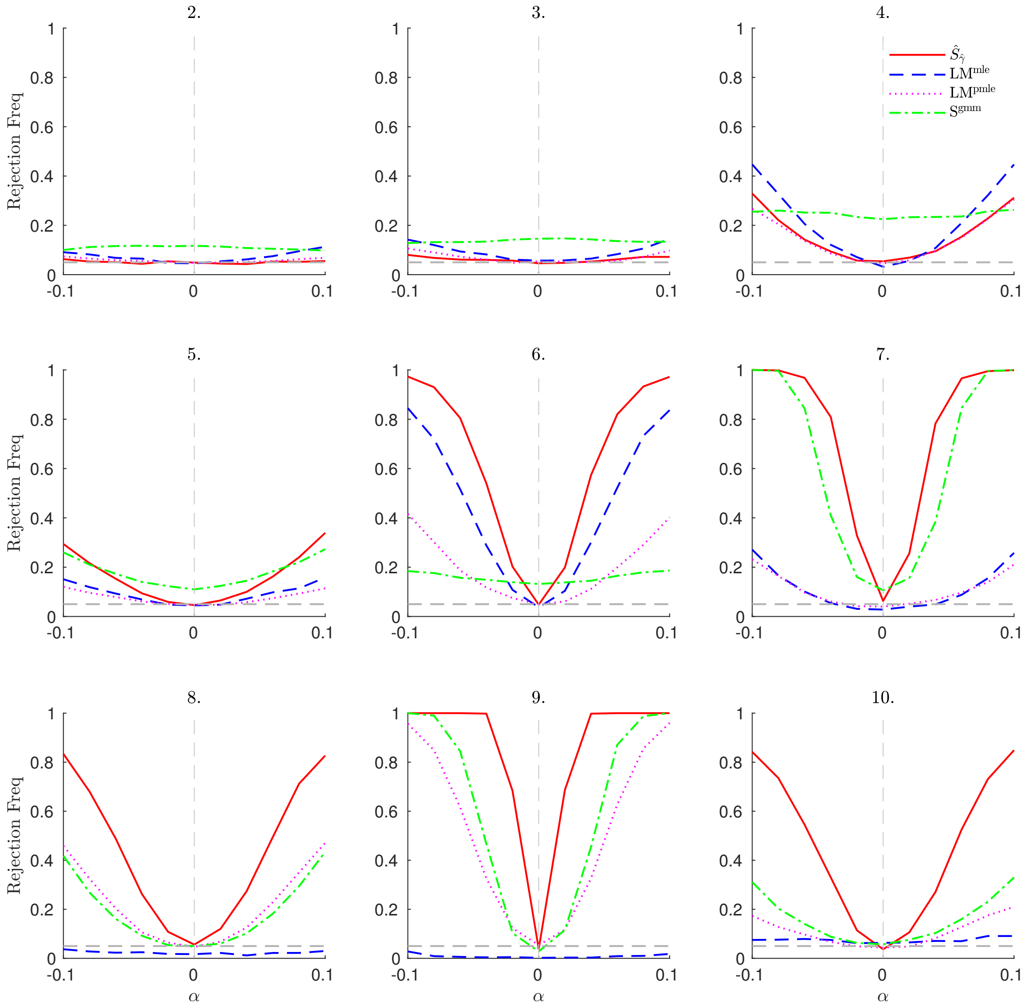
Notes: We report the 95% confidence bands for the production function coefficients for labor and capital. The first three columns consider the bounds obtained by considering the three-variable LSEM (i.e. $Y_i = (\log O_i, \log L_i, \log K_i)'$) with different explanatory variables as indicated in the rows. The right-most column displays the baseline OLS estimates for comparison. The bottom row shows the p-value for the independence test proposed by Matteson and Tsay (2017) as performed on $\{\hat{A}(Y_i - \hat{B}X_i)\}_{i=1}^{n}$, where $\hat{A} = D(\hat{\sigma})^{-1}S(\hat{\alpha}, \hat{\sigma})$, with $\hat{\alpha}$ denoting the minimizer of $\hat{S}_{\hat{\gamma}}$ and $\hat{\sigma}$ and $\hat{B}$ the OLS estimates for σ and B.

Figure F.8: Power Comparison Baseline model $n = 200$



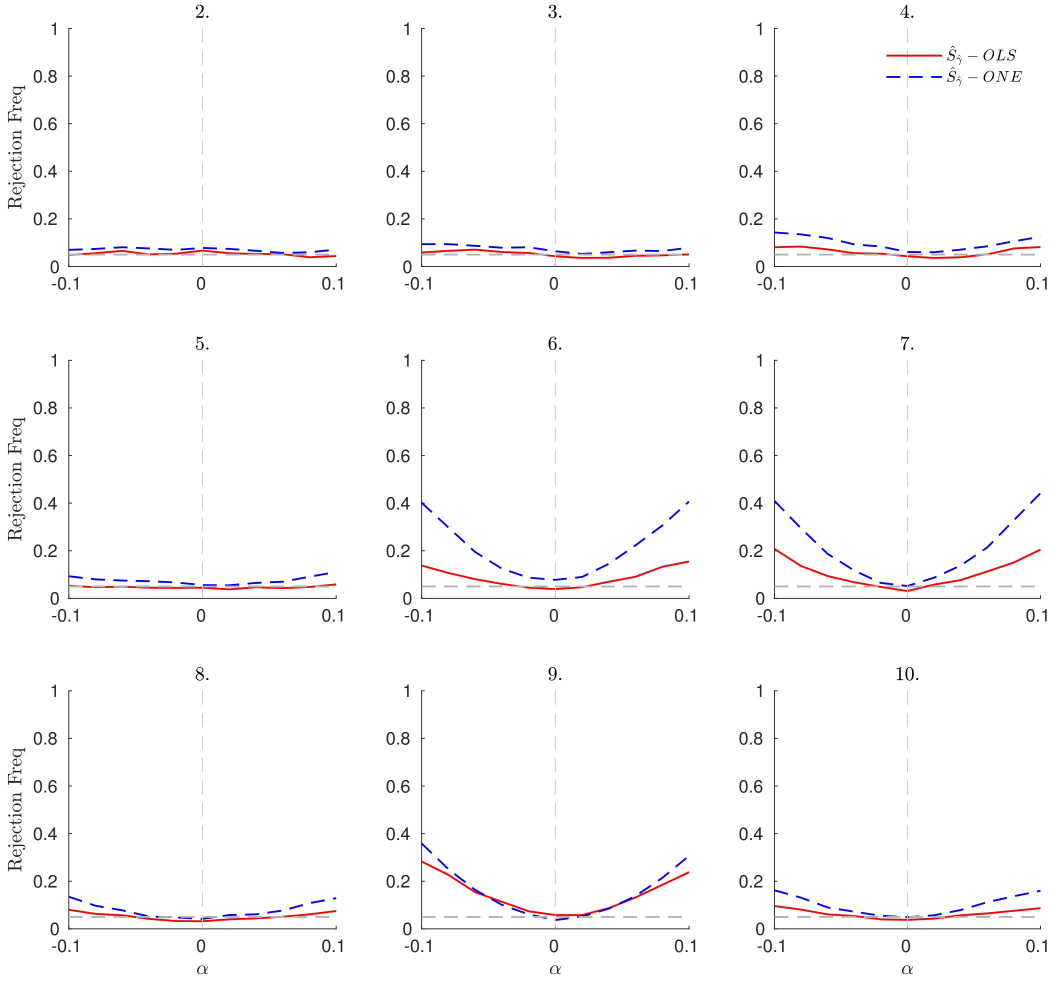
Notes: Empirical power curves for the baseline model with $k = 2$ and $n = 200$. Each plot corresponds to the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities in Figure E.3. The solid red line corresponds to $S_{\hat{\gamma}}$, the dashed blue line to LM$^{\text{mle}}$, the dotted pink line to LM$^{\text{pmle}}$ and the dot-dashed green line to S$^{\text{gmm}}$.

Figure F.9: Power Comparison Baseline model $n = 500$



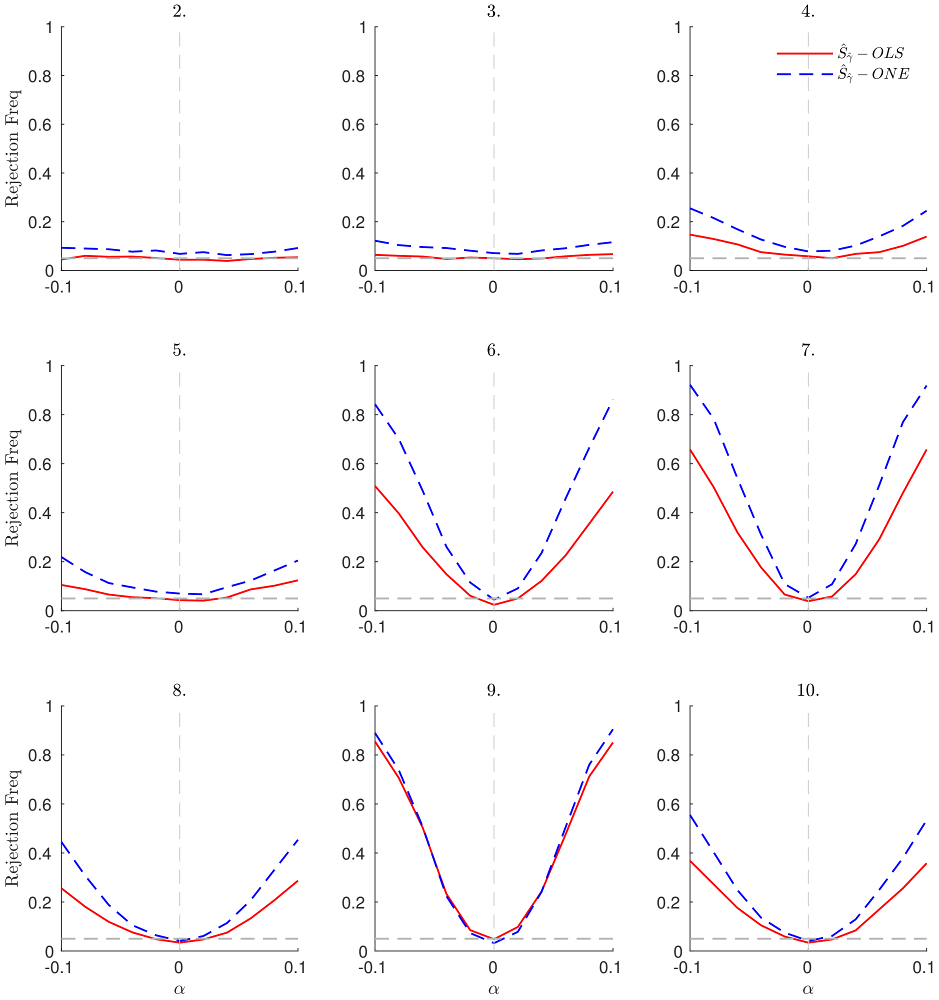
Notes: Empirical power curves for the baseline model with $k = 2$ and $n = 500$. Each plot corresponds to the choice for densities ϵ_k, for $k \geq 2$, where the numbers correspond to the different densities in Figure E.3. The solid red line corresponds to $S_{\hat{\gamma}}$, the dashed blue line to LM$^{\text{mle}}$, the dotted pink line to LM$^{\text{pmle}}$ and the dot-dashed green line to S$^{\text{gmm}}$.

Figure F.10: Power LSEM $n = 200$



Notes: Empirical power curves for the LSEM model with $k = 2$, $d = 2$ and $n = 200$. Each plot corresponds to the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure E.3. The solid red line corresponds to the empirical rejection frequency of the $\hat{S}_{\hat{\gamma}}$ test where $\hat{\gamma} = (\alpha_0, \hat{\beta})$, with $\hat{\beta}$ the OLS estimator. The dashed blue line corresponds to the empirical rejection frequency of the $\hat{S}_{\hat{\gamma}}$ test where $\hat{\gamma} = (\alpha_0, \hat{\beta})$, with $\hat{\beta}$ the one-step efficient MLE estimator.

Figure F.11: Power LSEM $n = 500$



Notes: Empirical power curves for the LSEM model with $k = 2$, $d = 2$ and $n = 500$. Each plot corresponds to the choice for densities $\epsilon_{i,k}$, for $k \geq 2$, where the numbers correspond to the different densities shown in Figure E.3. The solid red line corresponds to the empirical rejection frequency of the $\hat{S}_{\hat{\gamma}}$ test where $\hat{\gamma} = (\alpha_0, \hat{\beta})$, with $\hat{\beta}$ the OLS estimator. The dashed blue line corresponds to the empirical rejection frequency of the $\hat{S}_{\hat{\gamma}}$ test where $\hat{\gamma} = (\alpha_0, \hat{\beta})$, with $\hat{\beta}$ the one-step efficient MLE estimator.

Chapter 3

Robust Inference in Structural VAR Models Identified by Non-Gaussianity

This chapter was co-authored with Lukas Hoesch and Geert Mesters.

3.1. Introduction

In this paper we develop robust inference methods for non-Gaussian structural vector autoregressive (SVAR) models. To outline our contribution, consider the K-dimensional SVAR model

$$Y_t = c + B_1 Y_{t-1} + \cdots + B_p Y_{t-p} + A^{-1}\epsilon_t \ , \tag{3.1}$$

where Y_t is a $K \times 1$ vector of variables, c is an intercept, $B_1, \ldots, B_p$ are the $(K \times K)$ autoregressive matrices, A is the $(K \times K)$ invertible contemporaneous effect matrix and ϵ_t is the $K \times 1$ vector of independent structural shocks with mean zero and unit variance.

It is well known that, without further restrictions, the first and second moments of the process $\{Y_t\}$ are insufficient to identify all parameters in A (e.g. Kilian and Lütkepohl, 2017). Interestingly, non-Gaussian distributions for the structural shocks can be exploited to (locally) identify A. The most well known result follows from the Darmois–Skitovich theorem and is central to the literature on independent components analysis (ICA): if the components of ϵ_t are independent and at least $K - 1$ have a non-Gaussian distribution, A can be recovered up to sign and permutation of its columns (e.g. Comon, 1994). Based on this result several recent works have exploited non-Gaussianity to improve identification and conduct inference in the SVAR model (e.g. Lanne and Lütkepohl, 2010; Moneta et al., 2013; Lanne et al., 2017; Kilian and Lütkepohl, 2017; Maxand, 2018; Lanne and Luoto, 2021; Gouriéroux et al., 2017, 2019; Tank et al., 2019; Herwartz, 2019; Bekaert et al., 2019, 2020; Fiorentini and Sentana, 2022; Braun, 2021; Sims, 2021; Brunnermeier et al., 2021; Drautzburg and Wright, 2021; Davis and Ng, 2022).[1]

Unfortunately, as we show in this paper, standard inference methods for non-Gaussian SVARs are not robust in situations where the densities of the structural errors that generated the data are too "close" to the Gaussian density. Intuitively, what matters for correctly sized inference is not non-Gaussianity per se, but a sufficient distance from the Gaussian distribution. When the true distributions of the structural errors are close to the Gaussian distribution, local identification deteriorates and coverage distortions occur in confidence sets for structural functions (prominent examples include the coverage of structural impulse response functions and forecast error variance decompositions).[2] The problem is somewhat analogous to the weak instruments problem where it is well known that non-zero correlation between the instruments and the endogenous variables is not sufficient to conduct standard inference, but that the correlation must be sufficiently large in order for conventional IV asymptotic theory to be used.[3] Similarly, in our setting, non-Gaussianity alone is not sufficient for standard (pseudo) maximum likelihood or generalised method of moments approaches to yield correct inference when the distance to the Gaussian distribution is not sufficiently large. As such we term this phenomenon "weak non-Gaussianity".

In this paper, we propose a solution to this problem by combining insights from the econometric literature on weak identification robust hypothesis testing as well as the statistical literature on semiparametric modelling. Specifically, we treat the SVAR model with independent structural shocks as a semiparametric model where the densities of the structural errors form the non-parametric part.

For this set-up we provide two main results. First, we use a semi-parametric generalisation of Neyman's C(α) statistic in order to test the possibly weakly identified (or under / unidentified) parameters of the SVAR. More precisely, the semi-parametric score statistic that we employ is based on a quadratic form of the efficient score function, which projects out all scores for the nuisance parameters, including the scores of the density functions, from the conventional score function. This projection, along with the fact that our potentially weakly/non- identified parameter is fixed under the null when conducting the test (as is standard in score-type testing procedures), enables us to circumvent the identification problem and we show that the semi-parametric score test has a χ^2 limit under the null hypothesis. The testing procedure has some semi-parametric efficiency properties when the efficient information matrix is nonsingular (Choi et al., 1996), and for reduced rank information matrices the test remains minimax optimal (Lee, 2022).

[1] See Montiel Olea et al. (2022) for a recent review of this approach.

[2] Simulation studies in, among others, Gouriéroux et al. (2017, 2019) and Lanne and Luoto (2021) have previously highlighted such coverage distortions for parameter estimates in the case of "weakly" non-Gaussian distributions, see also Lee (2022); Lee and Mesters (2022a) for more discussion of the same issue in static ICA models.

[3] See e.g. the recent review by Andrews et al. (2019).

Second, we propose a method for constructing confidence sets for smooth structural functions. Examples include structural impulse responses and forecast error variance decompositions. Specifically, we utilise our proposed test for the weakly identified parameters in a Bonferroni-based procedure (cf. Granziera et al., 2018; Drautzburg and Wright, 2021) which is guaranteed to provide correct coverage level asymptotically, regardless of the level of non-Gaussianity in the errors.

Overall, our method is computationally simple as the implementation of the semiparametric score test only requires estimating regression coefficients, a covariance matrix and the log density scores of the structural errors. For the latter, we use B-spline regressions as developed in Jin (1992) and also considered in Chen and Bickel (2006) for independent component analysis. This approach is computationally convenient and allows our methodology to work under a wide variety of possible distributions for the structural errors.[4]

We assess the finite-sample performance of our method in a large simulation study and find that the empirical rejection frequencies of the semi-parametric score test are always close to the nominal size. This is in contrast to existing methods that are not robust to weak non-Gaussianity, which show substantial size distortions for some of the non-Gaussian distributions considered. We also analyze power of the proposed procedure and find that the power of the semi-parametric score test generally exceeds that of competitor methods which have been proposed in the literature.

Finally, in our empirical studies we revisit two canonical SVAR models and ask whether non-Gaussian distributions can help to identify the structural impulse responses of interest. Specifically, we revisit (i) the labour supply-demand model of Baumeister and Hamilton (2015) and (ii) the oil price model of Kilian and Murphy (2012).[5]

Our findings are mixed. In both applications we find that whilst non-Gaussianity does provide some identifying information, it is unable to pin down all parameter values and/or impulse responses precisely. In the first application, this is in contrast to the findings of Lanne and Luoto (2021) who utilise a GMM approach based on high-order moments to estimate the structural parameters and generally obtain much tighter confidence sets than the weak – identification robust confidence sets we obtain using the methodology of this paper. This is suggestive of the importance of using such robust confidence sets when assessing uncertainty around parameter estimates obtained using non-Gaussianity as an identifying assumption.

This paper relates to several strands of literature. First and foremost, the paper contributes to the literature that exploits non-Gaussianity of the structural shocks for identification (see

[4] The general approach is applicable with other choices of density score estimators provided they satisfy a particular high-level condition. Cf. Lemma A.1.

[5] The assumption of independence among the structural shocks is maintained throughout this paper. Therefore in each application we test for the existence of independent components following Matteson and Tsay (2017), see also Montiel Olea et al. (2022).

the references above). There are two papers that are specifically related to the current paper.

First, Drautzburg and Wright (2021) similarly observe that modest deviations from the Gaussian distribution may invalidate standard inference methods based on an assumption of non-Gaussianity. To circumvent distortions they exploit higher order moments moment restrictions in combination with the identification robust S-statistic of Stock and Wright (2000) for conducting inference. The benefit of their approach is that it is not necessary to assume full independence of the structural shocks. Instead, typically only the third and fourth order higher cross moments are set to zero, leaving all higher order moments unrestricted. A potential downside of such a robust moment approach is that it requires many finite moments. For instance, when using fourth order restrictions the convergence of the S-statistic requires the existence of at least eight moments. We provide a detailed comparison between the approaches in our simulation study.

Second, this paper builds on Lee and Mesters (2022a) and Lee (2022) who consider a similar score testing approach. The crucial differences are that (a) those papers require that the observations are independent and identically distributed across time and (b) they focus on testing a hypothesis on / constructing a confidence set for a potentially weakly- or unidenified parameter. Relaxing the independence assumption is non-trivial in this context; we show a new (uniform) local asymptotic normality result for semi-parametric SVARs. Further, in the SVAR setting the objects of economic interest, such as IRFs, are typically functions of both well-identified parameters and those which may suffer from identification problems. This paper provides a robust inference procedure for such objects.

Besides the non-Gaussian SVAR literature, we note that our approach is inspired by the identification robust inference literature in econometrics (e.g. Stock and Wright, 2000; Kleibergen, 2005; Andrews and Cheng, 2012). The crucial difference in our setting is that the nuisance parameters which determine identification status are infinite dimensional, i.e. the densities of the structural shocks. Despite this difference, conceptually our approach is similar to the score testing approach developed for weakly identified parametric models in Chamberlain (1986). To handle infinite dimensional nuisance parameters we build on the general statistical theory discussed in Bickel et al. (1998) and van der Vaart (2002). While the majority of the statistical literature focuses on efficient estimation in semi-parametric models, a few papers have contributed to testing in well identified models (e.g. Choi et al., 1996; Bickel et al., 2006). The major difference with our paper is that in our setting, a subset of the parameters of interest are possibly weakly- or un- / under- identified, which violates a key regularity condition assumed in this literature.

The remainder of this paper is organized as follows. In Section 3.2, we briefly illustrate how non-Gaussian distributions can help with identification and how the weak identification problem arises. Section 3.3 casts the SVAR model as a semi-parametric model and Section 3.4 establishes a number of preliminary results. The semi-parametric score testing approach is presented in Section 3.5 and inference for smooth structural functions is covered in Section 3.6. Section 3.7 evaluates the finite-sample performance of the proposed methodology and Section 3.8 discusses the results from the empirical studies. Section 3.9 concludes. All proofs are in the appendix.

3.2. Illustration of non-Gaussianity identification

In this section, we illustrate briefly how non-Gaussian distributed structural shocks can help to identify the parameters of the SVAR model. Furthermore, we provide an intuitive explanation for the weak identification problem that arises when the errors are close to Gaussian.

As an example, consider a bivariate SVAR model as defined in equation (3.1), but assume for simplicity that (i) the number of lags is zero ($p = 0$) and (ii) that the contemporaneous effect matrix A is orthonormal.[6] Under these assumptions, the (2×2) matrix A can be parameterized by a scalar parameter α and the model can be written as follows:

$$Y_t = A^{-1}\epsilon_t \qquad \text{where} \qquad A^{-1} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} .$$

In this model, the parameter of interest is the scalar α that determines the angle of the rotation matrix A. If for example $\alpha = 0$, A equals the identity matrix and each of the structural shocks only impacts its respective component in Y_t. For $0 < \alpha < \pi$, or integer multiples thereof, the off-diagonal elements are non-zero so that the shocks affect all variables, with signs depending on the value of α.

To illustrate how non-Gaussian distributions for ϵ_t may help to identify α, we study the expected log-likelihood $\mathbb{E}\ell_\alpha(Y_t)$ in the model above for different distributions of the structural shocks $\epsilon_{k,t}$. For instance, if $\epsilon_{k,t} \sim N(0,1)$ for all k we have

$$\mathbb{E}\ell_\alpha(Y_t) \propto -\frac{1}{2}\mathbb{E}(A^{-1}\epsilon_t)'A^{-1}\epsilon_t = -1 ,$$

and the log likelihood takes the same value for all α. This reflects the standard identification problem: without additional identifying restrictions, the impact effects of the structural shocks are not identifiable when the errors are Gaussian.

Figure 3.1 visually illustrates this result and shows how it changes when we move away from the Gaussian distribution. The left panels shows the expected log likelihood as a function of α, whereas the right panels show the contour plots of the log-likelihood together with a red and a blue line indicating the vector Y_t (i.e. a linear combination of the structural errors ϵ_t), corresponding to two different choices for α.

[6] Note, that the assumption of an orthonormal A matrix can be asymptotically justified if the data Y_t is jointly re-scaled to have mean zero and identity variance matrix (pre-whitening). For details, see Gouriéroux et al. (2017).

We find that as we move away from the Gaussian distribution local identification for α occurs, i.e. the expected gradient of $\ell_\alpha(Y_t)$ with respect to α becomes non-zero in the vicinity of the true parameter (here set as $\alpha = \pi$). Equivalently, in the contour plots the red and blue lines reach different level curves. This means that different choices of α lead to different values of the log-likelihood and hence, α is identifiable. Further we immediately see that only local identification occurs as the same level curves are reached in each quadrant of the contour plot, with each quadrant corresponding to a permutation and/or sign change of the columns. These examples illustrate how non-Gaussianity of the structural errors can help to identify parameters up to permutation and sign changes of the columns.

The problem of weak non-Gaussianity arises when the distance from the Gaussian distribution is not very large. In such scenarios, changes in α only imply minor changes in the level of the likelihood, so that the likelihood ends up being rather flat around the true parameter α. Compare for instance, the panels corresponding to the $t(5)$ density and the $t(15)$ density. In the case of the $t(5)$ density, the red and blue vectors end on clearly distinguishable contour lines of the log-likelihood and the value of the log likelihood varies substantially around $\alpha = \pi$. In contrast, for the $t(15)$ density, the differences are small and the red and blue vectors almost reach the same contour line. In the extreme case of no identification (i.e. the Gaussian case in the upper panels of figure 3.1) we find that α is completely unidentified.

Whilst in population we will always be able to locally identify α when the densities of the structural shocks differ from Gaussian, in the finite sample setting, if the densities of the structural errors are close to Gaussianity, the available identifying information may be small relative to sampling variability. This creates a problem when standard test statistics are used as, in such a setting, standard asymptotics provide a poor approximation to the finite sample behaviour of test statistics.

To remedy this problem we will develop a robust semi-parametric score test for constructing confidence bands for α, based on a test statistic which retains an accurate asymptotic approximation when the structural errors are close to Gaussianity.

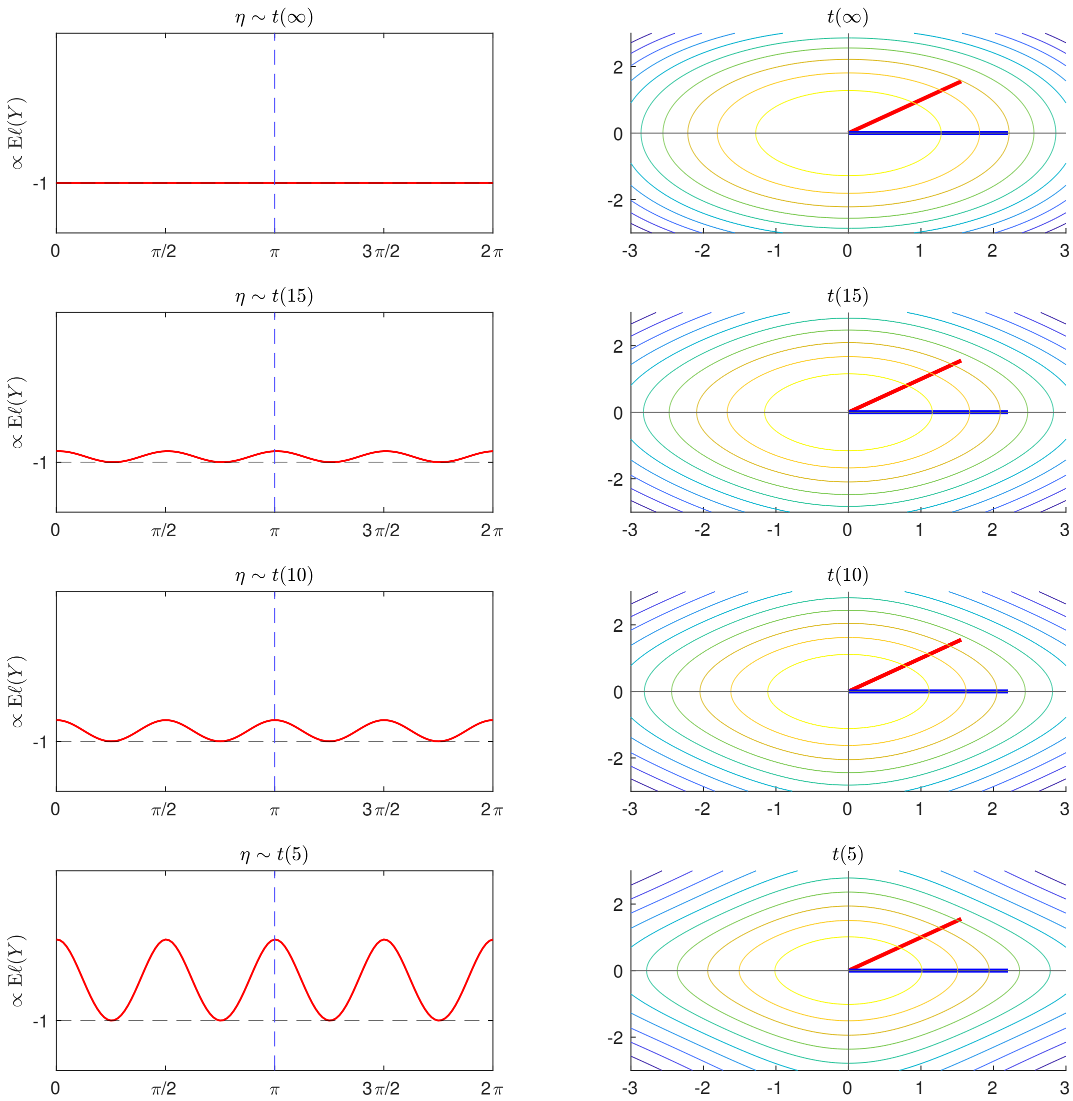
3.3. Semi-parametric SVAR model

In this section, we cast the SVAR model as a semi-parametric model and impose some low-level assumptions which will be maintained throughout. For convenience, we adopt the following notation

$$Y_t = BX_t + A^{-1}(\alpha, \sigma)\epsilon_t \, , \qquad t \in \mathbb{Z} \, , \tag{3.2}$$

where $X_t := (\iota_K', Y_{t-1}', \dots, Y_{t-p}')'$, $B := (c, B_1, \dots, B_p)$ and $A(\alpha, \sigma)$ is a $K \times K$ invertible matrix that is parametrized by the vectors α and σ.

Figure 3.1: Identification with Non-Gaussian Distributions



Notes: The figure shows the log-likelihood contours of Y_t in the SVAR(0) model with scalar parameter α for different distributions of the structural shocks, $\epsilon_{k,t}$. The red and blue lines in each plot denote the vector Y_t corresponding to two different choices for α.

We will leave the choice for the specific parametrization of $A(\alpha, \sigma)$ largely open to the researcher. The key restriction is that σ should be recoverable from the variance of $Y_t - BX_t$, whereas α may be unidentified depending on the distribution of the structural shocks. A canonical choice in this context sets $A^{-1}(\alpha, \sigma) = \Sigma^{1/2}(\sigma)R(\alpha)$, where $\Sigma^{1/2}(\sigma)$ is a lower triangular matrix with positive diagonal elements defined by the vector σ and $R(\alpha)$ is a rotation matrix that is parametrized by the vector α. That said, different parametrizations are often used in practice (cf the empirical section 3.8) and our general formulation allows for such alternatives.

We let $\eta = (\eta_1, \dots, \eta_K)$ correspond to the density functions of $\epsilon_t = (\epsilon_{1,t}, \dots, \epsilon_{K,t})'$ and summarize the parameters in

$$\theta = (\gamma, \eta) , \qquad \gamma = (\alpha, \beta) , \qquad \beta = (\sigma, b) , \tag{3.3}$$

where $b = \mathrm{vec}(B)$. It is clear that the model is semi-parametric with γ the parametric part and η the non-parametric part.

Let $Y^n = (Y_1, \dots, Y_n)'$ and let P_θ^n denote the distribution of Y^n conditional on the initial values $(Y_{1-p}, \dots, Y_0)$. Throughout we work with these conditional distributions; see Hallin and Werker (1999) for a similar setup. For a sample of size n, our semi-parametric SVAR model is the collection

$$\mathcal{P}_\Theta^n = \{P_\theta^n : \theta \in \Theta\} \qquad \Theta = \underbrace{\mathcal{A} \times \mathcal{B}}_{\Gamma} \times \mathcal{H} , \tag{3.4}$$

where $\Gamma \subset \mathbb{R}^L$, with $L = L_\alpha + L_\sigma + L_b$ corresponding to the dimensions of (α, σ, b), and $\mathcal{H} \subset \prod_{k=1}^K \mathsf{H}$ with

$$\mathsf{H} := \left\{ g \in L_1(\lambda) \cap \mathcal{C}^1(\lambda) : \; g(z) \geq 0, \int g(z) \, \mathrm{d}z = 1, \int z g(z) dz = 0, \int \kappa(z) g(z) \, \mathrm{d}z = 0 \right\} ,$$

where λ denotes Lebesgue measure on $\mathbb{R}$, $\mathcal{C}^1(\lambda)$ is the class of real functions on $\mathbb{R}$ which are continuously differentiable λ-a.e. and $\kappa(z) = z^2 - 1$. The parameter space for the densities η_k is restricted such that $\epsilon_{k,t}$ has mean zero and variance one. Further restrictions will be placed on the parameter spaces in the next subsection.

Assumptions

Having defined the semi-parametric SVAR model, we now proceed to formulate the required assumptions that will be maintained throughout the paper. Broadly speaking, we have two types of assumptions: (i) the main assumptions that allow us to establish the properties of the semi-parametric score test and (ii) an additional assumption that defines a set of regularity conditions under which the log density scores of the structural shocks can be consistently estimated. The latter assumption can be appropriately replaced whenever a different density score estimator is used.

Our main assumption is stated as follows.

Assumption 3.3.1. *For model* (3.2)*, we assume that*

1. *For all $\beta \in \mathcal{B}$, $|I_K - \sum_{j=1}^{p} B_j z^j| \neq 0$ for all $|z| \leq 1$ with $z \in \mathbb{C}$*

2. *Conditional on the initial values $(Y'_{-p+1}, \dots, Y'_0)'$, $\epsilon_t = (\epsilon_{1,t}, \dots, \epsilon_{K,t})'$ is independently and identically distributed across t, with independent components $\epsilon_{k,t}$. Each $\eta = (\eta_1, \dots, \eta_K) \in \mathcal{H}$ is such that each η_k is nowhere vanishing, dominated by Lebesgue measure on $\mathbb{R}$, continuously differentiable with log density scores denoted by $\phi_k(z) := \partial \log \eta_k(z)/\partial z$, and for all $k = 1, \dots, K$*

   a) *$\mathbb{E}\epsilon_{k,t} = 0$, $\mathbb{E}\epsilon_{k,t}^2 = 1$, $\mathbb{E}\epsilon_{k,t}^{4+\delta} < \infty$, $\mathbb{E}(\epsilon_{k,t}^4) - 1 > \mathbb{E}(\epsilon_{k,t}^3)^2$, and $\mathbb{E}\phi_k^{4+\delta}(\epsilon_{k,t}) < \infty$ (for some $\delta > 0$);*

   b) *$\mathbb{E}\phi_k(\epsilon_{k,t}) = 0$, $\mathbb{E}\phi_k^2(\epsilon_{k,t}) > 0$, $\mathbb{E}\phi_k(\epsilon_{k,t})\epsilon_{k,t} = -1$, $\mathbb{E}\phi_k(\epsilon_{k,t})\epsilon_{k,t}^2 = 0$ and $\mathbb{E}\phi_k(\epsilon_{k,t})\epsilon_{k,t}^3 = -3$;*

3. *For all $(\alpha, \beta) \in \Gamma$ we have that*

   a) *$A(\alpha, \sigma)$ is positive definite*

   b) *$(\alpha, \sigma) \to A(\alpha, \sigma)$ is continuously differentiable*

   c) *$(\alpha, \sigma) \to [D_{\alpha_l}(\alpha, \sigma)]_{k\bullet} A(\alpha, \sigma)^{-1}_{\bullet j}$ and $(\alpha, \sigma) \to [D_{\sigma_m}(\alpha, \sigma)]_{k\bullet} A(\alpha, \sigma)^{-1}_{\bullet j}$ are Lipschitz continuous for all $l = 1, \dots, L_\alpha$, $m = 1, \dots, L_\sigma$ and $j, k = 1, \dots, K$, where the notation $B_{\bullet j}$ or $B_{j\bullet}$ denotes the jth column or row of a matrix B.*

Part (i) imposes that (3.2) admits a stationary and causal solution. Part (ii) imposes that the densities of the errors are continuously differentiable and certain moment conditions hold. Specifically, part (a) normalises the errors to have mean zero, variance one and finite four+δ moments.[7] Additionally, we require the log density scores $\phi_k(x) = \partial \log \eta_k(x)/\partial x$ evaluated at the errors to have finite $4 + \delta$ moments. Part (b) simplifies the construction of the efficient score functions. Whilst this may at first glance appear a strong condition, lemma S12 in Lee and Mesters (2022a) shows that if the first part holds, then a simple sufficient condition is that the tails of the densities η_k converge to zero at a polynomial rate.[8] The final part (iii) of the assumption imposes that $A(\alpha, \sigma)$ is invertible and that the parametrisation chosen by the researcher is sufficiently smooth. For instance, for $A^{-1}(\alpha, \sigma) = \Sigma^{1/2}(\sigma)R(\alpha)$ when we model $R(\alpha)$ by the Cayley or trigonometric transformation parts (b) and (c) can easily be verified to hold. Verifying these conditions is generally straightforward.

Assumption 3.3.1 ensures that (1) a (uniform) local asymptotic normality [(U)LAN] result holds and (2) the efficient score function can be derived analytically.[9]

[7] $\mathbb{E}(\epsilon_{k,t}^4) - 1 \geq \mathbb{E}(\epsilon_{k,t}^3)^2$ always holds; this is known as Pearson's inequality. See e.g. result 1 in Sen (2012). Assuming that $\mathbb{E}(\epsilon_{k,t}^4) - 1 > \mathbb{E}(\epsilon_{k,t}^3)^2$ rules out (only) cases where $1, \epsilon_{k,t}$ and $\epsilon_{k,t}^2$ are linearly dependent when considered as elements of L_2. See e.g. Theorem 7.2.10 in Horn and Johnson (2013).

[8] Alternatively, these moment conditions hold if one can interchange integration and differentiation appropriately.

[9] See e.g. Le Cam and Yang (2000); van der Vaart (1998); Bickel et al. (1998) for general discussions on uniform local asymptotic normality.

Next, we impose a number of smoothness conditions on the densities η_k. These assumptions facilitate the estimation of the log density scores $\phi_k(z) = \nabla_z \log \eta_k(z)$, which are an important ingredient for the efficient score test.

Assumption 3.3.2. *Let $\phi_{k,n} := \phi_k \mathbf{1}_{[\Xi^L_{k,n}, \Xi^U_{k,n}]}$, $\Delta_{k,n} := \Xi^U_{k,n} - \Xi^L_{k,n}$ and $\nu_n = \nu^2_{n,p}$ with $1 < p \leq 1 + \delta/4$ and $n^{-1/2(1-1/p)} = o(\nu_{n,p})$. Suppose that for $[\Xi^L_{k,n}, \Xi^U_{k,n}] \uparrow \tilde{\Xi} \supset \operatorname{supp}(\eta_k)$ and $\delta_{k,n} \downarrow 0$ it holds that*

(I) *$P(\epsilon_k \notin [\Xi^L_{k,n}, \Xi^U_{k,n}]) = o(\nu^2_n)$;*

(II) *For some $\iota > 0$, $n^{-1}\Delta^{2+2\iota}_{k,n}\delta^{-(8+2\iota)}_{k,n} = o(\nu_n)$;*

(III) *η_k is bounded ($\|\eta_k\|_\infty < \infty$) and differentiable, with a bounded derivative: $\|\eta'_k\|_\infty < \infty$;*

(IV) *For each n, $\phi_{k,n}$ is three-times continuously differentiable on $[\Xi^L_{k,n}, \Xi^U_{k,n}]$ and $\|\phi^{(3)}_{k,n}\|^2_\infty \delta^6_{k,n} = o(\nu_n)$;*[10]

(V) *There are $c > 0$ and $N \in \mathbb{N}$ such that for $n \geq N$ we have $\inf_{t \in [\Xi^L_{k,n}, \Xi^U_{k,n}]} |\eta_k(t)| \geq c\delta_{k,n}$.*

These assumptions are similar to those considered in Chen and Bickel (2006) and Lee and Mesters (2022a). They ensure that the log density scores can be estimated sufficiently accurately using B-spline regressions (for details see section 3.5).[11] Formally, we require that the support of the density η_k is contained with high probability in the interval $[\Xi^L_{k,n}, \Xi^U_{k,n}]$. These lower and upper points will correspond to the smallest and largest knots of the B-splines. Second, condition (ii) ensures that the number of knots does not increase too fast, relative to the sample size n. Conditions (iii) and (iv) impose that the density is sufficiently smooth, such that it can be well-fitted by B-splines. The final condition restricts the tails of the density.

3.4. Preliminary results for semi-parametric SVARs

In this section we present two preliminary results for semi-parametric SVAR models that we believe are more broadly useful. First, we provide a (uniform) local asymptotic normality result for the semi-parametric SVAR model.[12] The primary difference with existing results is that we explicitly perturb the non-parametric part of the model, i.e. the densities η_k, whereas existing (U)LAN results for VARs hold this fixed (e.g. Hallin and Saidi, 2007).

[10] The differentiability and continuity requirements at the end-points are one-sided.

[11] These assumptions are tailored to the specific density score estimator we propose in this paper. Nevertheless, in principle, other density score estimators may be used. Inspection of the proofs reveals that any such estimator which satisfies the conclusions of Lemma A.1 can be adopted.

[12] The uniformity here is over the finite dimensional parameters, γ. For our results in the present paper we only require uniformity over α, but the additional uniformity over β follows at essentially no additional cost.

This extension is necessary for deriving the form of the score test proposed in this paper and can be used in other applications. Second, we analytically derive the efficient score function for the semi-parametric SVAR model. This derivation is more standard as similar efficient score functions are derived for ICA models in Amari and Cardoso (1997); Chen and Bickel (2006) and for linear simultaneous equations models in Lee and Mesters (2022a). Readers who are mainly interested in the practical implementation of the methodology for the semi-parametric SVAR model can safely skip this section.

3.4.1. Uniform Local Asymptotic Normality

Let $(\gamma_n)_{n\in\mathbb{N}} \subset \Gamma$ be such that $\gamma_n \to \gamma \in \Gamma$, fix $\eta \in \mathcal{H}$ and put $\theta := (\gamma, \eta)$. Let G_k denote the law on $\mathbb{R}$ corresponding to η_k $(k = 1, \ldots, K)$ and define

$$\dot{\mathscr{H}}_k := \left\{ h_k \in C_b^1(\lambda) : \int h_k \, \mathrm{d}G_k = \int h_k \iota \, \mathrm{d}G_k = \int h_k \kappa \, \mathrm{d}G_k = 0 \right\}, \quad \dot{\mathscr{H}} := \prod_{k=1}^{K} \dot{\mathscr{H}}_k, \tag{3.5}$$

where ι is the identity funcion, $\kappa(z) = z^2 - 1$ (as defined above) and $C_b^1(\lambda)$ denotes the class of real functions on $\mathbb{R}$ which are bounded, continuously differentiable and have bounded derivatives λ-a.e.. Note that $\mathbb{R}^{L_\alpha + L_\beta} \times \dot{\mathscr{H}}$ is a linear subspace of $\mathbb{R}^{L_\alpha + L_\beta} \times L_\infty(\lambda)^K$. We make this into a normed space by equiping it with the norm $\|(c, h)\| := \|c\|_2 + \sum_{k=1}^{K} \|h_k\|_{\lambda,\infty}$ where $\|\cdot\|_2$ denotes the Euclidean norm.

For an arbitrary sequence $(c_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{L_\alpha + L_\beta}$ such that $c_n := (a_n', d_n')' \to (a', d')' =: c$ let $\tilde{\gamma}_n := \gamma_n + c_n/\sqrt{n}$ and for an arbitrary $(h_n)_{n\in\mathbb{N}} \subset \dot{\mathscr{H}}$ with $h_n \to h \in \dot{\mathscr{H}}$ let $\tilde{\eta}_n := \eta(1 + h_n/\sqrt{n})$. Collect these parameters into $\theta_n := (\gamma_n, \eta)$ and $\tilde{\theta}_n := (\tilde{\gamma}_n, \eta_n)$ respectively. Denote by p_θ^n the density of P_θ^n with respect to λ^n and $\Lambda^n_{\tilde{\theta}_n/\theta_n}$ the (conditional) log-likelihood ratio

$$\Lambda^n_{\tilde{\theta}_n/\theta_n} := \log\left(\frac{p^n_{\tilde{\theta}_n}}{p^n_{\theta_n}}\right) = \sum_{t=1}^{n} \ell_{\tilde{\theta}_n}(Y_t, X_t) - \ell_{\theta_n}(Y_t, X_t)\,, \tag{3.6}$$

where $\ell_\theta(Y_t, X_t)$ denotes the tth contribution to the conditional log likelihood for the SVAR model evaluated at θ. We note that this can be explicitly written as

$$\ell_\theta(Y_t, X_t) = \log|\det(A(\alpha, \sigma))| + \sum_{k=1}^{K} \eta_k(A_{k\bullet}(\alpha, \sigma)(Y_t - BX_t))\,.$$

With this notation established we first derive the scores for the finite dimensional parameters

$\gamma = (\alpha, \sigma, b)$. The score functions with respect to the components α_l, σ_l and b_l are given by

$$\dot{\ell}_{\theta,\alpha_l}(Y_t, X_t) = \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta^{\alpha}_{l,k,j} \phi_k(A_{k\bullet} V_{\theta,t}) A_{j\bullet} V_{\theta,t} + \sum_{k=1}^{K} \zeta_{l,k,k} \left(\phi_k(A_{k\bullet} V_{\theta,t}) A_{k\bullet} V_{\theta,t} + 1\right), \tag{3.7}$$

$$\dot{\ell}_{\theta,\sigma_l}(Y_t, X_t) = \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \zeta^{\alpha}_{l,k,j} \phi_k(A_{k\bullet} V_{\theta,t}) A_{j\bullet} V_{\theta,t} + \sum_{k=1}^{K} \zeta_{l,k,k} \left(\phi_k(A_{k\bullet} V_{\theta,t}) A_{k\bullet} V_{\theta,t} + 1\right), \tag{3.8}$$

$$\dot{\ell}_{\theta,b_l}(Y_t, X_t) = \sum_{k=1}^{K} \phi_k(A_{k\bullet} V_{\theta,t}) \times \left[-A_{k\bullet} D_{b_l} X_t\right], \tag{3.9}$$

where $V_{\theta,t} := Y_t - BX_t$, $A := A(\alpha, \sigma)$, $D_{\alpha_l}(\alpha, \sigma) := \nabla_{\alpha_l} A(\alpha, \sigma)$, $D_{\sigma_l}(\alpha, \sigma) := \nabla_{\sigma_l} A(\alpha, \sigma)$, $D_{b_l} = \nabla_{b_l} B$, $\zeta^{\alpha}_{l,k,j} := [D_{\alpha_l}(\alpha, \sigma)]_{k\bullet} A^{-1}_{\bullet j}$, $\zeta^{\sigma}_{l,k,j} := [D_{\sigma_l}(\alpha, \sigma)]_{k\bullet} A^{-1}_{\bullet j}$ and $\phi_k(z) := \nabla_z \log \eta_k(z)$.

We collect these scores in the vector

$$\dot{\ell}_\theta(Y_t, X_t) := \left(\left(\dot{\ell}_{\theta,\alpha_l}(Y_t, X_t)\right)_{l=1}^{L_\alpha}, \left(\dot{\ell}_{\theta,\sigma_l}(Y_t, X_t)\right)_{l=1}^{L_\sigma}, \left(\dot{\ell}_{\theta,b_l}(Y_t, X_t)\right)_{l=1}^{L_b}, \right)'.$$

Under assumption 3.3.1, we have the following ULAN result.[13]

Proposition 3.4.1 (ULAN). *Suppose that assumption 3.3.1 holds. Then as $n \to \infty$,*

$$\Lambda^n_{\tilde{\theta}_n/\theta_n}(Y^n) = \mathrm{g}_n(Y^n) - \frac{1}{2}\mathbb{E}\left[\mathrm{g}_n(Y^n)^2\right] + o_{P^n_{\theta_n}}(1), \tag{3.10}$$

where the expectation is taken under $P^n_{\theta_n}$ and

$$\mathrm{g}_n(Y^n) := \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left[c' \dot{\ell}_{\theta_n}(Y_t, X_t) + \sum_{k=1}^{K} h_k(A_{k\bullet} V_{\theta_n,t}) \right].$$

Moreover, under $P^n_{\theta_n}$,

$$\mathrm{g}_n(Y^n) \rightsquigarrow \mathcal{N}(0, \Psi_\theta(c,h)), \qquad \Psi_\theta(c,h) := \lim_{n\to\infty} \mathbb{E}\left[\mathrm{g}_n(Y^n)^2\right].$$

The following corollary follows from Le Cam's first Lemma (e.g. van der Vaart, 1998, Example 6.5).

Corollary 3.4.2. *If assumption 3.3.1 holds, then the sequences $(P^n_{\theta_n})_{n\in\mathbb{N}}$ and $(P^n_{\tilde{\theta}_n})_{n\in\mathbb{N}}$ are mutually contiguous.*

[13] The proof is based on verifying the conditions of Theorem 2.1.2 in Taniguchi and Kakizawa (2000), which is due to Swensen (1985, Lemma 1).

The importance of this result is that the semi-parametric SVAR model can be locally asymptotically approximated by a Gaussian shift experiment, uniformly in γ. This local approximation can be exploited to derive the form of the score test below as well as its limiting distribution under local alternatives, but can be more broadly used for other inference problems. One example is a setting where α is well identified, say by assuming non-Gaussian structural shocks, then the (U)LAN result may be used to obtain semi-parametrically efficient parameter estimates similarly to as was done in Chen and Bickel (2006) for the ICA model.

3.4.2. Efficient score function

One of the key ingredients in our framework is the efficient score function for the parameter of interest, α. Loosely speaking this is defined as the projection of the score function for α on the orthogonal complement of the space spanned by the score functions for the nuisance parameters (β, η) (e.g. Bickel et al., 1998; van der Vaart, 2002; Newey, 1990; Choi et al., 1996).

In the case of interest here, where the nuisance parameter contains both finite (β) and infinite-dimensional (η) components, this can be calculated in two steps: (1) compute the projection of the score for $\gamma = (\alpha, \beta)$ on the orthocomplement of the space spanned by the score functions for η. (2) Partition the resulting object into the components corresponding to α and β and project the former onto the orthocomplement of the latter.

We will proceed according to this two-step procedure and now establish the form of the first projection.

Lemma 3.4.3. *Given Assumption 3.3.1 the efficient score function for γ in the semi-parametric SVAR model $\mathcal{P}_\Theta^n$ at any $\theta = (\gamma, \eta)$ with $\gamma = (\alpha, \beta)$, $\alpha \in \mathcal{A}$, $\beta = (\sigma, b) \in \mathcal{B}$ and $\eta \in \mathcal{H}$ is given by $\tilde{\ell}_{n,\theta}(y^n) = \sum_{t=1}^n \tilde{\ell}_\theta(y_t, x_t)$, where*

$$\tilde{\ell}_\theta(y_t, x_t) = \left(\left(\tilde{\ell}_{\theta,\alpha_l}(y_t, x_t) \right)_{l=1}^{L_\alpha}, \left(\tilde{\ell}_{\theta,\sigma_l}(y_t, x_t) \right)_{l=1}^{L_\sigma}, \left(\tilde{\ell}_{\theta,b_l}(y_t, x_t) \right)_{l=1}^{L_b} \right)'$$

with components

$$\begin{aligned}
\tilde{\ell}_{\theta,\alpha_l}(y_t, x_t) &= \sum_{k=1}^K \sum_{j=1, j\neq k}^K \zeta_{l,k,j}^\alpha \phi_k(A_{k\bullet} v_t) A_{j\bullet} v_t + \sum_{k=1}^K \zeta_{l,k,k}^\alpha \left[\tau_{k,1} A_{k\bullet} v_t + \tau_{k,2} \kappa(A_{k\bullet} v_t) \right] \\
\tilde{\ell}_{\theta,\sigma_l}(y_t, x_t) &= \sum_{k=1}^K \sum_{j=1, j\neq k}^K \zeta_{l,k,j}^\sigma \phi_k(A_{k\bullet} v_t) A_{j\bullet} v_t + \sum_{k=1}^K \zeta_{l,k,k}^\sigma \left[\tau_{k,1} A_{k\bullet} v_t + \tau_{k,2} \kappa(A_{k\bullet} v_t) \right] \\
\tilde{\ell}_{\theta,b_l}(y_t, x_t) &= \sum_{k=1}^K -A_{k\bullet} D_{b_l} \left[(x_t - \mu) \phi_k(A_{k\bullet} v_t) - \mu(\varsigma_{k,1} A_{k\bullet} v_t + \varsigma_{k,2} \kappa(A_{k\bullet} v_t)) \right]
\end{aligned}$$

where $v_t = y_t - Bx_t$*,* $\zeta^{\alpha}_{l,k,j} := [D_{\alpha_l}(\alpha,\sigma)]_{k\bullet}A^{-1}_{\bullet j}$ *with* $D_{\alpha_l}(\alpha,\sigma) := \partial A(\alpha,\sigma)/\partial\alpha_l$*,* $\zeta^{\sigma}_{l,k,j} := [D_{\sigma_l}(\alpha,\sigma)]_{k\bullet}A^{-1}_{\bullet j}$ *with* $D_{\sigma_l}(\alpha,\sigma) := \partial A(\alpha,\sigma)/\partial\sigma_l$*,* $D_{b_l} := \partial B/\partial b_l$*,* $\mu := \mathrm{vec}(\iota_K, (\iota_p \otimes (I_K - B_1 - \ldots - B_p)^{-1}c))$*, and* $\tau_k := (\tau_{1,k}, \tau_{2,k})'$ *and* $\varsigma_k := (\varsigma_{1,k}, \varsigma_{2,k})'$ *are defined as*

$$\tau_k := M_k^{-1}\begin{pmatrix}0\\-2\end{pmatrix}, \quad \varsigma_k := M_k^{-1}\begin{pmatrix}1\\0\end{pmatrix} \quad \textit{where } M_k := \begin{pmatrix}1 & \mathbb{E}_\theta(A_{k\bullet}v_t)^3\\ \mathbb{E}_\theta(A_{k\bullet}v_t)^3 & \mathbb{E}_\theta(A_{k\bullet}v_t)^4 - 1\end{pmatrix}.$$

The derivation of the efficient scores $\tilde{\ell}_\theta(y_t, x_t)$ follows along the same lines as in Amari and Cardoso (1997); Chen and Bickel (2006). The dependence on η comes through the log density scores $\phi_k(z) = \nabla_z \log\eta_k(z)$, for $k = 1, \ldots, K$. All other components are simple functions of the parameters and the moments of of the structural shocks as defined by M_k.

For future reference, we partition

$$\tilde{\ell}_\theta(y_t, x_t) = \begin{pmatrix}\tilde{\ell}_{\theta,\alpha}(y_t, x_t)\\ \tilde{\ell}_{\theta,\beta}(y_t, x_t)\end{pmatrix},$$

where $\tilde{\ell}_{\theta,\alpha}(y_t, x_t) = (\tilde{\ell}_{\theta,\alpha_l}(y_t, x_t))_{l=1}^{L_\alpha}$ and $\tilde{\ell}_{\theta,\beta}(y_t, x_t) = \left((\tilde{\ell}_{\theta,\sigma_l}(y_t, x_t))_{l=1}^{L_\sigma}, (\tilde{\ell}_{\theta,b_l}(y_t, x_t))_{l=1}^{L_b}\right)'$.

Based on the efficient scores, we define the efficient information matrix for γ by

$$\tilde{I}_{n,\theta} := \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\tilde{\ell}_\theta(Y_t, X_t)\tilde{\ell}'_\theta(Y_t, X_t) \qquad \text{with partitioning} \quad \tilde{I}_{n,\theta} = \begin{pmatrix}\tilde{I}_{n,\theta,\alpha\alpha} & \tilde{I}_{n,\theta,\alpha\beta}\\ \tilde{I}_{n,\theta,\beta\alpha} & \tilde{I}_{n,\theta,\beta\beta}\end{pmatrix}. \tag{3.11}$$

With Lemma 3.4.3 and the efficient information matrix in place, we can define the efficient score function for α with respect to β and η. In particular this score can be computed by the second projection (e.g. Bickel et al., 1998, p. 74)

$$\tilde{\kappa}_{n,\theta}(y_t, x_t) := \tilde{\ell}_{\theta,\alpha}(y_t, x_t) - \tilde{I}_{n,\theta,\alpha\beta}\tilde{I}^{-1}_{n,\theta,\beta\beta}\tilde{\ell}_{\theta,\beta}(y_t, x_t), \tag{3.12}$$

which exists as long as $\tilde{I}_{\theta,\beta\beta}$ is positive definite. The corresponding efficient information matrix is given by

$$\tilde{\mathcal{I}}_{n,\theta} := \tilde{I}_{n,\theta,\alpha\alpha} - \tilde{I}_{n,\theta,\alpha\beta}\tilde{I}^{-1}_{n,\theta,\beta\beta}\tilde{I}_{n,\theta,\beta\alpha}. \tag{3.13}$$

We note that the efficient score function $\tilde{\kappa}_\theta(y_t, x_t)$ and the efficient information matrix $\tilde{\mathcal{I}}_{n,\theta}$ can be evaluated at any parameters $\theta = (\alpha, \beta, \eta)$ and variables (y_t, x_t).

Building tests or estimators based on the efficient score function is attractive as efficiency results are well established, see Choi et al. (1996), Bickel et al. (1998) and van der Vaart (2002). A crucial difference in our setting is that the efficient information matrix might be singular. For instance, if more than one component of ϵ_t follows an exact Gaussian distribution, $\tilde{\mathcal{I}}_{n,\theta}$ is singular, see Lemma S11 in Lee and Mesters (2022a). The singularity plays an important role in the construction of the semi-parametric score statistic below.

3.5. Inference for potentially non-identified parameters

In this section we consider conducting inference on α without assuming that α is identified, i.e. without assuming that at most one component of ϵ_t has a Gaussian distribution. To do so, we consider testing hypotheses of the following form.

$$H_0 : \alpha = \alpha_0 \ , \ \beta \in \mathcal{B} \ , \ \eta \in \mathcal{H} \qquad \text{against} \qquad H_1 : \alpha \neq \alpha_0 \ , \ \beta \in \mathcal{B} \ , \ \eta \in \mathcal{H} \ . \tag{3.14}$$

The main idea is to consider test statistics whose computation does not require evaluation under the alternative H_1, thus avoiding the need to estimate α. Clearly, based on the trinity of classical tests, the score test is the only viable candidate and we will proceed by constructing score tests in the spirit of Neyman-Rao, but adapted for the semi-parametric setting (e.g. Choi et al., 1996). Such test statistics can then be inverted to yield a confidence region for α with correct coverage. This confidence region then form the basis for constructing confidence intervals for the structural impulse responses as we show in the next section.

In our setting, we rely on the efficient score functions for the SVAR model to construct test statistics. The functional form of the efficient scores $\tilde{\ell}_\theta(y_t, x_t)$ was analytically derived in Lemma 3.4.3. These scores can be estimated by replacing the population quantities by their sample equivalents. We have

$$\hat{\ell}_\gamma(Y_t, X_t) = \left(\left(\hat{\ell}_{\gamma,\alpha_l}(Y_t, X_t) \right)_{l=1}^{L_\alpha}, \left(\hat{\ell}_{\gamma,\sigma_l}(Y_t, X_t) \right)_{l=1}^{L_\sigma}, \left(\hat{\ell}_{\gamma,b_l}(Y_t, X_t) \right)_{l=1}^{L_b} \right)' \tag{3.15}$$

with components

$$\begin{aligned}
\hat{\ell}_{\gamma,\alpha_l}(Y_t, X_t) &= \sum_{k=1}^{K} \sum_{j=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j} \hat{\phi}_{k,n}(A_{k\bullet}V_t) A_{j\bullet}V_t + \sum_{k=1}^{K} \zeta^{\alpha}_{l,k,k} \left[\hat{\tau}_{k,1} A_{k\bullet}V_t + \hat{\tau}_{k,2} \kappa(A_{k\bullet}V_t) \right] \\
\hat{\ell}_{\gamma,\sigma_l}(Y_t, X_t) &= \sum_{k=1}^{K} \sum_{j=1,j\neq k}^{K} \zeta^{\sigma}_{l,k,j} \hat{\phi}_{k,n}(A_{k\bullet}V_t) A_{j\bullet}V_t + \sum_{k=1}^{K} \zeta^{\sigma}_{l,k,k} \left[\hat{\tau}_{k,1} A_{k\bullet}V_t + \hat{\tau}_{k,2} \kappa(A_{k\bullet}V_t) \right] \\
\hat{\ell}_{\gamma,b_l}(Y_t, X_t) &= \sum_{k=1}^{K} -A_{k\bullet} D_{b_l} \left[(X_t - \bar{X}_n) \hat{\phi}_{k,n}(A_{k\bullet}V_t) - \bar{X}_n (\hat{\zeta}_{k,1} A_{k\bullet}V_t + \hat{\zeta}_{k,2} \kappa(A_{k\bullet}V_t)) \right]
\end{aligned}$$

where $V_t = Y_t - BX_t$ and $\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$. The estimates for the τ_k's and ζ_k's are obtained by replacing the population moments defined in Lemma 3.4.3 by their sample counterparts: $\hat{\tau}_k = \hat{M}_k(0, -2)'$ and $\hat{\zeta}_k = \hat{M}_k(1, 0)'$, where

$$\hat{M}_k := \begin{pmatrix} 1 & \frac{1}{n}\sum_{t=1}^{n}(A_{k\bullet}V_t)^3 \\ \frac{1}{n}\sum_{t=1}^{n}(A_{k\bullet}V_t)^3 & \frac{1}{n}\sum_{t=1}^{n}(A_{k\bullet}V_t)^4 - 1 \end{pmatrix}. \tag{3.16}$$

Finally, the estimates of $\hat{\ell}_\gamma(Y_t, X_t)$ depend on $\hat{\phi}_{k,n}(\cdot)$ which is the estimate for the log density scores $\phi_k(z) = \nabla_z \eta_k(z)$. In practice, we estimate these density scores using B-splines following the methodology of Jin (1992) and Chen and Bickel (2006). To set this up, let $b_{k,n} = (b_{k,n,1}, \ldots, b_{k,n,B_{k,n}})'$ be a collection of $B_{k,n}$ cubic B-splines and let $c_{k,n} = (c_{k,n,1}, \ldots, c_{k,n,B_{k,n}})'$ be their derivatives: $c_{k,n,i}(x) := \frac{\mathrm{d}b_{k,n,i}(x)}{\mathrm{d}x}$ for each $i \in [B_{k,n}]$. The knots of the splines, $\xi_{k,n} = (\xi_{k,n,i})_{i=1}^{K_{k,n}}$ are taken as equally spaced in $[\Xi^L_{k,n}, \Xi^U_{k,n}]$.[14]

Our estimate for the log density score ϕ_k is given by

$$\hat{\phi}_{k,n}(z) := \hat{\gamma}'_{k,n} b_{k,n}(z) \; , \tag{3.17}$$

where

$$\hat{\gamma}_{k,n} := -\left[\frac{1}{n}\sum_{t=1}^{n} b_{k,n}(A_{k\bullet}V_t) b_{k,n}(A_{k\bullet}V_t)'\right]^{-1} \frac{1}{n}\sum_{t=1}^{n} c_{k,n}(A_{k\bullet}V_t) \; . \tag{3.18}$$

It shows that computing the log density score estimate (3.17) only requires computing the B-spline regression coefficients $\hat{\gamma}_{k,n}$ in (3.18).

Having defined all the components of the efficient score estimates we may estimate the efficient information matrix for γ by

$$\hat{I}_{n,\gamma} = \frac{1}{n}\sum_{t=1}^{n} \hat{\ell}_\gamma(Y_t, X_t)\hat{\ell}_\gamma(Y_t, X_t)' \; . \tag{3.19}$$

With the estimates for the efficient scores and information for γ, we can estimate the efficient score and information for α. This amounts to replacing the population score $\tilde{\kappa}_{n,\theta}(y_t, x_t)$ and information $\tilde{\mathcal{I}}_{n,\theta}$ in (3.12) and (3.13) by their sample counterparts. We have that

$$\hat{\kappa}_{n,\gamma}(Y_t, X_t) = \hat{\ell}_{\gamma,\alpha}(Y_t, X_t) - \hat{I}_{n,\gamma,\alpha\beta}\hat{I}^{-1}_{n,\gamma,\beta\beta}\hat{\ell}_{\gamma,\beta}(Y_t, X_t) \tag{3.20}$$

and

$$\hat{\mathcal{I}}_{n,\gamma} = \hat{I}_{n,\gamma,\alpha\alpha} - \hat{I}_{n,\gamma,\alpha\beta}\hat{I}^{-1}_{n,\gamma,\beta\beta}\hat{I}_{n,\gamma,\beta\alpha} \; . \tag{3.21}$$

Since the information matrix may be singular, we need to make an adjustment. Specifically, given the truncation rate ν_n defined in Assumption 3.3.2, we define a truncated eigenvalue version of the information matrix estimate as

$$\hat{\mathcal{I}}^t_{n,\gamma} = \hat{U}_n \hat{\Lambda}_n(\nu_n)\hat{U}'_n \; , \tag{3.22}$$

where $\hat{\Lambda}_n(\nu_n)$ is a diagonal matrix with the ν_n-truncated eigenvalues of $\hat{\mathcal{I}}_{n,\gamma}$ on the main diagonal and $\hat{U}_n$ is the matrix of corresponding orthonormal eigenvectors. To be specific, let $\{\hat{\lambda}_{n,i}\}_{i=1}^{L}$ denote the non-increasing eigenvalues of $\hat{\mathcal{I}}_{n,\gamma}$, then the (i, i)th element of $\hat{\Lambda}_n(\nu_n)$

[14] In practice we take these points as the 95th and 5th percentile of the samples $\{A_{k\bullet}V_t\}_{i=1}^{n}$ adjusted by $\log(\log(n))$, where $A = A(\alpha, \sigma)$ and $V_t = Y_t - BX_t$ for a given parameter choice $\gamma = (\alpha, \beta)$.

is given by $\hat{\lambda}_{n,i}\mathbf{1}(\hat{\lambda}_{n,i} \geq \nu_n)$.

Based on this, we define the semi-parametric score statistic for the SVAR model as follows.

$$\hat{S}_{n,\gamma} := \left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\hat{\kappa}_{n,\gamma}(Y_t, X_t)\right)' \hat{\mathcal{I}}_{n,\gamma}^{t,\dagger} \left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\hat{\kappa}_{n,\gamma}(Y_t, X_t)\right), \tag{3.23}$$

where $\hat{\mathcal{I}}_{n,\gamma}^{t,\dagger}$ is the Moore-Penrose pseudo-inverse of $\hat{\mathcal{I}}_{n,\gamma}^{t}$. We note that the test statistic can be evaluated at any $\gamma = (\alpha, \beta)$. To evaluate the null hypothesis (3.14) we will use $\alpha = \alpha_0$, i.e. fixing the unidentified parameters under the null, and $\hat{\beta}$, some $\sqrt{n}$-consistent estimate for the finite dimensional nuisance parameters.

For such parameter choices, the limiting distribution of $\hat{S}_{n,\gamma}$ is derived in the following theorem.

Theorem 3.5.1. *Let $\gamma_n = (\alpha_n, \beta) \to \gamma$ with each γ_n, γ in Γ and let $\theta_n := (\gamma_n, \eta) \to (\gamma, \eta) = \theta$ for some $\eta \in \mathcal{H}$. Suppose that under $P_{\theta_n}^n$, $\hat{\beta}_n$ is a $\sqrt{n}-$consistent estimator of β. Define $\mathscr{S}_n = n^{-1/2}C\mathbb{Z}^{L_2}$ for some $C > 0$ and let $\bar{\beta}_n$ be a discretized version of $\hat{\beta}_n$ which replaces its value with the closest point in $\mathscr{S}_n$. Define $\bar{\gamma}_n = (\alpha_n, \bar{\beta}_n)$, suppose that assumptions 3.3.1 and 3.3.2 hold. Let $r_n = \operatorname{rank}(\hat{\mathcal{I}}_{n,\bar{\gamma}_n}^{t})$ and denote by c_n the $1-a$ quantile of the $\chi_{r_n}^2$ distribution, for any $a \in (0,1)$. Then if $\tilde{\theta}_n := (\alpha_n, \tilde{\beta}_n, \tilde{\eta}_n)$ where $\sqrt{n}\|\tilde{\beta}_n - \beta\| = O(1)$ and $\tilde{\eta}_n = \eta(1 + h_n/\sqrt{n})$ with h_n in some compact $\dot{\mathscr{H}}_\star \subset \dot{\mathscr{H}}$,*

$$\lim_{n\to\infty} P_{\tilde{\theta}_n}^n(\hat{S}_{n,\bar{\gamma}_n} > c_n) \leq a,$$

with inequality only if $\operatorname{rank}(\tilde{\mathcal{I}}_{\theta_0}) = 0$.

Several comments on this theorem are in order.

We do not impose which estimator $\hat{\beta}_n$ should be adopted as the theorem holds for any $\sqrt{n}$-consistent estimator. However, given that the efficient scores of γ need to be computed anyway, it is attractive to rely on one-step efficient estimates for β as discussed in van der Vaart (1998), as this typically improves the (finite sample) power of the test.[15] That said conventional OLS estimates for the regression coefficients B and the variance parameters σ can also be used.

Second, the score statistic is evaluated at the discretised estimator $\bar{\beta}_n$. This is a technical device due to Le Cam (1960) that allows the proof to go through under weak conditions, see Le Cam and Yang (2000, p. 125) or van der Vaart (1998, pp. 72 – 73) for further discussion. Since the discretisation can be arbitrarily fine, this has no practical implications.[16]

Third, the eigenvalue truncation rate appears to have little effect on the finite sample results. In our simulation studies and empirical applications, we always truncate at machine

[15] See the simulation results of section 3.7.

[16] Indeed, in practice, we always discretise at machine precision.

precision which implies that $\hat{\mathcal{I}}_{n,\gamma}^{t,\dagger}$ is similar to $\hat{\mathcal{I}}_{n,\gamma}^{\dagger}$, the Moore-Penrose inverse of $\hat{\mathcal{I}}_{n,\gamma}$. Experimenting with different, but small, truncation rates appears to matter little in practice.

Fourth, the theorem is proven to hold along sequences of parameter values $\tilde{\theta}_n$. By standard methods one can translate such limit statements along sequences to limit statements that hold uniformly over certain sets. In the present case a uniform statement would hold over, for example, sets of the form $\mathscr{P}_n := \{P^n_{\alpha,\beta+d/\sqrt{n},\eta(1+h/\sqrt{n})} : \alpha \in \mathcal{A}_\star, \|d\| \leq M, h \in \dot{\mathscr{H}_\star}\}$ where $\mathcal{A}_\star \subset \mathcal{A}$, $\dot{\mathscr{H}_\star} \subset \dot{\mathscr{H}}$ are compact and $M \in (0,\infty)$.

Finally, if $\tilde{\mathcal{I}}_\theta$ has full rank, the singularity adjusted score statistic is asymptotically equivalent to its non-singular version that is computed with $\hat{\mathcal{I}}_{n,\bar{\gamma}_n}^{-1}$ instead of $\hat{\mathcal{I}}_{n,\bar{\gamma}_n}^{t,\dagger}$. Given this equivalence, it follows from Choi et al. (1996) that tests based on $\hat{S}_{n,\bar{\gamma}_n}$ are asymptotically uniformly most powerful within the class of rotation invariant tests (when $L = 1$, the rotational invariance can be dropped for one-sided tests and replaced with unbiasedness for two-sided tests). This implies that asymptotically when testing the hypothesis (3.14), the power of the test is the greatest possible in the class of rotationally invariant tests. Lee (2022) shows that in the case where $\tilde{\mathcal{I}}_\theta$ has positive rank, the singularity adjusted score statistic is (locally asymptotically) minimax optimal.[17] These results make tests based on $\hat{S}_{n,\bar{\gamma}_n}$ attractive for scenarios where there is no explicit direction in which one wants to maximize power. When such directions are given, alternative test statistics, also based on the efficient score function, can be considered (e.g. Bickel et al., 2006).

A confidence set for the parameters α can be constructed by inverting the efficient score test $\hat{S}_{n,\gamma}$ over a grid of values for α. Formally, for any $a \in (0,1)$ we define the $1-a$ confidence set estimate for α as

$$\hat{C}_{n,1-a} := \{\alpha \in \mathcal{A} : S_{n,(\alpha,\bar{\beta}_n)} \leq c_{n,\alpha}\}\ ,$$

where $c_{n,\alpha}$ the $1-a$ quantile of the $\chi^2_{r_{n,\alpha}}$ distribution and $r_{n,\alpha} = \operatorname{rank}(\hat{\mathcal{I}}^t_{n,(\alpha,\bar{\beta}_n)})$. The following corollary establishes that the confidence set $\hat{C}_{n,1-a}$ has asymptotically correct coverage.

Corollary 3.5.2. *Let γ_n, θ_n, $\tilde{\theta}_n$, $\hat{\beta}_n$, $\bar{\beta}_n$ and $\bar{\gamma}_n$ be as in Theorem 3.5.1 and suppose that assumptions 3.3.1 and 3.3.2 hold. Then,*

$$\lim_{n\to\infty} P^n_{\theta_n}\left(\alpha_n \in \hat{C}_{n,1-a}\right) \geq 1-a. \tag{3.24}$$

The confidence set $\hat{C}_{n,1-a}$ is the main building block for constructing confidence bands for the structural functions in the next section. In addition, this set can be of interest in its own right as the coefficients α can have a direct structural interpretation, see for instance the labour supply-demand model of Baumeister and Hamilton (2015) that is considered in Section 3.8.

[17] This result is particularly relevant in the setting considered in this paper, since if α is multidimensional, under Gaussianity $\tilde{\mathcal{I}}_\theta$ has positive (but deficient) rank.

We finish this section by briefly summarising the practical implementation for the efficient score test and the construction of the confidence set.

Algorithm 1: Confidence set for α

1. Choose a set $\mathcal{A}$;
2. For each $\alpha \in \mathcal{A}$:

    1 Obtain estimates $\hat{\beta}_n = (\hat{\sigma}_n, \hat{b}_n)$, with $b_n = \mathrm{vec}(B_n)$, and set $\hat{V}_t = Y_t - \hat{B}_n X_t$;

    2 For $k = 1, \ldots, K$, compute the log density scores $\hat{\phi}_k(A(\alpha_0, \hat{\sigma}_n)_{k\bullet}\hat{V}_t)$ from (3.17);

    3 Compute the efficient scores $\hat{\ell}_{\hat{\gamma}_n}(Y_t, X_t)$ from (3.15) and the information matrix $\hat{I}_{n,\hat{\gamma}_n}$ from (3.19) using $\hat{\gamma}_n = (\alpha_0, \hat{\beta}_n)$;

    4 Compute $\hat{\kappa}_{n,\hat{\gamma}_n}(Y_t, X_t)$ and $\hat{\mathcal{I}}_{n,\hat{\gamma}_n}$ from (3.20) and (3.21).

    5 Compute the score statistic $\hat{S}_{n,\hat{\gamma}_n}$ from (3.23) and accept $H_0 : \alpha = \alpha_0$ if $\hat{S}_{n,\hat{\gamma}_n} \leq c_n$, where c_n is the $1 - a$ quantile of the $\chi^2_{r_n}$ distribution with $r_n = \mathrm{rank}(\hat{\mathcal{I}}^t_{n,\hat{\gamma}_n})$.

3. Collect the accepted values for α to form $\hat{C}_{n,1-a}$.

The algorithm highlights that the computation costs for computing the confidence set are modest. In fact, the costs are similar to those for constructing standard weak instrument robust confidence sets, such as those based on the Anderson-Rubin statistic (e.g. Andrews et al., 2019) for instance. The only difference is that we require K regression estimates (to estimate the log density scores) as opposed to one. Further, note that for the implementation we do not need to concern ourselves with explicitly discretising the estimator $\hat{\beta}_n$.

3.6. Robust inference for smooth functions

In this section we discuss the methodology for conducting robust inference on smooth functions of the finite dimensional parameters $\gamma = (\alpha, \beta)$. The main functions of interest are the structural impulse response functions (sIRF), but also forecast error variance decompositions and forecast scenarios can be considered (e.g. Kilian and Lütkepohl, 2017). The difference with the preceding section is that we are now explicitly interested in conducting inference on functions of *both* α and β.

The general function of interest is defined as

$$g(\alpha, \beta) \,:\, D_g \to \mathbb{R}^{d_g} \quad \text{with} \quad D_g \supset \mathcal{A} \times \mathcal{B}\,, \tag{3.25}$$

where D_g is the domain of g and d_g is some integer. The following assumption restricts the class of functions that we consider.

Assumption 3.6.1. *$g : D_g \to \mathbb{R}^{d_g}$ is continuously differentiable with respect to β and the Jacobian matrix $J_\gamma := \nabla_{\beta'} g(\alpha, \beta)$ has full column rank on D_g.*

For concreteness the next example provides the details for a vector of structural impulse response functions.

Example 3. *Consider the vector that collects all sIRF at horizon l*

$$\text{IRF}(l) = g(\alpha, \beta) := \text{vec}\left(D\mathsf{B}(b)^l D' A(\alpha, \sigma)^{-1}\right),$$

where

$$D := \begin{bmatrix} I_K & 0_{K\times K(p-1)} \end{bmatrix}, \qquad \textit{and} \qquad \mathsf{B}(b) := \begin{bmatrix} B_1 & B_2 & \cdots & B_{p-1} & B_p \\ I_K & 0 & \cdots & 0 & 0 \\ 0 & I_K & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_K & 0 \end{bmatrix}.$$

In our general notation we have $d_g = K^2$ and we note that, given Assumption 3.3.1, this function is continuously differentiable with respect to β. The Jacobian $J_\gamma \in \mathbb{R}^{K^2 \times L_\beta}$ has the form $J_\gamma = [J_{\gamma,1}, J_{\gamma,2}]$ where

$$J_{\gamma,1} := \left[(A(\alpha,\sigma)^{-1})' \otimes I_K\right] \left\{ \sum_{j=0}^{h-1} \left[D(\mathsf{B}(b)')^{h-1-j} \otimes (D\mathsf{B}(b)^j D')\right] \right\}$$
$$J_{\gamma,2} := \left[I_K \otimes D\mathsf{B}(b)^h D'\right] \nabla_\sigma \text{vec}(A(\alpha,\sigma)^{-1}).$$

△

Our objective is to construct valid $1-q$ confidence sets for $g(\alpha, \beta)$. Intuitively, we proceed in two steps: first we construct a valid confidence set for α using the methodology of the previous section, and second, for each included α we construct a confidence set for $g(\alpha, \hat{\beta}_n)$. The union over the latter sets provides the final set. Overall, this two-step Bonferroni approach is similar to the approach utilised by Granziera et al. (2018) and Drautzburg and Wright (2021).

Formally, let $q_1, q_2 \in (0,1)$ such that $q_1 + q_2 = q \in (0,1)$. In the first step we construct a $1-q_1$ confidence set $\hat{C}_{n,1-q_1}$ for α using **Algorithm 1**. The asymptotic validity of this set is proven in Corollary 3.5.2. Second, for each $\alpha \in \hat{C}_{n,1-q_1}$ we compute $\hat{\nu}_{\alpha,n} := g(\alpha, \hat{\beta}_n)$. The confidence set for $\hat{\nu}_{\alpha,n}$ is given by

$$\hat{C}_{n,g,\alpha,1-q_2} := \left\{\nu : n(\hat{\nu}_{\alpha,n} - \nu)'\hat{V}_{n,\alpha}^{-1}(\hat{\nu}_{\alpha,n} - \nu) \le c_{q_2}\right\}, \tag{3.26}$$

where $\nu := g(\alpha, \beta)$ and $\hat{V}_{n,\alpha} = J_{\hat{\gamma}} \hat{\Sigma}_n J'_{\hat{\gamma}}$, with $\hat{\gamma} = (\alpha, \hat{\beta}_n)$ and $\hat{\Sigma}_n$ a consistent estimate for the asymptotic variance of $\hat{\beta}_n$. The critical value c_{q_2} corresponds to the $1 - q_2$ quantile of a $\chi^2_{1-q_2}$ random variable. The following proposition establishes the conditions on the estimates $\hat{\beta}_n$ that ensure that the confidence set (3.26) is valid.

Proposition 3.6.1. *Suppose that assumption 3.6.1 holds and let γ_n, θ_n, $\tilde{\theta}_n$ be as in Theorem 3.5.1. Suppose $\hat{\beta}_n$ is a sequence of estimates such that*

$$\sqrt{n}(\hat{\beta}_n - \tilde{\beta}_n) \overset{P^n_{\tilde{\theta}_n}}{\rightsquigarrow} \mathcal{N}(0, \Sigma)\,, \quad \textit{with} \quad \Sigma \succ 0$$

and $\hat{\Sigma}_n$ is a sequence of estimates such that $\hat{\Sigma}_n \xrightarrow{P^n_{\tilde{\theta}_n}} \Sigma$, then the confidence set $\hat{C}_{n,g,\alpha}$ in (3.26) *satisfies*

$$\lim_{n\to\infty} P^n_{\tilde{\theta}_n} \left(g(\alpha_n, \tilde{\beta}_n) \in \hat{C}_{n,g,\alpha_n,1-q_2} \right) = 1 - q_2. \tag{3.27}$$

The proof of this Proposition is a straightforward application of the (uniform) delta method.[18] Under Assumption 3.3.1 both OLS / moment – based and one-step efficient estimates for the parameters β satisfy the required conditions on $\hat{\beta}_n$. Moreover, conventional variance estimators for Σ can be adopted to satisfy the consistency of $\hat{\Sigma}_n$.

The final confidence set, $\hat{C}_{n,g}$ is formed by taking the union of the sets $\hat{C}_{n,g,\alpha,1-q_2}$ over $\alpha \in \hat{C}_{n,1-q_1}$:

$$\hat{C}_{n,g} := \bigcup_{\alpha \in \hat{C}_{n,1-q_1}} \hat{C}_{n,g,\alpha,1-q_2} \tag{3.28}$$

The confidence set $\hat{C}_{n,g}$ is a valid $1 - q$ confidence set as we formally establish in the following proposition.

Proposition 3.6.2. *Let $\tilde{\theta}_n$ be as in Theorem 3.5.1, $\hat{C}_{n,1-q_1}$ satisfies Corollary 3.5.2 and $\hat{C}_{n,g,\alpha_n,1-q_2}$ satisfies proposition 3.6.1, then*

$$\liminf_{n\to\infty} P^n_{\tilde{\theta}_n} \left(g(\alpha_n, \tilde{\beta}_n) \in \hat{C}_{n,g} \right) \geq 1 - q.$$

For convenience we summarise the practical implementation in the following algorithm.

Algorithm 2: Robust confidence sets for smooth functions

1. Obtain the confidence set $\hat{C}_{n,1-q_1}$ for α using **Algorithm 1**;
2. For each $\alpha \in \hat{C}_{n,1-q_1}$
   a) Estimate $\hat{\beta}_n$ and $\hat{\Sigma}_n$;
   b) Compute $\hat{V}_{n,\alpha} = J_{\hat{\gamma}} \hat{\Sigma} J'_{\hat{\gamma}}$ with $J_{\hat{\gamma}}$ and $\hat{\gamma} = (\alpha, \hat{\beta}_n)$

[18] See Theorem A.19 for the statement.

Table 3.1: Distributions for Structural Shocks

Abbreviation	Name	Definition
$\mathcal{N}(0,1)$	Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$
$t(\nu)$, $\nu = 15, 10, 5$	Student's t	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{\left(-\frac{\nu+1}{2}\right)}$
SKU	Skewed Unimodal	$\frac{1}{5}\mathcal{N}\big(0,1\big) + \frac{1}{5}\mathcal{N}\big(\frac{1}{2},(\frac{2}{3})^2\big) + \frac{3}{5}\mathcal{N}\big(\frac{13}{12},(\frac{5}{9})^2\big)$
KU	Kurtotic Unimodal	$\frac{2}{3}\mathcal{N}\big(0,1\big) + \frac{1}{3}\mathcal{N}\big(0,(\frac{1}{10})^2\big)$
BM	Bimodal	$\frac{1}{2}\mathcal{N}\big(-1,(\frac{2}{3})^2\big) + \frac{1}{2}\mathcal{N}\big(1,(\frac{2}{3})^2\big)$
SPB	Separated Bimodal	$\frac{1}{2}\mathcal{N}\big(-\frac{3}{2},(\frac{1}{2})^2\big) + \frac{1}{2}\mathcal{N}\big(\frac{3}{2},(\frac{1}{2})^2\big)$
SKB	Skewed Bimodal	$\frac{3}{4}\mathcal{N}\big(0,1\big) + \frac{1}{4}\mathcal{N}\big(\frac{3}{2},(\frac{1}{3})^2\big)$
TRI	Trimodal	$\frac{9}{20}\mathcal{N}\big(-\frac{6}{5},(\frac{3}{5})^2\big) + \frac{9}{20}\mathcal{N}\big(\frac{6}{5},(\frac{3}{5})^2\big) + \frac{1}{10}\mathcal{N}\big(0,(\frac{1}{4})^2\big)$

Note: The table reports the distributions that are used in the simulation studies in section 3.7 to draw the structural errors. The mixture distributions are taken from Marron and Wand (1992), see their table 1.

c) Construct the confidence set $\hat{C}_{n,g,\alpha,1-q_2}$ as in (3.26);

3. Construct $\hat{C}_{n,g}$ from (3.28).

As is demonstrated in the subsequent section, for structural impulse responses this approach often provides confidence sets with shorter average length when compared to alternative robust confidence set constructions proposed in the literature.

3.7. Finite sample performance

This section presents the results from a collection of simulation studies that are designed to evaluate the size and power of the proposed hypothesis testing procedure for different densities of the structural shocks. We also compare the performance of the test to existing approaches available in the literature. Finally, we evaluate the coverage and length of the confidence intervals for the structural impulse responses.

3.7.1. Size of semi-parametric score test

We start by evaluating the finite sample size of the score test $\hat{S}_{n,\hat{\gamma}_n}$ in the semi-parametric SVAR model. We consider SVAR(p) specifications with $p = 1, 2, 4, 8, 12$ lags, $K = 2, 3$ variables and $T = 200, 500, 1000$. We simulate the SVAR(p) model for ten different choices for the distributions of the structural errors $\epsilon_{k,t}$ with $k = 1, \ldots, K$. The density functions that we consider and their abbreviated names are reported in Table 3.1. We standardised the draws to have mean zero and unit variance.

We parameterize the contemporaneous effect matrix by $A(\alpha, \sigma)^{-1} = \Sigma^{1/2}(\sigma)R(\alpha)$ where $\Sigma^{1/2}(\sigma)$ is lower triangular and the rotation matrix $R(\alpha)$ is parameterized using the trigonometric transformation as in section 3.2. In the bivariate case, $L_\alpha = 1$ and we choose $\alpha_0 = \pi/5$ for the data-generating process. In the trivariate SVAR, $L_\alpha = 3$ and we use $\alpha_0 = (3\pi/5, 2\pi/5, -\pi/5)'$. Furthermore, we choose $\Sigma^{1/2}$ such that the diagonal elements are equal to one, $\sigma_{ii} = 1$ for $i = 1, \ldots, K$, and we set the off-diagonal elements to $\sigma_{ij} = 0.2$ for $i > j$. The SVAR coefficient matrices, $A_1, \ldots, A_p$ are generated as diagonal matrices with diagonal elements drawn from a $\mathcal{N}(0, 1)$ distribution.[19] Importantly, even though the data-generating process assumes diagonal coefficient matrices, the test is carried out treating the coefficient matrices as full $K \times K$ matrices. We use 250 burn-in periods to simulate the SVAR(p) model and use $M = 5,000$ Monte Carlo replications to compute the finite-sample rejection rates of the test procedure.

Tables 3.2-3.3 report the empirical rejection frequencies of the semi-parametric score test defined in Section 3.5 for testing the hypothesis $H_0 : \alpha = \alpha_0$ vs. $H_1 : \alpha \neq \alpha_0$. The test is implemented following steps 1-5 in **Algorithm 1** for $\alpha = \alpha_0$ and using $B = 6$ cubic B-splines for the estimation of the log density scores. Table 3.2 reports the results when estimating the nuisance parameters β using OLS while table 3.3 reports the results from using the one-step efficient estimates for β which update the OLS estimates using one Gauss-Newton iteration. All tests are conducted at 5% nominal size.

Table 3.2 shows the empirical rejection frequencies of the proposed test procedure when using OLS estimates for the nuisance parameters. The results for $T = 200$ are reported in the first panel of the table. For the SVAR(p) with $K = 2$ variables, the size of the test is generally very close to the nominal size of 5$\%$. Importantly, this holds even for densities where we may expect identification failures – even when the shocks are normally distributed and hence α is not identified, the test is correctly sized. The size remains correct for all densities that are close to Gaussian, such as the $t(15)$ and the skewed-unimodal density. For densities that are far from Gaussian such as the seperate bi-modal density, some under-rejection is observed.

As the number of parameters in the SVAR increases with the lag size p or the number of variables K, the rejection rates increase and the test starts to over-reject in small samples. For an increase in the number of lags, rejection rates only increase slightly, but when the number of variables increases, the number of parameters grows quadratically and hence rejection rates show a more substantial increase. Importantly, this holds regardless of the true underlying density considered and is caused by the rather imprecise OLS estimates that are plugged into the score test statistic.

When we increase the sample size ($T = 500$, $T = 1,000$) these size distortions quickly disappear and the rejection frequencies converge to the nominal size of the test. Thus, even

[19] To ensure stationarity of the SVAR(p) model, the coefficient matrices are sampled until inspection of the roots of the corresponding SVAR(1) companion matrix indicates that the SVAR passes the stationarity condition.

in the case of an SVAR with a larger lag length, the testing procedure gives correct inference, as long as the sample size is not too small. We note that we continue to see under-rejection for some of the densities far from Gaussianity.

Table 3.3 reports the empirical rejection frequencies for the same simulations when one-step efficient estimates are used for the nuisance parameters. The one-step efficient estimates of β are obtained by updating the OLS estimates of the nuisance parameters β towards the efficient estimates by one Gauss-Newton iteration. Comparing the rejection rates in table 3.3 with those reported in the case of OLS estimates of the nuisance parameters in table 3.2, shows that using the one-step estimates yields substantial improvements in the size of the test in small samples, especially when the number of lags is large. For example, for the case of an SVAR with three variables and 12 lags, the size of the rejection rates are very close to the nominal size of 5%. As the sample size grows, the difference between the two approaches is less pronounced and the procedures yield comparable rejection rates. For medium and large samples, either of the procedure can result in rejection rates closer to the nominal size, depending on the number of lags, the number of variables and the distribution of the structural errors that generated the data.

We note here that using one-step efficient updates of β also remedies the under-rejection observed for some of the Gaussian mixture distributions in Table 3.2.

Overall, we may conclude that the empirical size of the test is close to the nominal size regardless of the choice for the true densities, i.e. Gaussian, close to Gaussian, or far from Gaussian. Finite sample size distortions can be largely overcome by using one-step efficient estimates.

3.7.2. Comparison to alternative approaches

Next, we compare the performance of the semiparametric score test to a variety of alternative methods that have been proposed in the literature based on size and power. We distinguish between two types of tests: (i) tests that do not fix α under the null (e.g. Wald and Likelihood ratio type tests) and (ii) tests that fix under the null (e.g. score type tests). Clearly, from the discussion in Section 3.2 it follows that we expect the tests in the first category to perform poorly as they are vulnerable to identification failures.[20]

In the first category, we consider three tests: the psuedo maximum likelihood Wald test ($\mathrm{W}^{\mathrm{PML}}$) of Gouriéroux et al. (2017), which we implement using one (standardised) $t(7)$ density and a (standardised) $t(12)$ density for the second shock, as in Gouriéroux et al. (2017). We additionally consider two tests based on the work of Lanne and Luoto (2021): these are the GMM Wald (W^{LL}) and distance metric ($\mathrm{DM}^{\mathrm{LL}}$) tests based on higher (third & fourth) order moment conditions.

[20] Simulation evidence in Lee and Mesters (2022a) has shown that tests that do not fix α under the null often show severe over-rejection in ICA models when the errors are close to Gaussian.

Table 3.2: Empirical rejection frequencies using OLS Estimates

K	p	N(0,1)	t(15)	t(10)	t(5)	SKU	KU	BM	SPB	SKB	TRI
$T = 200$											
2	1	4.56	5.18	4.90	4.74	4.00	4.54	1.64	2.22	4.02	1.98
2	2	4.84	4.86	4.90	5.00	3.94	4.74	2.48	2.50	3.46	1.88
2	4	5.68	5.50	5.16	5.56	4.34	4.58	2.66	3.32	4.64	1.60
2	8	6.24	6.94	6.68	5.82	5.54	5.58	3.42	3.38	5.52	2.20
2	12	7.78	7.42	7.32	7.40	5.64	6.04	4.04	4.18	6.84	3.52
3	1	5.10	5.38	6.32	7.12	5.36	5.98	5.44	4.80	5.46	5.30
3	2	6.36	6.68	7.14	6.96	6.04	4.88	5.72	3.98	5.92	4.36
3	4	8.00	8.44	8.90	9.20	7.18	5.34	5.90	4.10	6.66	4.42
3	8	11.30	12.28	11.72	12.52	8.74	7.32	7.30	4.76	9.88	6.46
3	12	16.32	16.90	17.26	15.28	10.92	11.32	11.06	7.36	13.84	8.28
$T = 500$											
2	1	5.08	4.78	5.30	4.60	3.92	4.42	1.78	1.48	3.08	1.76
2	2	4.86	5.16	4.24	4.02	4.04	4.92	1.96	1.64	3.62	1.66
2	4	5.16	5.02	5.24	4.68	4.24	5.34	2.28	2.04	3.64	1.46
2	8	5.40	5.38	5.02	4.80	5.12	5.70	2.42	3.14	4.16	1.54
2	12	6.50	5.94	5.72	5.34	5.18	7.12	3.04	4.22	4.34	1.82
3	1	4.84	5.84	5.56	6.40	5.12	6.08	4.64	5.18	4.96	5.56
3	2	5.56	5.80	6.40	5.70	6.16	5.02	4.28	5.28	4.80	4.94
3	4	6.14	6.66	6.58	6.72	5.82	4.38	4.66	4.44	4.76	4.40
3	8	7.74	8.06	8.22	8.50	7.68	5.72	5.66	4.32	6.42	4.36
3	12	9.86	9.84	10.04	9.74	8.82	5.86	5.70	4.46	7.66	4.50
$T = 1,000$											
2	1	5.24	4.52	3.90	4.20	3.86	4.36	1.82	1.14	3.00	1.18
2	2	4.38	4.84	4.58	4.18	3.82	4.84	1.48	1.84	2.82	1.68
2	4	5.00	4.92	4.42	4.28	4.08	4.94	2.04	1.88	2.92	1.52
2	8	5.18	4.88	4.84	4.08	4.92	6.58	2.32	2.44	2.78	1.56
2	12	5.68	5.34	5.42	4.46	5.28	7.68	2.76	3.60	3.82	1.74
3	1	4.68	4.88	5.38	5.08	4.94	5.58	5.22	5.04	5.00	5.14
3	2	4.62	4.72	5.76	5.12	5.64	5.76	4.46	5.08	4.26	4.60
3	4	4.92	4.84	5.06	5.72	6.24	5.02	4.66	4.60	4.30	5.08
3	8	7.10	5.92	5.84	6.16	6.54	5.64	5.08	4.66	5.08	4.14
3	12	7.70	7.28	7.62	7.18	7.04	5.36	4.86	4.44	5.42	4.44

Note: The table reports empirical rejection frequencies for the semi-parametric score test of the hypothesis $H_0 : \alpha = \alpha_0$ vs. $H_1 : \alpha \neq \alpha_0$ in the K-variable SVAR(p) model with nominal size 5%. The nuisance parameters β are estimated by OLS. The columns correspond to different choices for the distributions of the structural shocks, $\epsilon_{k,t}$ for $k = 1, \ldots, K$. The distributions are reported in Appendix XX. Rejection rates are computed based on $M = 5,000$ Monte Carlo replications.

Table 3.3: Empirical rejection frequencies using One-step Estimates

K	p	N(0,1)	t(15)	t(10)	t(5)	SKU	KU	BM	SPB	SKB	TRI
$T = 200$											
2	1	5.94	6.26	6.48	5.34	5.46	4.94	4.12	5.28	4.48	4.26
2	2	5.94	5.18	5.68	5.26	5.16	4.20	5.16	5.56	4.12	4.98
2	4	4.86	5.12	4.20	4.34	4.82	3.98	4.48	5.36	4.54	5.36
2	8	4.24	4.30	4.76	4.32	4.46	3.70	5.08	6.42	4.04	5.54
2	12	3.92	3.72	3.52	4.18	4.06	3.58	5.26	6.58	3.94	5.88
3	1	7.36	7.26	7.60	7.50	7.12	6.36	6.60	6.44	5.72	6.78
3	2	7.42	7.50	7.70	7.98	7.44	7.40	6.46	6.38	6.64	6.30
3	4	6.56	8.12	7.70	8.20	6.98	6.20	6.46	6.76	5.86	6.04
3	8	4.26	4.78	4.74	5.60	4.36	4.16	3.34	4.06	3.70	3.92
3	12	2.20	2.40	2.48	2.58	2.80	3.00	2.40	2.58	2.04	2.86
$T = 500$											
2	1	6.64	6.60	6.92	6.26	5.42	4.60	5.58	6.16	4.54	5.62
2	2	6.20	6.72	5.64	5.34	5.54	4.76	6.20	6.18	4.90	5.38
2	4	6.20	6.74	6.26	5.72	5.40	4.40	6.22	6.12	4.94	6.34
2	8	5.58	5.92	5.68	5.76	5.08	4.46	6.10	6.82	4.96	6.34
2	12	6.04	5.48	5.12	5.00	4.70	5.88	6.60	7.70	4.24	7.72
3	1	8.04	8.66	7.68	7.68	5.80	6.30	5.40	6.08	5.64	5.70
3	2	7.74	7.66	8.38	6.94	6.18	6.72	5.56	6.20	5.92	5.90
3	4	7.74	8.24	7.62	7.54	6.60	6.72	6.24	6.78	5.62	6.38
3	8	7.86	7.98	8.96	7.68	6.22	7.12	6.86	8.24	6.56	6.82
3	12	7.86	8.22	7.60	7.36	6.40	6.36	6.86	8.34	6.92	5.98
$T = 1,000$											
2	1	6.94	5.82	5.60	5.96	5.16	4.58	5.76	5.58	4.74	5.12
2	2	5.92	6.12	6.28	6.22	4.94	4.88	5.50	5.44	4.30	5.58
2	4	6.36	6.12	5.80	6.16	4.88	4.56	6.14	6.12	4.06	5.32
2	8	6.08	5.94	6.36	5.70	5.60	5.10	6.24	6.56	4.22	6.14
2	12	6.38	5.72	6.54	5.48	5.92	5.00	6.06	7.58	5.06	7.26
3	1	7.64	7.12	7.38	6.62	5.06	6.02	5.70	5.56	5.94	5.32
3	2	7.80	7.10	7.72	6.54	5.84	5.40	5.64	5.82	5.16	4.64
3	4	7.36	7.20	7.02	7.00	6.14	5.94	6.22	6.24	5.40	5.44
3	8	8.86	8.08	7.10	7.48	5.50	7.06	6.88	6.26	5.98	6.86
3	12	8.34	8.38	8.74	7.58	6.28	7.44	7.76	7.38	6.42	8.34

Note: The table reports empirical rejection frequencies for the semi-parametric score test of the hypothesis $H_0 : \alpha = \alpha_0$ vs. $H_1 : \alpha \neq \alpha_0$ in the K-variable SVAR(p) model with nominal size 5%. The nuisance parameters β are estimated by the one-step efficient procedure. The columns correspond to different choices for the distributions of the structural shocks, $\epsilon_{k,t}$ for $k = 1, \dots, K$. The distributions are reported in Appendix XX. Rejection rates are computed based on $M = 5,000$ Monte Carlo replications.

In the second category we consider four tests. Firstly we have the pseudo maximum likelihood Lagrange Multiplier test ($\mathrm{LM}^{\mathrm{PML}}$) that is based on work of Gouriéroux et al. (2017). This test is based on the score of the pseudo log likelihood which we take, following Gouriéroux et al. (2017), to be the Student's t with degrees of freedom fixed at $\nu = 7$ and $\nu = 12$ for the first and second shocks respectively.[21] Secondly, we consider the LM test corresponding to the GMM setup of Lanne and Luoto (2021) ($\mathrm{LM}^{\mathrm{LL}}$). Lastly, we compare to the recently proposed robust GMM methods of Drautzburg and Wright (2021). We include both tests that they propose. The first is based on the S-statistic of Stock and Wright (2000) which sets the cross third and fourth order moments to zero (S^{DW}). Second, we include their non-parametric test which is based on Hoeffding (1948) and Blum et al. (1961) and sets all higher order cross moments to zero ($\mathrm{BKR}^{\mathrm{DW}}$). The S^{DW} has the benefit that it does not require a full independence assumption, whereas the $\mathrm{BKR}^{\mathrm{DW}}$ test, similarly to our semi-parametric score test, requires full independence of the structural shocks. We implement the S^{DW} and $\mathrm{BKR}^{\mathrm{DW}}$ tests using the bootstrap procedure described in Drautzburg and Wright (2021).

To evaluate the finite-sample performance, we focus on an SVAR(1) model with $K = 2$ variables and a sample size of $T = 500$. We use the same parameterization and parameter values as described in the previous subsection to generate the data. However, different to the previous simulation study evaluating the size of the score test, we report results both for the case where the structural errors $\epsilon_{1,t}, \epsilon_{2,t}$ are identically distributed, but also for the case where the first error is fixed to have a Gaussian distribution while the distribution of the second structural error varies. Note that in the latter case, theoretically non-Gaussianity can still be used to identify the parameters of the SVAR if the second structural error does not follow a Gaussian distribution. However, we suspect identification to be weaker in this case.

Size comparison

Table 3.4 reports the results from the simulation study and compares the size of the alternative testing procedures to the size of the score test. The first panel reports the case where the two structural errors, $\epsilon_{1,t}, \epsilon_{2,t}$ are drawn from the same (non-Gaussian) distribution while the second panel reports the results where $\epsilon_{1,t}$ is fixed to have a Gaussian distribution.

The results reconfirm that the tests based on the efficient score function indeed control the size of the test well. The same is true of the competitor tests in group (ii) for most (but not all) of the shock distributions considered. In contrast the tests in group (i) typically perform poorly, often displaying severe over-rejection.

[21] Note that this test is not actually used in Gouriéroux et al. (2017), but the simulations in Lee and Mesters (2022a) show that it has reliable size for i.i.d. linear simultaneous equations models.

First, note that the rejection rates for the two efficient score tests ($\hat{S}$) in the case of identically distributed shocks are close to the nominal size of 5%, regardless of the distribution of the structural shocks (as in table 3.3). Inspecting the second panel of the table, we note that the performance of the score test does not deteriorate when the first structural error is Gaussian; the rejection rates continue to be close to the nominal size of 5% regardless of the distribution of the second error.

Next consider the LM test based on Gouriéroux et al. (2017) ($\mathrm{LM}^{\mathrm{PML}}$): in the case with one Gaussian density, this test is able to control size for all choices of the second density considered. In the case where both shocks are drawn from the same distribution, this test is able to control size for most of the distributions, however over-rejects somewhat for the BM, SPB and TRI distributions.

The LM test based on Lanne and Luoto (2021) ($\mathrm{LM}^{\mathrm{LL}}$) displays slightly worse performance, with over-rejections for about half of the distributions considered. Interestingly many of these over-rejections occur in the first panel, where we may expect that identification is somewhat stronger. The tests of Drautzburg and Wright (2021) ($\mathrm{GMM}^{\mathrm{DW}}$ and $\mathrm{BKR}^{\mathrm{DW}}$) generally perform well, with the former always controlling size correctly and the latter over-rejecting only in a few cases (e.g. the kurtotic unimodal distribution).

As expected, the tests in group (i) tend to perform very poorly, with the simulation results demonstrating both substantial over-rejection and extremely conservative performance, depending on the test and distribution pair.

To summarise, most of the alternative procedures lead to incorrect inference if the distribution of the structural shocks is not "sufficiently" non-Gaussian. Furthermore, the identity of the best-performing alternative procedure crucially depends on which non-Gaussian distribution generated the data. In contrast, the semi-parametric score test proposed in this paper gives correct inference regardless of the distribution of the structural errors and whether one or both errors are non-Gaussian.

Power comparison

Next, we compare the power of the tests that control the size well. We again focus on an SVAR(1) model with $K = 2$ variables a sample size of $T = 500$, and two independent errors drawn from the same distribution.

Figure 3.2 reports the raw (i.e not size-adjusted) power for the semi-parametric score test using one-step nuisance parameter estimates (red solid line), the semi-parametric score test using OLS nuisance parameter estimates (black sold line), the psuedo maximum likelihood LM test (dot - dashed blue line), the Drautzburg and Wright (2021) GMM test (dotted green line) and the non-parametric Drautzburg and Wright (2021) test (dot - dashed purple line). For the t distributions in the first row of the figure, the best performing test is the

Table 3.4: Empirical rejection frequencies for alternative tests

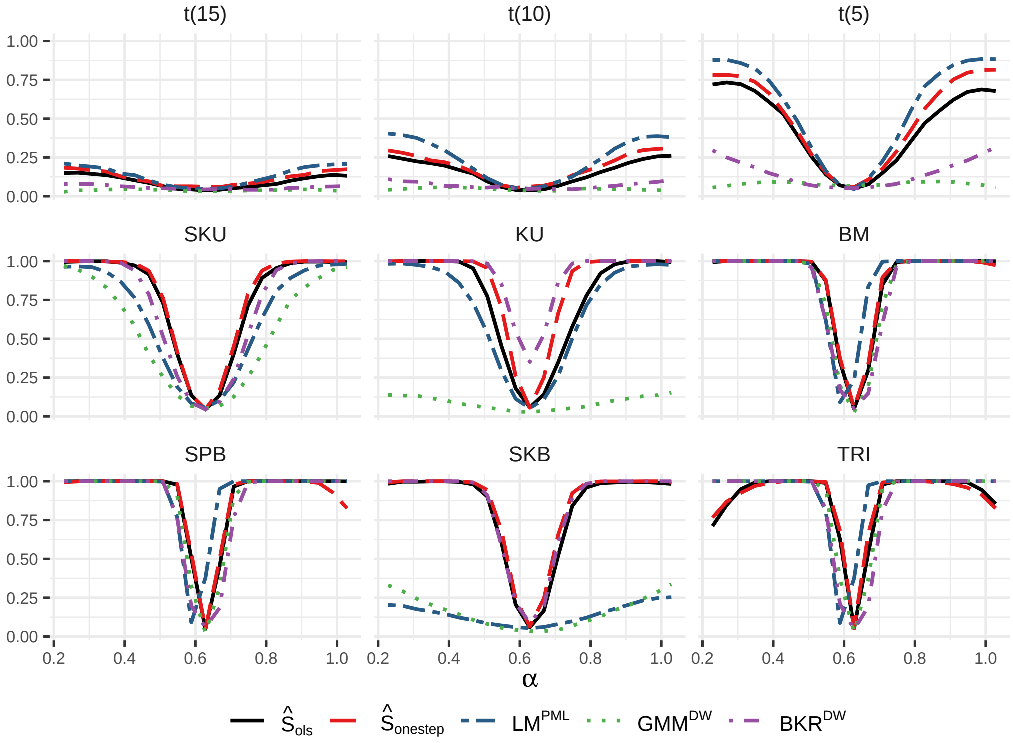
Test	N(0,1)	t(15)	t(10)	t(5)	SKU	KU	BM	SPB	SKB	TRI
$\epsilon_{1,t} \sim \epsilon_{2,t}$										
$\hat{S}_{ols}$	4.56	6.24	4.72	4.56	5.16	5.16	4.28	4.40	4.16	4.56
$\hat{S}_{onestep}$	5.88	7.28	6.28	4.92	5.28	5.20	4.92	4.48	4.64	5.20
$\mathrm{LM}^{\mathrm{PML}}$	4.48	4.84	4.96	4.84	6.36	5.76	20.44	31.68	5.68	32.36
$\mathrm{LM}^{\mathrm{LL}}$	6.04	9.88	13.20	25.88	22.36	14.96	5.64	4.72	11.32	5.28
$\mathrm{GMM}^{\mathrm{DW}}$	3.40	4.04	3.92	5.24	4.88	4.36	3.04	2.36	3.56	2.96
$\mathrm{BKR}^{\mathrm{DW}}$	5.00	4.64	4.00	5.24	6.76	30.56	4.80	4.76	6.44	4.80
$\mathrm{W}^{\mathrm{PML}}$	20.44	3.16	1.60	2.40	3.36	3.32	100.00	100.00	3.12	100.00
W^{LL}	74.96	44.08	22.64	1.00	0.44	2.40	0.00	0.00	50.00	0.00
$\mathrm{DM}^{\mathrm{LL}}$	11.80	12.56	13.60	14.28	11.96	10.68	5.48	4.92	13.72	4.28
$\epsilon_{1,t} \sim \mathcal{N}(0,1)$										
$\hat{S}_{ols}$	5.12	4.52	4.64	4.40	4.16	4.36	1.60	1.12	3.48	1.88
$\hat{S}_{onestep}$	6.72	6.32	6.20	5.76	5.08	4.56	5.04	5.00	5.24	6.00
$\mathrm{LM}^{\mathrm{PML}}$	5.56	6.28	5.68	6.08	9.04	6.80	5.68	6.68	5.04	5.68
$\mathrm{LM}^{\mathrm{LL}}$	7.36	6.12	6.40	6.56	7.12	8.08	12.36	13.60	6.24	12.36
$\mathrm{GMM}^{\mathrm{DW}}$	3.00	3.84	4.36	5.56	3.60	3.20	3.04	4.52	3.32	4.08
$\mathrm{BKR}^{\mathrm{DW}}$	4.52	5.24	5.28	5.88	9.84	49.72	7.56	9.20	13.44	9.32
$\mathrm{W}^{\mathrm{PML}}$	22.20	10.40	7.64	2.04	1.88	1.44	95.08	97.68	11.20	97.92
W^{LL}	74.88	67.40	58.64	24.64	14.80	43.84	56.08	50.88	72.36	54.28
$\mathrm{DM}^{\mathrm{LL}}$	12.04	11.96	11.48	9.08	9.24	11.64	6.20	5.04	12.72	5.20

Note: The table reports empirical rejection frequencies for tests of the hypothesis $H_0 : \alpha = \alpha_0$ vs. $H_1 : \alpha \neq \alpha_0$ with 5% nominal size for the SVAR(1) model with $K = 2$ and $T = 400$, and $\alpha_0 = \pi/5$. $\hat{S}_{ols}$ denotes the semi-parametric score test using OLS estimates for β, $\hat{S}_{onestep}$ uses one-step efficient estimates. $\mathrm{LM}^{\mathrm{LL}}$, W^{LL} and $\mathrm{DM}^{\mathrm{LL}}$ denote the GMM-based LM, Wald and distance metric tests of Lanne and Luoto (2021). $\mathrm{LM}^{\mathrm{PML}}$ and $\mathrm{W}^{\mathrm{PML}}$ denote the pseudo-maximum likelihood LM and Wald tests of Gouriéroux et al. (2017), $\mathrm{GMM}^{\mathrm{DW}}$ denotes the GMM-based test of Drautzburg and Wright (2021), $\mathrm{BKR}^{\mathrm{DW}}$ denotes the non-parametric test of Drautzburg and Wright (2021). The columns correspond to different choices for the distributions of the structural shocks, $\epsilon_{k,t}$ for $k = 1, \ldots, K$. The distributions are reported in Table 3.1. The tests of Drautzburg and Wright (2021) use 500 bootstrap replications to simulate the null distribution of the test statistics. Rejection rates are computed based on $M = 2,500$ Monte Carlo replications.

psuedo maximum likelihood LM test, which is based on a t – density and therefore has the correct shape. Nevertheless, the efficient score tests are not far behind, offering almost comparable power. Moreover, in the other panels, the efficient score tests are typically the most powerful tests (that also control size), with the one-step update version performing slightly better. The quality of the other three tests depends to a large extent on the underlying density. For example, the tests of Drautzburg and Wright (2021) offer very little power in the t-distribution cases, but for the other distributions considered their non-parametric test has power curves which are not much below those of the efficient score test.[22]

[22] For the kurtotic unimodal distribution the power curve of this test is higher, however this test is substantially oversized for this density. It should also be noted that the tests of Drautzburg and Wright (2021) are substantially more computationally demanding than the efficient score based approaches, as they utilise a bootstrap approach to obtain the critical value.

Figure 3.2: Power in the SVAR(1) model



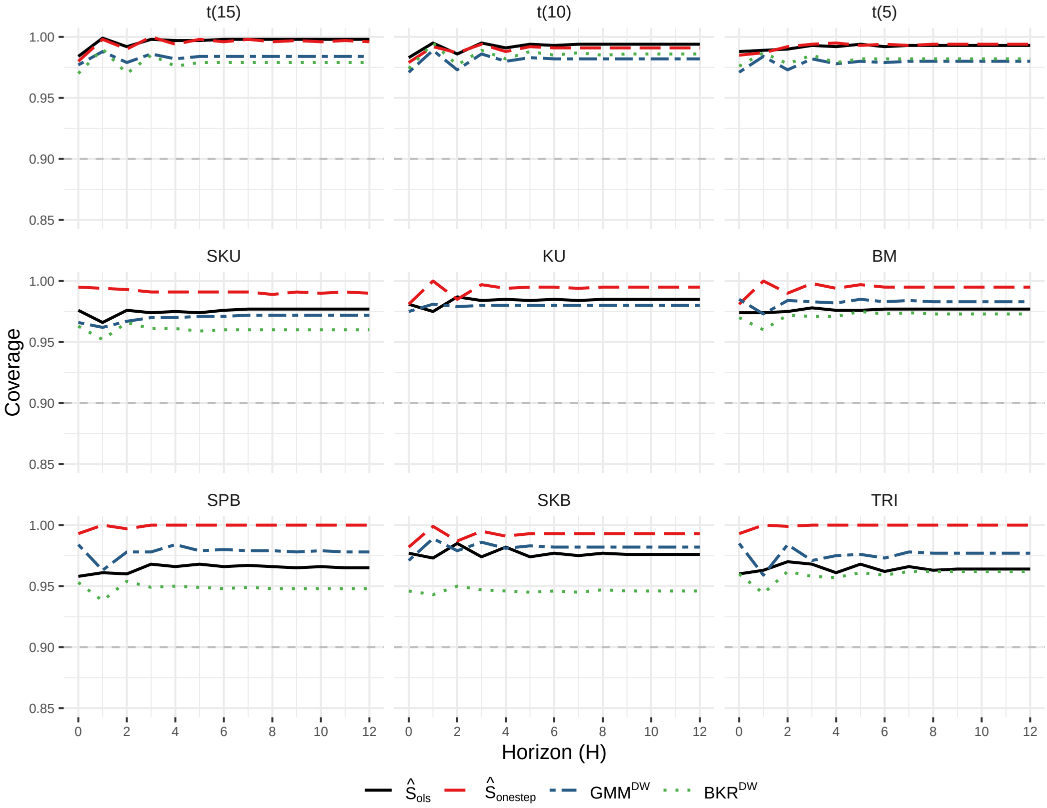
Notes: The figure reports unadjusted empirical power curves for tests of the hypothesis $H_0 : \alpha = \alpha_0$ vs. $H_1 : \alpha \neq \alpha_0$ with 5% nominal size for the SVAR(1) model with $K = 2$ and $T = 500$. The x-axis corresponds to different alternatives for α around $\alpha_0 = \pi/5$. $\hat{S}_{ols}$ denotes the semi-parametric score test using OLS estimates for β, $\hat{S}_{onestep}$ uses one-step efficient estimates. LM^{PML} denotes the pseudo-maximum likelihood test of Gouriéroux et al. (2017), GMM^{DW} denotes the GMM-based test of Drautzburg and Wright (2021), BKR^{DW} denotes the non-parametric test of Drautzburg and Wright (2021). The tests of Drautzburg and Wright (2021) use 500 bootstrap replications to simulate the null distribution of the test statistics. Rejection frequencies are computed using $M = 1,000$ Monte Carlo replications.

Coverage & Average length of confidence sets

Figures 3.3 and 3.4 compare the coverage rate of confidence intervals for the structural impulse response of the first variable to the first shock constructed based on the procedure outlined in Section 3.6 and those of Drautzburg and Wright (2021). These results are based on a SVAR(1) model with $K = 2$, $T = 500$, and two independent errors drawn from the same distribution. In each case, the coverage rate and length are calculated as that of the convex hull of the confidence set proposed in Section 3.6.

Figure 3.3 demonstrates that, as expected, all of these procedures provide correct coverage rates, except for the non-parametric approach of Drautzburg and Wright (2021) with the kurtotic unimodal density. Figure 3.4 demonstrates that whilst the average length of these confidence sets is generally within the same ballpark, there are some differences between the methods depending on the underling density. In general the efficient score based approaches outperform the GMM based approach of Drautzburg and Wright (2021) in all cases except

Figure 3.3: Coverage rates of $\hat{C}_{n,g,\alpha,0.9}$

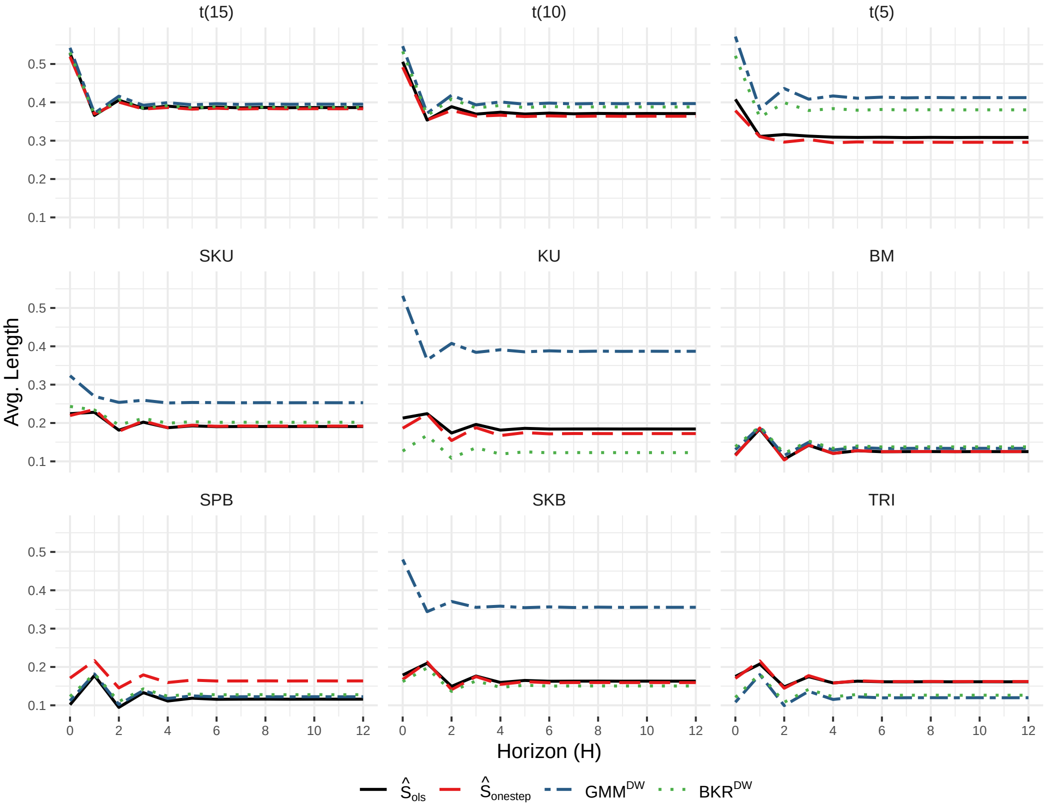


Notes: The figure reports empirical coverage rates of confidence intervals at individual horizons for the impulse response of the first variable to the first shock 90% nominal coverage for the SVAR(1) model with $K = 2$ and $T = 500$. $\hat{S}_{ols}$ denotes the semi-parametric score test using OLS estimates for β, $\hat{S}_{onestep}$ uses one-step efficient estimates. GMM^{DW} denotes the GMM-based test of Drautzburg and Wright (2021) and BKR^{DW} denotes the non-parametric test of Drautzburg and Wright (2021).The tests of Drautzburg and Wright (2021) use 500 bootstrap replications to simulate the null distribution of the test statistics. The tests of Drautzburg and Wright (2021) use 500 bootstrap replications to simulate the null distribution of the test statistics. Average length is computed using $M = 1,000$ Monte Carlo replications.

for the trimodal density (and the separated bimodal case for the one-step efficient version). Against the non-parametric approach of Drautzburg and Wright (2021) the comparison is closer, again with the exception of trimodal and separated bimodal densities.[23] For the t-densities the efficient score approaches attain shorter lengths. There is no substantial difference in these conclusions across horizons in either figure.

[23] Recall that for the kurtotic unimodal density the non-parametric approach from Drautzburg and Wright (2021) does not have correct coverage.

Figure 3.4: Average length of $\hat{C}_{n,g,\alpha,0.9}$



Notes: □ Lukas: TODO: Add description here The figure reports average length of confidence intervals at individual horizons for the impulse response of the first variable to the first shock 90% nominal coverage for the SVAR(1) model with $K = 2$ and $T = 500$. $\hat{S}_{ols}$ denotes the semi-parametric score test using OLS estimates for β, $\hat{S}_{onestep}$ uses one-step efficient estimates. GMM^{DW} denotes the GMM-based test of Drautzburg and Wright (2021) and BKR^{DW} denotes the non-parametric test of Drautzburg and Wright (2021).The tests of Drautzburg and Wright (2021) use 500 bootstrap replications to simulate the null distribution of the test statistics. The tests of Drautzburg and Wright (2021) use 500 bootstrap replications to simulate the null distribution of the test statistics. Average length is computed using $M = 1,000$ Monte Carlo replications.

3.8. Applications

3.8.1. Labor supply/demand model of Baumeister and Hamilton (2015)

Recall the bivariate SVAR(p) model of the U.S. labor market of Baumeister and Hamilton (2015). We have $Y_t = (\Delta w_t, \Delta \eta_t)'$, where Δw_t is the growth rate of real compensation per hour and $\Delta \eta_t$ is the growth rate of total U.S. employment.

$$Y_t = c + B_1 Y_{t-1} + \cdots + B_p Y_{t-p} + B_0^{-1} \Sigma^{1/2}\, \epsilon_t, \qquad B_0 \equiv \begin{bmatrix} -\alpha^d & 1 \\ -\alpha^s & 1 \end{bmatrix},\ \Sigma^{1/2} \equiv \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \tag{3.29}$$

In the model, α^d is the short-run wage elasticity of demand, and α^s is the short-run wage elasticity of supply. The number of lags used in the SVAR is $p = 8$ and sign restrictions imposed on the supply and demand elasticities require that $\alpha^s > 0$ and $\alpha^d < 0$. It is further assumed that $\sigma_1, \sigma_2 > 0$.

In addition to the sign restrictions discussed above, Baumeister and Hamilton (2015)'s identification approach includes carefully motivated priors on the short-run labor supply and demand elasticities, based on estimates from the micro-econometric and macroeconomic literature, as well as a long-run restriction on the effect of labor-demand shocks on employment. The latter restriction was recently criticized by Lanne and Luoto (2021) who revisited the application using a non-Gaussianity identification strategy that is not robust to weak identification.

To address the identification problem, we consider the robust SVAR approach described in this paper that exploits potential non-Gaussianity in the structural shocks to simultaneously test $(\alpha^d, \alpha^s) = (\alpha_0^d, \alpha_0^s)$ for different choices of α_0 to obtain confidence sets for the elasticity parameters as well as confidence bands for the impulse responses to labor supply and labor demand shocks. Specifically, we construct confidence sets for α using Algorithm 1 of Section 3.5 and confidence bands for the impulse responses using Algorithm 2 of Section 3.6. To this end, we estimate $b = (c, \operatorname{vec} {B_1}', \ldots, \operatorname{vec} {B_p}')'$ in (3.29) equation-by-equation by OLS. Given $\hat{b}$, as well as a hypothesised α_0, we can estimate the variance parameters $\sigma = (\sigma_1, \sigma_2)$ in (3.29) by methods of moments, solving the system of equations $\operatorname{vech}\left(B_0(\alpha_0)\hat{\Sigma}_u B_0(\alpha_0)'\right) = \operatorname{vech}\left(\Sigma^{1/2}\Sigma^{1/2'}\right)$ for σ. Confidence bands for the impulse responses are then constructed using the usual asymptotic Delta method approach with a Bonferroni correction, as in section 3.6.[24]

[24] For each α, the estimates of $\Sigma_{\hat{b}}$ and $\Sigma_{\hat{\sigma}}$ required for the Delta method are constructed using the usual asymptotic formulas, see Kilian and Lütkepohl (2017). We also considered an alternative version using standard errors obtained using a standard recursive non-parametric bootstrap with minor differences in results.

Confidence Sets for (α^d, α^s)

We start by testing for independent components using the permutation test of Matteson and Tsay (2017). The test does not reject that ϵ_t has independent components (p-value = 0.43), hence we conclude that our main identifying assumption is likely to hold and proceed with constructing confidence sets for the elasticity parameters.

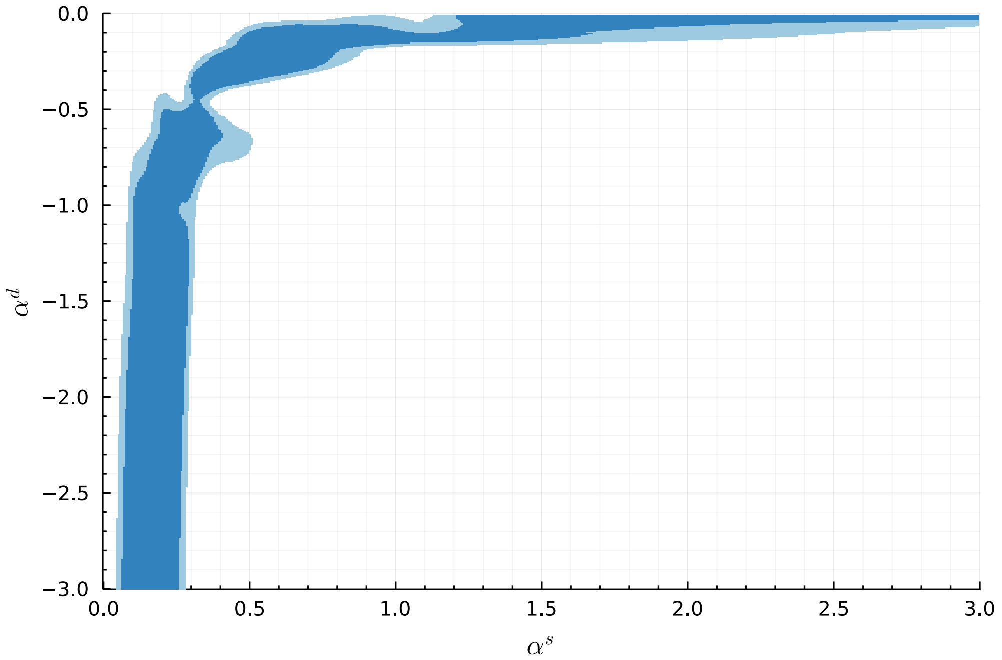
Figure 3.5 shows the 95% and 84% joint confidence sets for labor demand (α^d) and labor supply (α^s) parameters obtained using Algorithm 1 of Section 3.5. The confidence sets are constructed based on a grid of 250,000 equally spaced points for $(\alpha^d, \alpha^s) \in [-3, 0) \times (0, 3]$ which covers the majority of elasticity estimates reported in the microeconometric literature, as well as findings from theoretical macroeconomic models (see the discussion in Baumeister and Hamilton (2015)). The figure shows that overall, non-Gaussianity is not sufficient to pin down a precise region for the elasticities. For sufficiently negative values of the short-run demand elasticity, the short-run supply elasticity is reasonably well identified from non-Gaussianity with confidence sets indicating that α^s lies in the 0 - 0.3 range for both 95% and 84% confidence level. In contrast, for values of α^d that are less negative (smaller absolute value), the confidence sets support a wide range of values for the supply elasticity, spanning almost all values in the inspected grid. Our results match the findings of Baumeister and Hamilton (2015) who report that the main posterior mass for α^s lies in the 0 - 0.5 range while the posterior for α^d indicates that demand elasticities between -3 and 0 are well supported by the model.

Note that the estimate of Lanne and Luoto (2021) obtained using non-Gaussianity identification ($\alpha^d = 0.765, \alpha^s = -0.197$) falls within our confidence set at both levels. However, they find narrow confidence sets for the elasticity parameters (asymptotic standard errors of 0.196 for α^d and 0.057 for α^s, respectively) while our weak-identification robust approach results in much wider confidence sets, similar to the credible sets of Baumeister and Hamilton (2015).

Confidence Sets for impulse responses

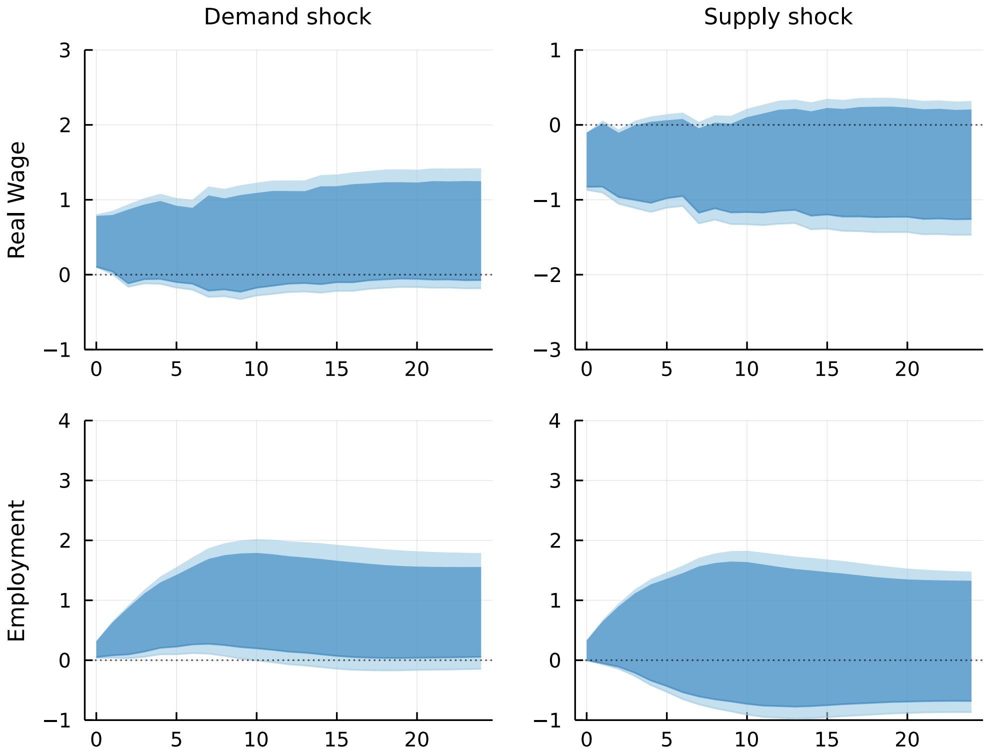
Figure 3.6 shows our identification-robust 90% and 67% confidence sets for the impulse responses to labor-demand and labor-supply shocks. Comparing the impulse response bands to the posterior credible sets reported by Baumeister and Hamilton (2015), we note that the implied impulse responses are, overall, very similar and show long and persistent responses to the supply and demand shocks. The main differences are that our 90% identification-robust bands support slightly negative long-run responses of the real wage and employment to a demand shock as well as a more pronounced negative long-run response of employment to a supply shock while Baumeister and Hamilton (2015)'s credible sets contain only (weakly) positive responses. Comparing our results to Lanne and Luoto (2021), we note several differences. First, Lanne and Luoto (2021) find a significant negative long-run response of the real wage to a supply shock while our confidence sets do not rule out that the long-run response is weakly positive. Second, and most important, they find a strong and significant dynamic response of both the real wage and employment to the labor demand shock, inconsistent with the tight prior variance Baumeister and Hamilton (2015) impose on the long-run response of employment to a demand shock. In contrast to their findings, our 90% identification-robust confidence bands do not rule out that the long-run response of either variable to the demand shock is zero. This evidence suggests that the long-run restriction of Baumeister and Hamilton (2015) cannot be rejected solely on the basis of non-Gaussianity.

Figure 3.5: Confidence Sets for Labor Demand and Supply Elasticities



Notes: 95% (light blue) and 84% (dark blue) confidence regions for labor demand and supply elasticities obtained using 250,000 equally-spaced grid points for $(\alpha^d, \alpha^s) \in [-3, 0) \times (0, 3]$.

Figure 3.6: IRF confidence bands for labor demand and supply shocks



Notes: 90% (light blue) and 67% (dark blue) identification-robust confidence bands for impulse responses to labor supply and labor demand shocks, obtained using 250,000 equally-spaced grid points for $(\alpha^d, \alpha^s) \in [-3, 0) \times (0, 3]$.

3.8.2. Oil price model of Kilian and Murphy (2012)

Next, we revisit the tri-variate oil market SVAR(p) model of Kilian and Murphy (2012). We have $Y_t = (\Delta q_t, x_t, p_t)'$ where Δq_t is the percent change in global crude oil production, x_t is an index of real economic activity representing the global business cycle and p_t is the log of the real price of oil.

$$Y_t = c + B_1 y_{t-1} + \cdots + B_p Y_{t-p} + A^{-1}(\alpha, \sigma)\, \epsilon_t, \quad A^{-1}(\alpha, \sigma) = \begin{bmatrix} \sigma_1 & \alpha_1 & \alpha_2 \\ \sigma_2 & \sigma_4 & \alpha_3 \\ \sigma_3 & \sigma_5 & \sigma_6 \end{bmatrix} \tag{3.30}$$

where $p = 24$. In this model, ϵ_t consists of a shock to the world production of crude oil (*"oil supply shock"*), a shock to the demand for crude oil and other industrial commodities associated with the global business cycle (*"aggregate demand shock"*), and a shock to demand for oil that is specific to the oil market (*"oil-market-specific demand shock"*).

The baseline model of Kilian and Murphy (2012) makes use of the following sign restrictions on the impact responses in A^{-1} to identify impulse responses:

$$A^{-1}(\alpha, \sigma) = \begin{bmatrix} - & + & + \\ - & + & - \\ + & + & + \end{bmatrix} \tag{3.31}$$

In addition, Kilian and Murphy (2012) impose a set of upper bounds on the short-run oil supply elasticities implied by the model to shrink the identified set for the impulse responses. Specifically, they assume that the short-run (impact) price elasticity of oil supply (α_2/σ_6) as well as the short-run (impact) demand elasticity of oil supply (α_1/σ_5) are smaller than 0.0258 and that $-1.5 < \alpha_3 < 0$. These restrictions, in particular the elasticity bounds, have been criticized by Baumeister and Hamilton (2019) as being too restrictive, and there is an active debate around which values for these bounds are warranted by the data (see Herrera and Rangaraju (2020) for an overview).

To address the identification problem, we consider the robust SVAR approach described in this paper that exploits possible non-Gaussianity in the structural shocks to simultaneously test $(\alpha_1, \alpha_2, \alpha_3) = (\alpha_{0,1}, \alpha_{0,2}, \alpha_{0,3})$ for different choices of α_0 to obtain confidence sets for α and confidence bands for the impulse responses to the oil supply shock, the aggregate demand shock and the oil-market-specific demand shock.

Specifically, we construct confidence sets for $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ using Algorithm 1 of Section 3.5 and confidence bands for the impulse responses using Algorithm 2 of Section 3.6. To this end, we estimate $b = (c, \operatorname{vec} B_1', \ldots, \operatorname{vec} B_p')'$ by equation-by-equation OLS, and given α_0 and $\hat{b}$, we estimate the variances $\sigma = (\sigma_1, \ldots, \sigma_6)$ by method of moments, solving the system of 6 equations implied by $\text{vech}\left(A^{-1}(\alpha_0,\sigma)A^{-1}(\alpha_0,\sigma)'\right) = \text{vech}\left(\hat{\Sigma}_u\right)$ for σ.[25] Confidence bands for the impulse responses are then constructed using the usual asymptotic Delta method approach with a Bonferroni correction, as in Section 3.6 .[26]

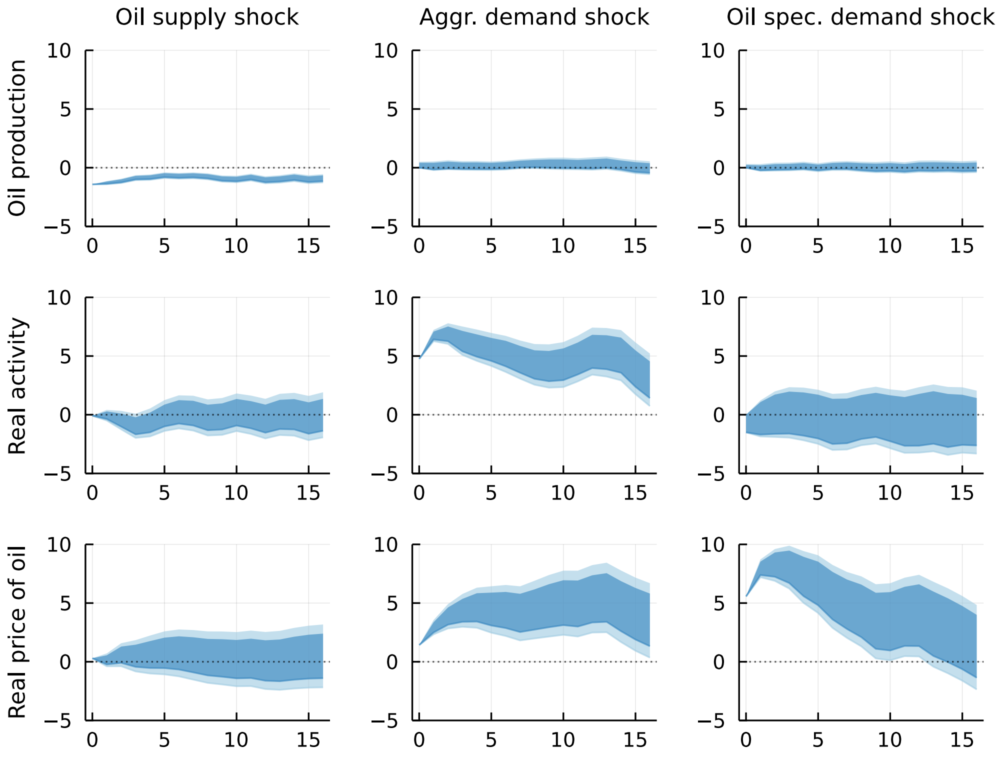
Confidence sets for oil supply elasticities

We start by testing for independent components using the permutation test of Matteson and Tsay (2017). The test does not reject that ϵ_t has independent components (p-value = 0.77), hence we conclude that our main identifying assumption is likely to hold and proceed with constructing confidence sets for the elasticity parameters.

Next, we construct confidence sets for α based on a grid of 8,000 equally spaced points for for $(\alpha_1, \alpha_2, \alpha_3) \in (0, 0.5] \times (0, 1.5] \times [-1.5, 0)$. The end points of the grid were chosen based on Kilian and Murphy (2012)'s assumption that $-1.5 < \alpha_3 < 0$ while allowing for a wider range of implied supply elasticities to address the critique of Baumeister and Hamilton (2019). In particular, note that based on the reduced-form estimate of Σ_u and the estimators $\hat{\sigma}(\alpha_0)$ for each α_0 in the grid, the maximum short-run supply elasticities supported by this grid are $(\alpha_1/\sigma_5,\ \alpha_2/\sigma_6) = (0.35,\ 0.27)$, well in excess of the bounds imposed by Kilian and Murphy (2012) and nesting a large share of supply elasticity estimates previously reported in the literature (Herrera and Rangaraju, 2020).

The resulting 84% confidence sets for α imply upper bounds on the elasticities $(\alpha_1/\sigma_5, \alpha_2/\sigma_6)$ equal to $(0.29, 0.04)$. We note that non-Gaussianity helps to identify the impact price elasticity of the oil supply since the upper bound implied by the confidence set is significantly reduced to about double the original bound considered by Kilian and Murphy (2012). In contrast, the upper bound on the impact elasticity of oil supply with respect to economic activity can not be well identified using non-Gaussianity as the upper bound remains close to the maximum value implied by the grid. Finally, we note that non-Gaussianity alone is not sufficient to pin down the response of real economic activity to an oil-specific supply shock, since the 84% confidence set for α_3 includes the bound of the grid ($\alpha_3 = -1.5$). Overall, based on the robust confidence sets, we conclude that relying on non-Gaussianity and sign restrictions alone is not sufficient to identify supply elasticities and unable to settle the current debate in the literature.

Figure 3.7: IRF Confidence Bands in the Oil Market Model



Notes: 90% (light blue) and 67% (dark blue) identification-robust confidence bands for the impulse responses to oil supply, aggregate demand and oil-specific demand shocks, obtained using 8,000 equally-spaced gird points for $(\alpha_1, \alpha_2, \alpha_3) \in (0, 0.5] \times (0, 1.5] \times [-1.5, 0)$.

Confidence Sets for Impulse Responses

Finally, we turn to inspecting the 90% and 67% confidence bands for impulse responses to oil supply, aggregate demand and oil-specific supply shocks which are depicted in Figure 3.7. We note that our confidence bands exhibit response patterns that are very close to the results reported in Kilian and Murphy (2012) based on sign restrictions and the more restrictive elasticity bounds. In particular, the responses of oil production are identified precisely while the responses of global real activity and of the real price of oil exhibit more uncertainty with insignificant and flat responses to the oil supply shock, significant positive hump-shaped responses to the aggregate demand shock and mixed response patterns to the oil-specific demand shock. While our impulse response bands agree with the results in Kilian and Murphy (2012), it is important to keep in mind that our confidence sets for α implied that the restriction we imposed on α_3 is binding. Hence, based on sign restrictions

[25] This system does not have a closed-form solution and we employ a numeric gradient-based algorithm to obtain $\hat{\sigma}$ with initial values corresponding to the recursive identification solution, $\alpha = (0, 0, 0)$, that recovers σ as the (sign-corrected) lower triangular Cholesky factor of $\hat{\Sigma}_u$.

[26] For each α, the estimates of $\Sigma_{\hat{b}}$ and $\Sigma_{\hat{\sigma}}$ required for the Delta method are constructed using the usual asymptotic formulas, see Kilian and Lütkepohl (2017). We also considered an alternative version using standard errors obtained using a standard recursive non-parametric bootstrap with minor differences in results.

and non-Gaussianity alone it is not possible to confirm the response patterns identified by Kilian and Murphy (2012) without making additional restrictions on impact responses.

3.9. Conclusion

This paper develops robust inference methods for structural vector autoregressive (SVAR) models that are identified via non-Gaussianity in the distributions of the structural errors. We treat the SVAR model as a semi-parametric model where the densities of the structural errors form the non-parametric part and conduct inference on the possibly weakly identified or non identified parameters of the SVAR, using a semi-parametric generalisation of Neyman's $\mathrm{C}(\alpha)$ statistic. We additionally provide a two-step Bonferroni-based approach to conduct inference on smooth functions of all the finite-dimension parameters of the model. We assess the finite-sample performance of our method in a large simulation study and find that the empirical rejection frequencies of the semi-parametric score test are always close to the nominal size, regardless of the true distribution of the errors. Moreover, the power of the test is typically higher than alternative methods that have been proposed in the literature. Finally, we employ the proposed approach in a number of empirical studies. Overall our findings are mixed. Whilst non-Gaussianity does provide some identifying information for the structural parameters of interest, it is unable to pin down all parameter values and/or impulse responses precisely. These exercises also highlight the importance of using weak identification robust methods to asses estimation uncertainty when using non-Gaussianity for identification.

Acknowledgements

I am grateful to my co-authors for their help in writing this chapter. I would also like to thank Christian Brownlees, Bjarni G. Einarsson, Kirill Evdokimov, Katerina Petrova, Barbara Rossi, André B. M. Souza and Philipp Tiozzo for helpful comments and discussions. I am also grateful to participants at the 12th Workshop on Time Series Econometrics (Zaragoza) for helpful comments and discussions.

Appendices

A. Proofs and additional results

A.1. Density score estimation

Lemma A.1. *Suppose Assumptions 3.3.1 and 3.3.2 hold. Let* $\tilde{\theta}_n = (\alpha_n, \tilde{\beta}_n, \eta) \to \theta$ *where* $\sqrt{n}\|\tilde{\beta}_n - \beta\| = O(1)$. *Then the log density score estimates* $\hat{\phi}_{k,n}$ *defined as in* (3.17) *satisfy for* $j, k = 1, \ldots, K$, $k \neq j$

$$\frac{1}{n}\sum_{t=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k\bullet}(Y_t - B_n X_t)) - \phi_k(A_{n,k\bullet}(Y_t - B_n X_t))\right] W_{n,t} = o_{P^n_{\tilde{\theta}_n}}(n^{-1/2}), \quad (32)$$

where $A_n := A(\alpha_n, \tilde{\beta}_n)$ *and* $B_n := B(\tilde{\beta}_n)$ *and for* $\nu_n = \nu_{n,p}^2$ *with* $1 < p \leq 1 + \delta/4$ *and* $n^{-1/2(1-1/p)} = o(\nu_{n,p})$ *we have*

$$\frac{1}{n}\sum_{t=1}^{n}\left(\left[\hat{\phi}_{k,n}(A_{n,k\bullet}(Y_t - B_n X_t)) - \phi_k(A_{n,k\bullet}(Y_t - B_n X_t))\right] W_{n,t}\right)^2 = o_{P^n_{\theta'_n}}(\nu_n). \quad (33)$$

where $W_{n,t}$ *are any random variables independent from all* $A_{n,k\bullet}(Y_s - c_n - B_n X_s)$ *with* $s > t$ *and such that* $\sup_{n\in\mathbb{N}, 1\leq t\leq n} \mathbb{E}_{\tilde{\theta}_n} W_{n,t}^2 < \infty$ *and* $\frac{1}{n}\sum_{t=1}^{n} W_{n,t}^2 - \mathbb{E}_{\tilde{\theta}_n} W_{n,t}^2 \xrightarrow{P^n_{\tilde{\theta}_n}} 0$.

Proof of Lemma A.1. The proof follows by an argument analogous used to prove Lemma 3 of Lee and Mesters (2022a); see Lee and Mesters (2022b) for the proof. □

A.2. Main proofs

Proof of Proposition 3.4.1. Throughout we work conditional on $(Y_{-p+1}, \ldots, Y_0)'$. Define

$$W_{n,t} := \frac{1}{2\sqrt{n}}\left[c'\dot{\ell}_{\theta_n}(Y_t, X_t) + \sum_{k=1}^{K} h_k(A_{n,k\bullet}V_{\theta_n,t})\right],$$

where $A_n := A(\alpha_n, \sigma_n)$, $\mathcal{F}_{n,t} := \sigma(Y_t, X_t)$, $\mathcal{F}_n := \mathcal{F}_{n,n}$ and note that $(W_{n,t}, \mathcal{F}_{n,t})_{n\in\mathbb{N},\, t\in[n]}$ forms an adapted stochastic process. Moreover it is clear that given assumption 3.3.12,

$$\mathbb{E}\left[W_{n,t}|\mathcal{F}_{n,t-1}\right] = \frac{1}{2\sqrt{n}}\left[c'\mathbb{E}\left[\dot{\ell}_{\theta}(Y_t, X_t)|\mathcal{F}_{n,t-1}\right] + \sum_{k=1}^{K}\mathbb{E}[h_k(A_{n,k\bullet}V_{\theta_n,t})|\mathcal{F}_{n,t-1}]\right] = 0, \quad (34)$$

almost surely, where the expectation is taken under $P^n_{\theta_n}$.

Next define $Z_{n,t} := (z_{n,t}/z_{n,t-1})^{1/2} - 1$ where $z_{n,0} = 1$ and else

$$z_{n,j} := \left(\frac{|\tilde{A}_n|}{|A_n|}\right)^j \times \prod_{t=1}^{j}\prod_{k=1}^{K} \frac{\eta_k(\tilde{A}_{n,k\bullet}\tilde{V}_{n,t})}{\eta_k(A_{n,k\bullet}V_{n,t})}\left(1 + h_{n,k}(\tilde{A}_{n,k\bullet}\tilde{V}_{n,t})/\sqrt{n}\right),$$

i.e.,

$$Z_{n,t} := \left[\frac{|\tilde{A}_n|}{|A_n|}\prod_{k=1}^{K} \frac{\eta_k(\tilde{A}_{n,k\bullet}\tilde{V}_{n,t})}{\eta_k(A_{n,k\bullet}V_{n,t})}\left(1 + h_{n,k}(\tilde{A}_{n,k\bullet}\tilde{V}_{n,t})/\sqrt{n}\right)\right]^{1/2} - 1.$$

We now verify conditions (S2) – (S6) of Theorem 2.1.2 in Taniguchi and Kakizawa (2000), having shown (S1) to hold above. (S2), i.e. that $\mathbb{E}\sum_{t=1}^{n}[W_{n,t} - Z_{n,t}]^2 \to 0$, where the expectation is taken under $P^n_{\theta_n}$ is shown to hold in Lemma A.5 below. (S3) – (S6) follow from Lemmas A.9 and A.10. (S3) follows immediately from Lemma A.9; (S5) follows from Lemma A.10 by Markov's inequality. For (S4), use the uniform integrability given by Lemma A.9 and Markov's inequality to obtain that for any $\varepsilon > 0$, as $n \to \infty$

$$\begin{aligned} P^n_{\theta_n}\left(\max_{1\le t\le n}|W_{n,t}| > \varepsilon\right) &\le P^n_{\theta_n}\left(\sum_{t=1}^{n} W^2_{n,t}\mathbf{1}\{|W_{n,t}| > \varepsilon\} > \varepsilon^2\right) \\ &\le \varepsilon^{-2}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[nW^2_{n,t}\mathbf{1}\{\sqrt{n}|W_{n,t}| > \varepsilon\sqrt{n}\}\right] \\ &\to 0. \end{aligned}$$

For (S6), note that the same UI argument as just used yields that

$$\lim_{n\to\infty}\sum_{t=1}^{n}\mathbb{E}\left[W^2_{n,t}\mathbf{1}\{|W_{n,t}| > \delta\}\right] = 0,$$

for some $\delta > 0$ and hence as conditional expectations are contractions in L_1,

$$\lim_{n\to\infty}\mathbb{E}\left|\sum_{t=1}^{n}\mathbb{E}\left[W^2_{n,t}\mathbf{1}\{|W_{n,t}| > \delta\}|\mathcal{F}_{n,t-1}\right]\right| = 0,$$

implying (S6). (L3) of Theorem 2.1.1 in Taniguchi and Kakizawa (2000) holds since the relevant measures are both absolutely continuous with respect to Lebesgue measure (cf. Taniguchi and Kakizawa, 2000, p. 34). By Theorem 2.1.2 of Taniguchi and Kakizawa (2000), under $P^n_{\theta_n}$:

$$\Lambda^n_{\tilde{\theta}_n/\theta_n}(Y^n) \rightsquigarrow \mathcal{N}(-\tau^2/2, \tau^2). \tag{35}$$

In view of Lemma A.10 and (S1) we have that $\Psi_\theta(c,h) := \lim_{n\to\infty}\mathbb{E}\left[\mathrm{g}_n(Y^n)^2\right] = \tau^2$ (in which the dependence on c, h is notationally supressed on the right hand side). Let $\varepsilon \in (0,1)$ be fixed and define $E_n := \{\max_{1\le t\le n}|Z_{n,t}| \le \varepsilon\}$ and note that by Theorem 2.1.2 of Taniguchi and Kakizawa (2000) $P^n_{\theta_n}E_n \to 1$. By Taylor expansion of $\log(1 + x)$, on E_n we have

$$\log(1 + Z_{n,t}) = Z_{n,t} - \frac{1}{2}Z^2_{n,t} + Z^2_{n,t}R(Z_{n,t}),$$

where $R(x) \leq M|x|$ for some $M \in [0, \infty)$ and so by (S2), on E_n

$$\begin{aligned}\Lambda^n_{\tilde{\theta}_n/\theta_n}(Y^n) &= 2\sum_{t=1}^{n} \log(Z_{n,t}+1) \\ &= \sum_{t=1}^{n} 2Z_{n,t} - \frac{1}{2}\sum_{t=1}^{n} 2Z_{n,t}^2 + \sum_{t=1}^{n} Z_{n,t}^2 R(Z_{n,t}).\end{aligned}$$

Moreover, by Theorem 2.1.2 of Taniguchi and Kakizawa (2000),

$$\sum_{t=1}^{n} Z_{n,t}^2 R(Z_{n,t}) \leq M \max_{1\leq t\leq n} |Z_{n,t}| \sum_{t=1}^{n} W_{n,t}^2 = o_{P^n_{\theta_n}}(1),$$

and so using also Lemma A.6

$$\Lambda^n_{\tilde{\theta}_n/\theta_n}(Y^n) = \sum_{t=1}^{n} 2W_{n,t} - \tau^2/4 - \frac{1}{2}\sum_{t=1}^{n} 2W_{n,t}^2 + o_{P^n_{\theta_n}}(1).$$

Lemma A.10, comparison of $W_{n,t}$ and $\mathrm{g}_n(Y^n)$ and the fact that the above display holds with $P^n_{\theta_n}$–probability approaching 1 yields the asymptotic expansion (3.10). The weak convergence of $\mathrm{g}_n(Y^n)$ follows by combining (3.10), (35) and (S5). □

Proof of Corollary 3.4.2. Combine (35) with Example 6.5 in van der Vaart (1998). □

Proof of Lemma 3.4.3. Define

$$\mathcal{T}^{\eta|\gamma}_{P_\theta,H} := \left\{ \sum_{t=1}^{n}\sum_{k=1}^{K} h_k(A_{k\bullet}V_{\theta,t}) : h = (h_1, \ldots, h_K) \in \dot{\mathscr{H}} \right\}, \quad V_{\theta,t} := Y_t - B_\theta X_t. \tag{36}$$

It suffices to show that (a) $\tilde{\ell}_\theta(Y_t, X_t) \in \left[\mathcal{T}^{\eta|\gamma}_{P_\theta,H}\right]^\perp \subset L_2(P^n_\theta)$ (componentwise) and (b) under P^n_θ

$$\mathbb{E}\left[\left(\dot{\ell}_\theta(Y_s, X_s) - \tilde{\ell}_\theta(Y_t, X_t)\right) \sum_{t=1}^{n}\sum_{k=1}^{K} h_k(A_{k\bullet}V_{\theta,t})\right] = 0 \quad \text{for all } h \in \dot{\mathscr{H}}.$$

For (a), the fact that $\tilde{\ell}_\theta(Y_s, X_s) \in L_2(P^n_\theta)$ follows straightfowardly from its form and the moment conditions in assumption 3.3.12. Next note that for any $h \in \dot{\mathscr{H}}$, $1 \leq s \leq n$,

$$\sum_{t=1}^{n}\sum_{k=1}^{K} \mathbb{E}\left[\tilde{\ell}_\theta(Y_s, X_s) h_k(A_{k\bullet}V_{\theta,t})\right] = 0$$

will obtain under P^n_θ if we have that for all $k, j, m \in [K]$ with $m \neq j$ and all $1 \leq s \leq n$, $1 \leq t \leq n$,

$$\begin{aligned}&\mathbb{E}\left[\phi_l(\epsilon_{m,s})\epsilon_{j,s} h_k(\epsilon_{k,t})\right] = 0 \\ &\mathbb{E}\left[\epsilon_{m,s} h_k(\epsilon_{k,t})\right] = 0 \\ &\mathbb{E}\left[\kappa(\epsilon_{m,s}) h_k(\epsilon_{k,t})\right] = 0 \\ &\mathbb{E}\left[(X_s - \mu)\phi_m(\epsilon_{m,s}) h_k(\epsilon_{k,t})\right] = 0,\end{aligned}$$

the first three of which follow from the independence between components and across time of $(\epsilon_t)_{t\geq 1}$. If $s \leq t$, then by independence $\mathbb{E}\left[(X_s-\mu)\phi_m(\epsilon_{m,s})h_k(\epsilon_{k,t})\right] = \mathbb{E}\left[(X_s-\mu)\phi_m(\epsilon_{m,s})\right]\mathbb{E}\left[h_k(\epsilon_{k,t})\right] = 0$. If $s > t$, then $\mathbb{E}\left[(X_s-\mu)\phi_m(\epsilon_{m,s})h_k(\epsilon_{k,t})\right] = \mathbb{E}\left[(X_s-\mu)h_k(\epsilon_{k,t})\mathbb{E}\left[\phi_m(\epsilon_{m,s})|\sigma(\epsilon_1,\ldots,\epsilon_{s-1})\right]\right] = 0$ again by independence.

For (b), that $\dot{\ell}_\theta(Y_s,X_s) - \tilde{\ell}_\theta(Y_s,X_s) \in L_2(P_\theta^n)$ follows from $\tilde{\ell}_\theta(Y_s,X_s) \in L_2(P_\theta^n)$ (as noted above) and Lemma A.9. Note that for any $h \in \dot{\mathscr{H}}$, $1 \leq s \leq n$,

$$\sum_{t=1}^{n}\mathbb{E}\left[\left(\dot{\ell}_\theta(Y_s,X_s) - \tilde{\ell}_\theta(Y_s,X_s)\right)\sum_{k=1}^{K}h_k(A_{k\bullet}V_{\theta,t})\right] = 0$$

will obtain under P_θ^n if we have that for any $m \in [K]$, $1 \leq t \leq n$, $1 \leq s \leq n$ and

$$\mathbb{E}\left[(\phi_m(\epsilon_{m,s})\epsilon_{m,s} + 1 - \tau_{m,1}\epsilon_{m,s} - \tau_{m,2}\kappa(\epsilon_{m,s}))\sum_{k=1}^{K}h_k(\epsilon_{k,t})\right] = 0$$
$$\mathbb{E}\left[(\phi_m(\epsilon_{m,s}) + \varsigma_{m,1}\epsilon_{m,s} + \varsigma_{m,2}\kappa(\epsilon_{m,s}))\sum_{k=1}^{K}h_k(\epsilon_{k,t})\right] = 0.$$

If $s \neq t$, both terms follows by independence (over t) of $(\epsilon_t)_{t\geq 1}$ and the definition of $\dot{\mathscr{H}}$. If $s = t$ the first term follows from the fact that the projection of $\phi_m(\epsilon_{m,t})\epsilon_{k,t} + 1$ on $[\mathcal{T}^{\eta|\gamma}_{P_\theta,H}]^\perp$ is $\tau_{k,1}\epsilon_{k,t} + \tau_{k,2}\kappa(\epsilon_{k,t})$ as follows from the analogous result in the proof of Lemma 2 of Lee and Mesters (2022a).[27] For the second term, if $s \neq t$, then this follows by independence (over t) of $(\epsilon_t)_{t\geq 1}$ and the definition of $\dot{\mathscr{H}}$. If $s = t$, then define $q(e) := \phi_m(e) + \varsigma_{m,1}e + \varsigma_{m,2}\kappa(e)$. $q(\epsilon_m)$ belongs to $\operatorname{cl}\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ as $q(\epsilon_{m,t}) \in L_2(P_\theta^n)$ and the choice of ς ensures that

$$\mathbb{E}[q(\epsilon_{m,t})] = \mathbb{E}[q(\epsilon_{m,t})\epsilon_{m,t}] = \mathbb{E}[q(\epsilon_{m,t})\kappa(\epsilon_{m,t})] = 0,$$

as is easily verified.[28] Define also $r(e) := \varsigma_{m,1}e + \varsigma_{m,2}\kappa(e)$. Then, by definition of $\dot{\mathscr{H}}$ we have that $r(\epsilon_{m,t}) \in [\mathcal{T}^{\eta|\gamma}_{P_\theta,H}]^\perp$. Hence we can write

$$\phi_m(\epsilon_{m,t}) = q(\epsilon_{m,t}) - r(\epsilon_{m,t})$$

where the first right hand side term belongs to $\operatorname{cl}\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ and the second to its orthogonal complement. Therefore, by theorem 4.11 of Rudin (1987), $-r(\epsilon)_{m,t}$ is the orthogonal projection of $\phi_m(\epsilon_{m,t})$ onto $[\mathcal{T}^{\eta|\gamma}_{P_\theta,H}]^\perp$ which implies that $\mathbb{E}\left[(\phi_m(\epsilon_{m,t}) - (-r(\epsilon_{m,t})))\sum_{k=1}^{K}h_k(\epsilon_{k,t})\right] = 0$. □

Proof of Theorem 3.5.1. Define

$$R_{n,1}(\beta_\star) := \left\|\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\gamma_\star} - \tilde{\ell}_{\theta_\star}\right]\right\|$$
$$R_{n,2}(\beta_\star) := \left\|\sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\theta_\star} - \tilde{\ell}_{\theta_n}\right] + \sqrt{n}\tilde{I}_{n,\theta_n}(\gamma_\star - \gamma_n)'\right\|$$
$$R_{n,3}(\beta_\star) := \left\|\hat{I}_{n,\gamma_\star} - \tilde{I}_\theta\right\|,$$

[27] See Lee and Mesters (2022b) for the proof.

[28] That $\operatorname{cl}\mathcal{T}^{\eta|\gamma}_{P_\theta,H}$ is the set of L_2 random variables satisfying these equations can be shown by an argument analogous to that in footnote S5 of Lee and Mesters (2022b).

where $\gamma_\star := (\alpha_n, \beta_\star)$ and $\theta_\star := (\gamma_\star, \eta)$. We show that we have

$$R_{n,i}(\tilde{\theta}_n) \xrightarrow{P^n_{\tilde{\theta}_n}} 0 \quad \text{for } i = 1, 2, 3. \tag{37}$$

Define $b_n := \sqrt{n}(\beta'_n - \beta)$. We may assume without loss of generality that $b_n \to b$ and $h_n \to h$.[29]

Let Q_n denote the law of $(Y_t)_{t=1}^n$ corresponding to $\tilde{\theta}_n := (\alpha_n, \beta + b_n/\sqrt{n}, \eta(1 + h_n/\sqrt{n}))$ and P_n that corresponding to $\check{\theta}_n := (\alpha_n, \beta + b_n/\sqrt{n}, \eta)$ (both conditional on the initial observations). By Corollary 3.4.2 $Q_n \lhd\rhd P_n$ and hence (37) follows by Lemma A.12 and Le Cam's first Lemma (e.g. van der Vaart, 1998, Lemma 6.4).

Next we show that (37) continues to hold if the argument of the remainders $R_{n,i}$ is replaced by $\bar{\beta}_n$ as defined in the theorem. Since $\bar{\beta}_n$ remains $\sqrt{n}$-consistent there is an $M > 0$ such that $P^n_{\tilde{\theta}_n}\left(\sqrt{n}\|\bar{\beta}_n - \beta\| > M\right) < \varepsilon$. If $\sqrt{n}\|\bar{\beta}_n - \beta\| \le M$ then $\bar{\beta}_n$ is equal to one of the values in the finite set $\mathscr{S}_n^c = \{\beta' \in n^{-1/2}C\mathbb{Z}^{L_2} : \|\beta' - \beta\| \le n^{-1/2}M\}$. For each M this set has finite number of elements bounded independently of n, call this upper bound $\overline{B}$. Letting R_n denote any of $R_{n,1}$, $R_{n,2}$ or $R_{n,3}$ we have that for any $\upsilon > 0$

$$\begin{aligned}
P^n_{\tilde{\theta}_n}\left(\|R_n(\bar{\beta}_n)\| > \upsilon\right) &\le \varepsilon + \sum_{\beta_n \in \mathscr{S}_n^c} P^n_{\tilde{\theta}_n}\left(\{\|R_n(\beta_n)\| > \upsilon\} \cap \{\bar{\beta}_n = \beta_n\}\right) \\
&\le \varepsilon + \sum_{\beta_n \in \mathscr{S}_n^c} P^n_{\tilde{\theta}_n}\left(\|R_n(\beta_n)\| > \upsilon\right) \\
&\le \varepsilon + \overline{B} P^n_{\tilde{\theta}_n}\left(\|R_n(\beta_n^*)\| > \upsilon\right),
\end{aligned}$$

where $\beta_n^* \in B_n$ maximises $\beta \mapsto P^n_{\tilde{\theta}_n}\left(\|R_n(\beta)\| > \upsilon\right)$. As $(\beta_n^*)_{n\in\mathbb{N}}$ is a deterministic $\sqrt{n}$-consistent sequence for β we have that $P^n_{\tilde{\theta}_n}\left(\|R_n(\beta_n^*)\| > \upsilon\right) \to 0$ by (37).

It follows that

$$\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\gamma}_n} - \tilde{\ell}_{\theta_n}\right] = \sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\gamma}_n} - \tilde{\ell}_{\bar{\theta}_n}\right] + \sqrt{n}\mathbb{P}_n\left[\tilde{\ell}_{\bar{\theta}_n} - \tilde{\ell}_{\theta_n}\right] = -\tilde{I}_{n,\theta_n}(0, \sqrt{n}(\bar{\beta}_n - \beta)')' + o_{P^n_{\tilde{\theta}_n}}(1),$$

and $\hat{I}_{n,\bar{\theta}_n} \xrightarrow{P^n_{\tilde{\theta}_n}} \tilde{I}_\theta$ and so $\hat{\mathcal{K}}_{n,\bar{\theta}_n} \xrightarrow{P^n_{\tilde{\theta}_n}} \tilde{\mathcal{K}}_\theta$ for

$$\tilde{\mathcal{K}}_\theta := \begin{bmatrix} I & -\tilde{I}_{\theta,\alpha\beta}\tilde{I}^{-1}_{\theta,\beta\beta} \end{bmatrix}, \quad \hat{\mathcal{K}}_{n,\theta} := \begin{bmatrix} I & -\hat{I}_{n,\theta,\alpha\beta}\hat{I}^{-1}_{n,\theta,\beta\beta} \end{bmatrix}.$$

We combine these to obtain

$$\begin{aligned}
&\sqrt{n}\mathbb{P}_n\left[\hat{\kappa}_{n,\bar{\gamma}_n} - \tilde{\kappa}_{n,\theta_n}\right] \\
&\quad = \left(\hat{\mathcal{K}}_{n,\bar{\gamma}_n} - \tilde{\mathcal{K}}_{\theta_n}\right)\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\gamma}_n} - \tilde{\ell}_{\theta_n}\right] + \tilde{\mathcal{K}}_{\theta_n}\sqrt{n}\mathbb{P}_n\left[\hat{\ell}_{\bar{\gamma}_n} - \tilde{\ell}_{\theta_n}\right] + \left(\hat{\mathcal{K}}_{n,\bar{\gamma}_n} - \tilde{\mathcal{K}}_{\theta_n}\right)\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\theta_n} \\
&\quad = -\tilde{\mathcal{K}}_{\theta_n}\tilde{I}_{\theta_n}(0, \sqrt{n}(\bar{\beta}_n - \beta)')' + o_{P^n_{\tilde{\theta}_n}}(1) \\
&\quad = -\begin{bmatrix} I & -\tilde{I}_{\theta_n,\alpha\beta}\tilde{I}^{-1}_{\theta_n,\beta\beta} \end{bmatrix} \begin{bmatrix} \tilde{I}_{\theta_n,\alpha\alpha} & \tilde{I}_{\theta_n,\alpha\beta} \\ \tilde{I}_{\theta_n,\beta\alpha} & \tilde{I}_{\theta_n,\beta\beta} \end{bmatrix} \begin{bmatrix} 0 \\ \sqrt{n}(\bar{\beta}_n - \beta) \end{bmatrix} + o_{P^n_{\tilde{\theta}_n}}(1) \\
&\quad = o_{P^n_{\tilde{\theta}_n}}(1).
\end{aligned}$$

[29] Otherwise the same argument can proceed along appropriately chosen subsequences.

Next, let $Z_n := \frac{1}{\sqrt{n}}\sum_{t=1}^n \hat{\kappa}_{n,\bar{\gamma}_n}(Y_t, X_t)$ and re-write it as

$$Z_n = \frac{1}{\sqrt{n}}\sum_{t=1}^n \tilde{\kappa}_{n,\theta_n}(Y_t, X_t) + \frac{1}{\sqrt{n}}\sum_{t=1}^n (\hat{\kappa}_{n,\bar{\gamma}_n}(Y_t, X_t) - \tilde{\kappa}_{n,\theta_n}(Y_t, X_t)) = \frac{1}{\sqrt{n}}\sum_{t=1}^n \tilde{\kappa}_{n,\theta_n}(Y_t, X_t) + o_{P^n_{\tilde{\theta}_n}}(1).$$

By (i) of Lemma A.12 and Le Cam's third lemma (e.g. van der Vaart, 1998, Example 6.7)

$$\frac{1}{\sqrt{n}}\sum_{t=1}^n \tilde{\ell}_{\theta_n}(Y_t, X_t) \rightsquigarrow \mathcal{N}\left(\tilde{I}_\theta (0', b')', \tilde{I}_\theta\right) \qquad \text{under } P_{\tilde{\theta}_n},$$

and hence under $P_{\tilde{\theta}_n}$

$$Z_n = \frac{1}{\sqrt{n}}\sum_{t=1}^n \tilde{\ell}_{\theta_n,\alpha}(Y_t, X_t) - \tilde{I}_{n,\theta_n,\alpha\beta}\tilde{I}^{-1}_{n,\theta_n,\beta\beta}\tilde{\ell}_{\theta_n,\beta}(Y_t, X_t) + o_{P^n_{\tilde{\theta}_n}}(1) \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{\mathcal{I}}_\theta).$$

We additionally have

$$\left\|\hat{\mathcal{I}}_{n,\bar{\gamma}_n} - \tilde{\mathcal{I}}_\theta\right\|_2 \leq \left\|\hat{I}_{n,\bar{\gamma}_n,\alpha\alpha} - \tilde{I}_{\theta,\alpha\alpha}\right\|_2 + \left\|\hat{I}_{n,\bar{\gamma}_n,\alpha\beta}\hat{I}^{-1}_{n,\bar{\gamma}_n,\beta\beta}\hat{I}_{n,\bar{\gamma}_n,\beta\alpha} - \tilde{I}_{\theta,\alpha\beta}\tilde{I}^{-1}_{\theta,\beta\beta}\tilde{I}_{\theta,\beta\alpha}\right\|_2 .$$

By repeated addition and subtraction along with the observations that any submatrix has a smaller operator norm than the original matrix we obtain and the matrix inverse is Lipschitz continuous at a non-singular matrix we obtain

$$\left\|\hat{\mathcal{I}}_{n,\bar{\gamma}_n} - \tilde{\mathcal{I}}_\theta\right\|_2 \lesssim \left\|\hat{I}_{n,\bar{\gamma}_n} - \tilde{I}_\theta\right\|_2 .$$

Hence by equation (37) with $\bar{\gamma}_n$ replacing γ_n we have $P_{\tilde{\theta}_n}\left(\left\|\hat{\mathcal{I}}_{n,\bar{\gamma}_n} - \tilde{\mathcal{I}}_\theta\right\|_2 < \check{\nu}_n\right) \to 1$ where $\check{\nu}_n = C\nu_n$ for some positive constant $C \geq 1$. By Proposition 3.13 and Lemma C.6 of Lee (2022)

$$\hat{\mathcal{I}}^{t,\dagger}_{n,\bar{\gamma}_n} \xrightarrow{P^n_{\tilde{\theta}_n}} \tilde{\mathcal{I}}^\dagger_\theta \qquad \text{and} \qquad P^n_{\tilde{\theta}_n} R_n \to 1,$$

where $R_n := \{\operatorname{rank}(\tilde{\mathcal{I}}^t_{n,\bar{\gamma}_n}) = \operatorname{rank}(\tilde{\mathcal{I}}_\theta)\}$.

Suppose first that $r := \operatorname{rank}(\tilde{\mathcal{I}}_\theta) > 0$. By Slutsky's lemma and the continuous mapping theorem we have that

$$\hat{S}^{SR}_{n,\bar{\gamma}_n} = Z_n'\hat{\mathcal{I}}^{t,\dagger}_{n,\bar{\gamma}_n}Z_n \rightsquigarrow Z'\tilde{\mathcal{I}}^\dagger_\theta Z \sim \chi^2_r$$

where the distributional result $X := Z'\tilde{\mathcal{I}}^\dagger_\theta Z \sim \chi^2_r$, follows from e.g. Theorem 9.2.2 in Rao and Mitra (1971). On R_n c_n is the $1-a$ quantile of the χ^2_r distribution, which we will call c. Hence, we have $c_n \xrightarrow{P^n_{\tilde{\theta}_n}} c$ and as a result,$\hat{S}^{SR}_{n,\bar{\gamma}_n} - c_n \rightsquigarrow X - c$ where $X \sim \chi^2_r$. Since the χ^2_r distribution is continuous, we have by the Portmanteau theorem

$$P^n_{\tilde{\theta}_n}\left(\hat{S}^{SR}_{n,\bar{\gamma}_n} > c_n\right) = 1 - P^n_{\tilde{\theta}_n}\left(\hat{S}^{SR}_{n,\bar{\gamma}_n} - c_n \leq 0\right) \to 1 - \mathrm{P}\left(X - c \leq 0\right) = 1 - \mathrm{P}\left(X \leq c\right) = 1 - (1-a) = a\,,$$

which completes the proof in the case that $r > 0$.

It remains to handle the case with $r = 0$. We first note that $Z_n \rightsquigarrow Z \sim \mathcal{N}(0, \tilde{\mathcal{I}}_\theta)$ continues to hold by our assumptions, though in this case $\tilde{\mathcal{I}}_\theta$ is the zero matrix and hence the limiting distribution is degenerate: $Z = 0$.

On the sets R_n we have that $\hat{\mathcal{I}}^t_{n,\bar{\gamma}_n}$ is the zero matrix, whose Moore-Penrose inverse is also the zero matrix. Hence on these sets we have $\hat{S}^{SR}_{n,\bar{\gamma}_n} = 0$ and $c_n = 0$ and therefore do not reject, implying

$$P^n_{\tilde{\theta}_n}(\hat{S}^{SR}_{n,\bar{\gamma}_n} > c_n) \leq 1 - P^n_{\tilde{\theta}_n} R_n \to 0.$$

It follows that $P^n_{\tilde{\theta}_n}(\hat{S}^{SR}_{n,\bar{\gamma}_n} > c_n) \to 0$. □

Proof of Corollary 3.5.2. Apply Theorem 3.5.1 to conclude:

$$\lim_{n\to\infty} P^n_{\tilde{\theta}_n}(\alpha_n \in \hat{C}_n) \geq 1 - \lim_{n\to\infty} P^n_{\tilde{\theta}_n}(\hat{S}^{SR}_{n,\bar{\gamma}_n} > c_n) \geq 1 - \alpha.$$

□

Proof of Proposition 3.6.1. Let G be a convex, compact set with $G \supset \{\gamma_n : n \geq N_0\}$ for some $N_0 \in \mathbb{N}$. Since g is continuously differentiable and G is compact, $\{\|g'_\gamma\| : \gamma \in G\}$ is bounded and hence $\{g'_{\gamma_n} : n \in \mathbb{N}\}$ is uniformly equicontinuous (cf. Remark A.2). By compactness, $\gamma \mapsto g'_\gamma$ is uniformly continuous on G. Combined with the mean-value theorem (e.g. Drabek and Milota, 2007, Theorem 3.2.7) this implies that g is uniformly differentiable along $(\gamma_n)_{n\in\mathbb{N}}$. By Theorem A.19 and the fact that $\mathcal{N}(0, M_n) \xrightarrow{TV} \mathcal{N}(0, M)$ if $M_n \to M \succ 0$,

$$\sqrt{n}\left(g(\alpha_n, \check{\beta}_n) - g(\alpha_n, \tilde{\beta}_n)\right) \overset{P^n_{\tilde{\theta}_n}}{\rightsquigarrow} \mathcal{N}\left(0, J_\gamma V_\theta J'_\gamma\right).$$

This and the fact that $\check{V}_n \xrightarrow{P^n_{\tilde{\theta}_n}} J_\gamma V_\theta J'_\gamma$ imply that

$$ng(\alpha_n, \check{\beta}_n)'\check{V}_n^{-1} g(\alpha_n, \check{\beta}_n) \rightsquigarrow \chi^2_{d_g} \quad \text{under } P^n_{\tilde{\theta}_n}.$$

It follows that

$$\lim_{n\to\infty} P^n_{\tilde{\theta}_n}(g(\alpha_n, \tilde{\beta}_n) \in \check{C}_{n,g,\alpha_n}) = \lim_{n\to\infty} P^n_{\tilde{\theta}_n}\left(ng(\alpha_n, \check{\beta}_n)'\check{V}_n^{-1} g(\alpha_n, \check{\beta}_n) \leq c_a\right) = 1 - a.$$

□

Proof of Proposition 3.6.2. This follows directly from the hypotheses and the fact that

$$\begin{aligned} P^n_{\tilde{\theta}_n}\left(g(\alpha_n, \tilde{\beta}_n) \in \hat{C}_{n,g}\right) &\geq P^n_{\tilde{\theta}_n}\left(\{g(\alpha_n, \check{\beta}_n) \in \check{C}_{n,g,\alpha_n}\} \cap \left\{\alpha_n \in \hat{C}_n\right\}\right) \\ &\geq P^n_{\tilde{\theta}_n}\left(g(\alpha_n, \check{\beta}_n) \in \check{C}_{n,g,\alpha_n}\right) + P^n_{\tilde{\theta}_n}\left(\alpha_n \in \hat{C}_n\right) - 1. \end{aligned}$$

□

A.3. Auxilliary results

Here we record results relating to the model under study in the main text, which are used in establishing the main results which are proven above.

Define $Z_t := (Y_t', Y_{t-1}', \ldots, Y_{t-p+1}')'$, $\mathsf{C}_\theta := (c_\theta', 0', \ldots, 0')'$,

$$\mathsf{B}_\theta := \begin{bmatrix} B_{\theta,1} & B_{\theta,2} & \cdots & B_{\theta,p-1} & B_{\theta,p} \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}, \quad \mathsf{D}_\theta := \begin{bmatrix} A_\theta^{-1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and note that we can write

$$Z_t = \mathsf{C}_\theta + \mathsf{B}_\theta Z_{t-1} + \mathsf{D}_\theta \epsilon_t. \tag{38}$$

Proposition A.2. *Suppose that assumption 3.3.1 holds. Then $(Z_t)_{t\geq 0}$ (with initial value $Z_0 = z$) is a uniformly ergodic Markov chain on $\mathbb{R}^{Kp}$. Moreover for any compact set $K \subset \mathbb{R}^{d_\gamma}$, we have that for any (initial value) $z \in \mathbb{R}^{Kp}$,*

$$\sup_{\theta=(\gamma,\eta):\ \gamma\in K} \|Q_\theta^n(z,\cdot) - \pi_\theta(\cdot)\|_{TV} \leq (M_1 + \|z\|^2)\gamma^n, \quad \textit{for some } \gamma < 1,\ M_1 < \infty$$

and π_θ an invariant probability distribution for Φ (under θ) and for $M_2 < \infty$

$$\sup_{\theta=(\gamma,\eta):\ \gamma\in K} \beta_\theta(n) \leq (4M_1 + 3\|z\|^2 + M_2)\gamma^{\lfloor n/2 \rfloor},$$

where $\beta_\theta(n)$ are the β-mixing coefficients of Φ.

Proof. That $\Phi := (Z_t)_{t\geq 0}$ is a Markov chain follows from Proposition 11.6 in Kallenberg (2021). Explicit computation of the rank of the controllability matrix (Meyn and Tweedie, 2009, equation (4.13)) demonstrates that the associated linear control model is controllable. Moreover under assumption 3.3.1, (LSS4) and (LSS5) of Meyn and Tweedie (2009) hold and hence by Proposition 6.3.5 in Meyn and Tweedie (2009), Φ is a ψ-irreducible T-chain and every compact subset is a small set. Aperiodicity of Φ follows from the assumptions on the densities.

The 1-step transition probability is given by the density on $\mathbb{R}^{Kp} \times \mathbb{R}^{Kp}$ defined as

$$q_\theta(y,x) := |A_\theta| \prod_{k=1}^{K} \eta_k(A_{\theta,k} V_\theta), \quad V_\theta := y_1 - c_\theta - \sum_{l=1}^{p} B_{\theta,l} x_l,$$

where e.g. y_1 denotes the first K elements of y and similarly for x. By assumption 3.3.1, the map $(\gamma, y, x) \mapsto q_{(\gamma,\eta)}(y,x)$ is continuous and positive everywhere on $\Gamma \times \mathbb{R}^{Kp} \times \mathbb{R}^{Kp}$. For any compact $B \subset \mathbb{R}^{Kp}$ put $\varepsilon := \int \inf_{(\gamma,x)\in K\times B} q_{(\gamma,\eta)}(y,x)\,\mathrm{d}y$ and $\rho(y) := \inf_{(\gamma,x)\in K\times B} q_{(\gamma,\eta)}(y,x)/\varepsilon$.[30] Then for any $A \in \mathcal{B}(\mathbb{R}^{Kp})$ and any $x \in B$,

$$\int_A q_\theta(y,x)\,\mathrm{d}y \geq \varepsilon \int_A \rho(y)\,\mathrm{d}y.$$

Under assumption 3.3.1 the eigenvalues of B_θ are bounded above by some $\overline{\rho} < 1$ for all $\theta \in \mathsf{T} := \{(\gamma,\eta) : \gamma \in K\}$. Using this and the Gelfand formula (e.g. Horn and Johnson, 2013, Corollary 5.6.14) there exists a $\rho_\star < 1$ with $\|\mathsf{B}_\theta^n\| \leq \rho_\star^n$ on T. Since we can re-write

[30] Note that $\epsilon > 0$ by the positivity and continuity.

(38) as

$$Z_t - m_\theta = \mathsf{B}_\theta(Z_{t-1} - m_\theta) + \mathsf{D}_\theta\epsilon_t, \tag{39}$$

with $m_\theta := \left(\sum_{i=0}^{\infty} \mathsf{B}_\theta^i\right) \mathsf{C}_\theta$, we have

$$V_\theta(Z_t) = \|\mathsf{B}_\theta(Z_{t-1} - m_\theta)\|^2 + \|\mathsf{D}_\theta\epsilon_t\|^2 + 2[\mathsf{B}_\theta(Z_{t-1} - m_\theta)]'\mathsf{D}_\theta\epsilon_t + 1,$$

and since ϵ_t is independent of Z_{t-1}, and $\|\mathsf{D}_\theta\| \leq D_\star < \infty$ on T,

$$\mathbb{E}[V_\theta(Z_t)|Z_{t-1}] \leq \rho_\star^2\|Z_{t-1} - m_\theta\|^2 + D_\star^2 \leq \rho_\star^2 V_\theta(Z_{t-1}) + D_\star^2.$$

This, in conjunction with Proposition 5.5.3 and Lemmas 15.2.8 of Meyn and Tweedie (2009) establishes that the Markov chain satisfies the drift condition (10) in Roberts and Rosenthal (2004) with $\lambda = (1 + \rho_\star^2)/2 < 1$, $b = D_\star^2 < \infty$ and $C = C_\theta = \{z : V_\theta(z) \leq 2D_\star^2/(1 - \rho_\star^2)\}$. By Proposition 11 in Roberts and Rosenthal (2004) their bivariate drift condition (11) is satisfied with $h(x, y) = [V_\theta(x) + V_\theta(y)]/2$ and $\alpha^{-1} = \lambda + b/(d+1) < 1$. Moreover $b_{0,\theta} := \max\{1, \alpha(1 - \varepsilon)\sup_{(x,y)\in C_\theta\times C_\theta} \bar{R}_\theta h_\theta(x, y)\}$ is bounded above by $(1 - \varepsilon)D_\star^2/(1 - \rho_\star^2) < \infty$, where $\bar{R}_\theta h_\theta(x, y)$ is defined analogously to $\bar{R}h(x, y)$ on p. 41 of Roberts and Rosenthal (2004). By Theorem 16.0.2 of Meyn and Tweedie (2009) there exists an invariant π_θ with

$$\|Q_\theta^n(z, \cdot) - \pi_\theta\| \leq Rr^{-n}, \quad R < \infty,\ r > 1,$$

where $Q_\theta(z, \cdot)$ is the transition probability. That is, Φ is uniformly ergodic.

For the second claim, by Theorem 12 in Roberts and Rosenthal (2004) we have that for any (initial) $z \in \mathbb{R}^{Kp}$ and some $\gamma < 1$, for all $\theta \in \mathsf{T}$,

$$\|Q_\theta^n(z, \cdot) - \pi_\theta\|_{TV} \leq (M_1 + \|z\|^2)\gamma^n,$$

where[31]

$$M_1 = 1 + \sup_{\theta\in\mathsf{T}} \|m_\theta\|^2 + \sup_{\theta\in\mathsf{T}} \int \|z - m_\theta\|^2 \,\mathrm{d}\pi_\theta(z) < \infty.$$

The claim regarding the β-mixing coefficients then follows directly from Proposition 3 in Liebscher (2005), with $M_2 := \sup_{\theta\in\mathsf{T}} \int \|z\|^2 \,\mathrm{d}\pi_\theta(z) < \infty$.[32] □

Lemma A.3. *Suppose that assumption 3.3.1 holds. Define $U_{\theta,t}$ as the (unique, strictly) stationary solution to* (38) *(under θ). Then $U_{\theta,t}$ has the representation*

$$U_{\theta,t} = m_\theta + \sum_{j=0}^{\infty} \mathsf{B}_\theta^j \mathsf{D}_\theta\epsilon_{t-j}, \quad m_\theta := (I - \mathsf{B}_\theta)^{-1}\mathsf{C}_\theta, \quad \sum_{j=0}^{\infty} \|\mathsf{B}_\theta^j\| < \infty.$$

If ρ_θ is the largest absolute eigenvalue of the companion matrix B_θ and $v > 0$ is such that

[31] That the first supremum is finite is clear since $m_\theta = (I - \mathsf{B}_\theta)^{-1}\mathsf{C}_\theta$ which is evidently continuous. For the second supremum note that the integral is taking an expectation with respect to the distribution of the stationary solution of a VAR model. This is bounded uniformly over $\theta \in \mathsf{T}$ by Lemma A.3, the fact that $\|\mathsf{D}_\theta\|$ is uniformly bounded on T and the observation that since $M \mapsto \rho(M)$ is continuous and T is compact, there is a ρ and v with $\rho + v < 1$ such that $\rho \geq \rho(\mathsf{B}_\theta)$ for all $\theta \in \mathsf{T}$.

[32] The uniform boundedness of M_2 follows by an analogous argument as given in footnote 31.

$\rho_\theta + \upsilon < 1$, the for $\|\cdot\|$ the spectral norm,

$$\mathbb{E}\|U_{\theta,t} - m_\theta\|^{\rho} \leq \frac{\mathbb{E}\|\mathsf{D}_\theta \epsilon_t\|^{\rho}}{1 - (\rho_\theta + \upsilon)^{\rho}}, \quad \rho \in [1, 4+\delta].$$

Proof. Rewriting (38) as (39) and applying Theorem 11.3.1 in Brockwell and Davis (1991) yields the first part. For the second part, let $\mathcal{U}_\theta^* \mathcal{J}_\theta \mathcal{U}_\theta$ be a Schur decomposition of B_θ. Then

$$\|U_{\theta,t} - m_\theta\| \leq \sum_{j=0}^{\infty} \|\mathsf{B}_\theta^j\| \|\mathsf{D}_\theta \epsilon_{t-j}\| \leq \sum_{j=0}^{\infty} \|\mathcal{J}_\theta\|^j \|\mathsf{D}_\theta \epsilon_{t-j}\| \leq \sum_{j=0}^{\infty} (\rho_\theta + \nu)^j \|\mathsf{D}_\theta \epsilon_{t-j}\|.$$

Since $\mathbb{E}\|\mathsf{D}_\theta \epsilon_{t-j}\|^{\rho} = \mathbb{E}\|\mathsf{D}_\theta \epsilon_t\|^{\rho} < \infty$ for all $t \in \mathbb{N}$, all $j \geq 0$ and $\rho \in [1, 4+\delta]$, it follows that

$$\mathbb{E}\|U_{\theta,t} - m_\theta\|^{\rho} \leq \sum_{j=0}^{\infty} (\rho_\theta + \nu)^{j\rho} \mathbb{E}\|\mathsf{D}_\theta \epsilon_{t-j}\|^{\rho} = \frac{\mathbb{E}\|\mathsf{D}_\theta \epsilon_t\|^{\rho}}{1 - (\rho_\theta + \nu)^{\rho}}.$$

□

Corollary A.4. *Suppose that assumption 3.3.1 holds and $\theta_n = (\gamma_n, \eta) \to (\gamma, \eta) = \theta$. Define π_θ as in Proposition A.2 and let $\mathscr{G}_{\theta_n,n}$ be the measure corresponding to the density $\frac{1}{n}\sum_{t=1}^n \rho_{\theta_n,t}$ where $\rho_{\theta_n,t}$ is the density of the non-deterministic parts of X_t under $P_{\theta_n}^n$ $(1 \leq t \leq n)$. Then $\mathscr{G}_{\theta_n,n} \rightsquigarrow \pi_\theta$.*

Proof. By Proposition A.2, $\mathscr{G}_{\theta,n} \xrightarrow{TV} \pi_\theta$ uniformly on $\mathsf{T} := \{\theta_n : n \in \mathbb{N}\} \cup \{\theta\}$. We also have that $\pi_{\theta_n} \rightsquigarrow \pi_\theta$. To see this, use the representation in Lemma A.3 and the fact that we can uniformly bound $\|\mathsf{B}_\vartheta^j\|$ and $\|\mathsf{D}_\vartheta\|$ for $\vartheta \in \mathsf{T}$ and $j \in \mathbb{N}$ to obtain

$$\begin{aligned}
\mathbb{E}\,\|U_{\theta_n,t} - U_{\theta,t}\| &\leq \|m_{\theta_n} - m_\theta\| + \mathbb{E}\left\|\sum_{j=0}^{\infty} \mathsf{B}_{\theta_n}^j \mathsf{D}_{\theta_n} \epsilon_{t-j} - \mathsf{B}_\theta^j \mathsf{D}_\theta \epsilon_{t-j}\right\| \\
&= o(1) + \mathbb{E}\|\epsilon_t\| \sum_{j=0}^{\infty} \left(\|\mathsf{B}_{\theta_n}^j\| \|\mathsf{D}_{\theta_n} - \mathsf{D}_\theta\| + \|\mathsf{D}_\theta\| \|\mathsf{B}_{\theta_n}^j - \mathsf{B}_\theta^j\| \right) \\
&= o(1)
\end{aligned}$$

where the second equality uses the fact the ϵ_t are identically distributed and the third equality uses the dominated convergence theorem.[33] This implies that $U_{\theta_n,t} \rightsquigarrow U_{\theta,t}$ as $n \to \infty$, i.e. $\pi_{\theta_n} \rightsquigarrow \pi_\theta$. Combination of these results yields the claim. □

Lemma A.5 (UDQM). *Suppose that assumption 3.3.1 holds. Then, with $W_{n,t}$ and $Z_{n,t}$ defined as in the proof of Proposition 3.4.1,*

$$\lim_{n\to\infty} \mathbb{E} \sum_{t=1}^{n} (W_{n,t} - Z_{n,t})^2 = 0.$$

Proof. Write $Y_{n,t}$ and $X_{n,t}$ for random elements which have the same law as Y_t, X_t

[33] Note that $\|\mathsf{B}_{\theta_n}^j - \mathsf{B}_\theta^j\| \to 0$ pointwise in j and is dominated by $2\rho_\star^j$ where $\rho_\star < 1$ is a uniform upper bound on $\|B_\vartheta\|$ for $\vartheta \in \mathsf{T}$ and $\sum_{j=0}^{\infty} 2\rho_\star^j = 2\sum_{j=0}^{\infty} \rho_\star^j < \infty$.

(respectively) under $P^n_{\theta_n}$. Recall $V_{n,t} := Y_{n,t} - BX_{n,t}$ and define

$$q_\theta(Y_{n,t}, X_{n,t}) := |A| \prod_{k=1}^{K} \eta_k(A_{k\bullet}V_t), \qquad g_\theta(Y_{n,t}, X_{n,t}) := c'\dot{\ell}_\theta(Y_{n,t}, X_{n,t}) + \sum_{k=1}^{K} h_k(A_{k\bullet}V_{n,t}). \tag{40}$$

Let $\varphi(u) = (c, \eta_1 h_1, \ldots, \eta_K h_K)$ for $u = (c, h)$ with $c \in \mathbb{R}^{L_\alpha + L_\beta}$, $h \in \dot{\mathscr{H}}$. We initially suppose that $\theta_n = \theta$ for all $n \in \mathbb{N}$ and argue similarly to Lemma 7.6 in van der Vaart (1998). By Assumption 3.3.1 and standard computations, the derivative of $s \mapsto \sqrt{q_{\theta + s\varphi(u)}}$ at $s = \mathsf{s}$ is $\frac{1}{2} g_{\theta + \mathsf{s}\varphi(u)} \sqrt{q_{\theta + \mathsf{s}\varphi(u)}}$ (everywhere). Inspection reveals that this is continuous in s. Let $\rho_{\theta,t}$ be as defined in Corollary A.4. Define

$$I_{\theta,t} := \int g_\theta^2 q_\theta \rho_{\theta,t} \,\mathrm{d}\lambda.$$

By the mean-value theorem and Jensen's inequality we can write

$$\int \left(\frac{\sqrt{q_{\vartheta_{1,n}}} - \sqrt{q_\theta}}{1/\sqrt{n}} \right)^2 \rho_{\theta,t} \,\mathrm{d}\lambda \leq \frac{1}{4} \int \int_0^1 (g_{\vartheta_{v,n}} \sqrt{q_{\vartheta_{v,n}}})^2 \rho_{\theta,t} \,\mathrm{d}v \,\mathrm{d}\lambda = \frac{1}{4} \int_0^1 I_{\vartheta_{v,n},t} \,\mathrm{d}v \tag{41}$$

where $\vartheta_{v,n} := \theta + \frac{v}{\sqrt{n}} \varphi(u)$ and the last step follows by Tonelli's theorem.

It is shown in Lemma A.7 that as $n \to \infty$,

$$\frac{1}{n} \sum_{t=1}^{n} \int_0^1 I_{\vartheta_{v,n},t} \,\mathrm{d}v = \int_0^1 \int g^2_{\vartheta_{v,n}} \,\mathrm{d}G_{\theta_{v,n},n} \,\mathrm{d}v \to \int g_\theta^2 \,\mathrm{d}G_\theta < \infty, \tag{42}$$

where $G_{\theta,n}$ is as defined in Lemma A.11. Using this, we can re-write

$$\sum_{t=1}^{n} \int \left(\sqrt{q_{\vartheta_{1,n}}} - \sqrt{q_\theta} - \frac{1}{2\sqrt{n}} g_\theta \sqrt{q_\theta} \right)^2 p_{\theta,t} \,\mathrm{d}\lambda = \int \left(\sqrt{n} \left[\frac{\sqrt{q_{\vartheta_{1,n}}}}{\sqrt{q_\theta}} - 1 \right] - \frac{1}{2} g_\theta \right)^2 \mathrm{d}G_{\theta,n}. \tag{43}$$

By the assumed differentiability, the integrand in the last integral converges pointwise to zero. Combining this with (41), (42) and (43) with Proposition A.16 we have

$$\lim_{n \to \infty} \int \left(\sqrt{n} \left[\sqrt{q_{\vartheta_{1,n}}} - \sqrt{q_\theta} \right] - \frac{1}{2} \Delta_\theta(u) \sqrt{q_\theta} \right)^2 \bar{\rho}_{\theta,n} \,\mathrm{d}\lambda = 0, \tag{44}$$

where $\bar{\rho}_{\theta,n} := \frac{1}{n} \sum_{t=1}^{n} \rho_{\theta,t}$ and $\Delta_\theta(u) := g_\theta$, to emphasise the linearity in u of g_θ. We next show that any $u_n \to u$, $\mathsf{u}_n \to \mathsf{u}$ (all in U), and any $(v_n)_{n \in \mathbb{N}} \subset [0, \infty)$ with $v_n \downarrow 0$,

$$\lim_{n \to \infty} \int \left[\Delta_{\theta_n + v_n \varphi(\mathsf{u}_n)}(u_n) \sqrt{q_{\theta_n + v_n \varphi(\mathsf{u}_n)}} - \Delta_{\theta_n}(u) \sqrt{q_{\theta_n}} \right]^2 \bar{\rho}_{\theta_n,n} \,\mathrm{d}\lambda = 0. \tag{45}$$

We first note that for any (deterministic) convergent sequence $x_n \to x$, we have

$$[\Delta_{\theta_n + v_n \varphi(\mathsf{u}_n)}(u_n) \sqrt{q_{\theta_n + v_n \varphi(\mathsf{u}_n)}}](\cdot, x_n) - [\Delta_\theta(u) \sqrt{q_\theta}](\cdot, x) \to 0,$$

pontwise in y. This follows by the continuity of the relevant functions and that, for $\check{\vartheta}_n := \theta_n + v_n \varphi(\mathsf{u}_n)$, (i)

$$(y - B_{\check{\vartheta}_n} x_n) - (y - B_\theta x) = B_{\check{\vartheta}_n}(x_n - x) + (B_{\check{\vartheta}_n} - B_\theta)x \to 0,$$

since $\vartheta \mapsto B_\vartheta$ is continuous and (ii), since $\vartheta \mapsto A_\vartheta$ is continuous,

$$A_{\check{\vartheta}_n,k\bullet} D_{b_l} x_n - A_{\theta,k\bullet} D_{b_l} x = A_{\check{\vartheta}_n,k\bullet} D_{b_l}(x_n - x) + (A_{\check{\vartheta}_n,k\bullet} - A_{\theta,k\bullet}) D_{b_l} x \to 0.$$

The form of $\dot{\ell}_{\check{\vartheta}_n}$ is the same as that given in (3.7) – (3.9) once each ϕ_k is replaced by

$$\tilde{\phi}_{k,n} := \phi_k + \frac{v_n h_k/\sqrt{n}}{1 + v_n h_k/\sqrt{n}}, \tag{46}$$

and, moreover, since $\check{\vartheta}_n \to \theta$, the continuity and continuous differentiability conditions in assumption 3.3.1 ensure that all non-random terms in the expressions (3.7) – (3.9) converge and are thus bounded.[34] Noting this and directly integrating, it follows that

$$\lim_{n\to\infty} \int \left([\Delta_{\theta_n + v_n\varphi(\mathsf{u}_n)}(u_n)\sqrt{q_{\theta_n+v_n\varphi(\mathsf{u}_n)}}](y, x_n) \right)^2 \, \mathrm{d}y = \int ([\Delta_\theta(u)\sqrt{q_\theta}](y,x))^2 \, \mathrm{d}y < \infty,$$

and hence by Proposition 2.29 in van der Vaart (1998),

$$\int \left([\Delta_{\theta_n + v_n\varphi(\mathsf{u}_n)}(u_n)\sqrt{q_{\theta_n+v_n\varphi(\mathsf{u}_n)}}](y, x_n) - \Delta_\theta(u)\sqrt{q_\theta}(y,x) \right)^2 \, \mathrm{d}y \to 0.$$

Taking $v_n = 0$, $u_n = u$ and $\theta_n = \theta$ in the above yields also

$$\int ([\Delta_\theta(u_n)\sqrt{q_\theta}](y, x_n) - \Delta_\theta(u)\sqrt{q_\theta}(y,x))^2 \, \mathrm{d}y \to 0,$$

and hence we have that

$$\mathscr{Q}_n(x) := \int \left([\Delta_{\theta_n + v_n\varphi(\mathsf{u}_n)}(u_n)\sqrt{q_{\theta_n+v_n\varphi(\mathsf{u}_n)}}](y, x) - [\Delta_\theta(u_n)\sqrt{q_\theta}](y,x) \right)^2 \, \mathrm{d}y$$

converges continuously to 0. Using the form given in (54) for the (non-deterministic) parts of X_t and noting (as discussed following (54)) that $\{\rho(\mathsf{B}_\vartheta) : \vartheta \in \{\theta_n : n \in \mathbb{N}\} \cup \theta\}$ is bounded, and similarly that $\{\|A_\vartheta^{-1}\| : \vartheta \in \{\theta_n : n \in \mathbb{N}\} \cup \{\theta\}\}$ is bounded, it is easy to see that $\sup_{n\in\mathbb{N},1\leq t\leq n} \mathbb{E}\|X_t\| < \infty$. Hence by Markov's inequality for any $\varepsilon > 0$, there is an M such that $\sup_{n\in\mathbb{N},1\leq t\leq n} P^n_{\theta_n}(\|X_t\| \leq M) \geq 1 - \epsilon$ and so the family $\{X_{n,t} : n \in \mathbb{N}, 1 \leq t \leq n\}$ is uniformly tight, where each $X_{n,t}$ is a random variable (defined on a common probability space) with law $\mathscr{L}(X_t | P^n_{\theta_n})$. Let $(t_n)_{n\in\mathbb{N}}$ be an arbitrary sequence of positive integers satisfying $t_n \leq n$ and put $\tilde{X}_n := X_{n,t_n}$. The sequence $(\tilde{X}_n)_{n\in\mathbb{N}}$ is uniformly tight. It follows by Prohorov's theorem that any subsequence $(\tilde{X}_{k_n})_{n\in\mathbb{N}}$ contains a further subsequence $(\tilde{X}_{m_n})_{n\in\mathbb{N}}$ with $X_{m_n} \rightsquigarrow X$ for some random variable X. Since $(\mathscr{Q}_n)_{n\in\mathbb{N}}$ is continuously convergent to the zero function, it follows by the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1) that $\mathscr{Q}_{m_n}(\tilde{X}_{m_n}) \rightsquigarrow 0$. Equation (45) will then follow provided we show that $(\mathscr{Q}_n(\tilde{X}_n))_{n\in\mathbb{N}}$ is uniformly integrable. For this, dominate the n-th term by

$$\begin{aligned}
\mathscr{Q}_n(\tilde{X}_n) &\leq 2 \left[\int [\Delta_{\check{\vartheta}_n}(u_n)\sqrt{q_{\check{\vartheta}_n}}](y, \tilde{X}_n)^2 \, \mathrm{d}y + \int [\Delta_\theta(u)\sqrt{q_\theta}](y, \tilde{X}_n)^2 \, \mathrm{d}y \right] \\
&= 2 \left[\mathbb{E}\left[\Delta_{\check{\vartheta}_n}(u_n)(\tilde{Y}_n, \tilde{X}_n)^2 \middle| \tilde{X}_n \right] + \mathbb{E}\left[\Delta_\theta(u)(\tilde{Y}, \tilde{X}_n)^2 \middle| \tilde{X}_n \right] \right],
\end{aligned}$$

where $\tilde{Y}_n$ and $\tilde{Y}$ have laws such that their conditional density given $\tilde{X}_n$ is $q_{\check{\vartheta}_n}$ and q_θ,

[34] Cf. footnote 40.

respectively. Lemma A.9 and (46) ensure that $(\Delta_{\breve{\vartheta}_n}(u_n)(\tilde{Y}_n, \tilde{X}_n)^2)_{n\in\mathbb{N}}$ is UI. Combining this with the conditional Jensen inequality and the de la Valée Poussin criterion for uniform integrability (e.g. Bogachev, 2007, Theorem 4.5.9) yields that the first conditional expectation in the preceeding display is UI. That the second conditional expectation is also UI follows similarly.

To complete the proof, first let $\theta \in \Theta$ be arbitrary, $s_n := n^{-1/2}$, $u_n \to u$, and use (44), the mean-value theorem (e.g. Drabek and Milota, 2007, Theorem 3.2.7(i)) and (45) to obtain

$$\begin{aligned}
&\left\| \left(\frac{\sqrt{q_{\theta+s_n\varphi(u_n)}} - \sqrt{q_\theta}}{s_n} - \frac{1}{2}\Delta_\theta(u)\sqrt{q_\theta} \right) \sqrt{\bar{\rho}_{\theta,n}} \right\|_{\lambda,2} \\
&\quad \leq \left\| \left(\frac{\sqrt{q_{\theta+s_n\varphi(u_n)}} - \sqrt{q_{\theta+s_n\varphi(u)}}}{s_n} \right) \sqrt{\bar{\rho}_{\theta,n}} \right\|_{\lambda,2} \\
&\qquad + \left\| \left(\frac{\sqrt{q_{\theta+s_n\varphi(u)}} - \sqrt{q_\theta}}{s_n} - \frac{1}{2}\Delta_\theta(u)\sqrt{q_\theta} \right) \sqrt{\bar{\rho}_{\theta,n}} \right\|_{\lambda,2} \\
&\quad \leq \sup_{\delta\in[0,1]} \left\| \frac{1}{2}\Delta_{\theta+s_n\varphi(u)+s_n\delta\varphi(u_n-u)}(u_n-u)\sqrt{q_{\theta+s_n\varphi(u)+s_n\delta\varphi(u_n-u)}}\sqrt{\bar{\rho}_{\theta,n}} \right\|_{\lambda,2} + o(1) \\
&\quad = o(1).
\end{aligned}$$

Now return to our original setting with $\theta_n = (\gamma_n, \eta) \to \theta = (\gamma, \eta)$. By the preceding display, applying the mean-value theorem (e.g. Drabek and Milota, 2007, Theorem 3.2.7(ii)) at each n gives

$$\begin{aligned}
\sum_{t=1}^n \mathbb{E}\left[(W_{n,t} - Z_{n,t})^2\right] &= \left\| \left(\frac{\sqrt{q_{\theta_n+s_n\varphi(u_n)}} - \sqrt{q_{\theta_n}}}{s_n} - \frac{1}{2}\Delta_{\theta_n}(u)\sqrt{q_{\theta_n}} \right) \sqrt{\bar{\rho}_{\theta_n,n}} \right\|_{\lambda,2} \\
&\leq \sup_{\delta\in[0,1]} \left\| \frac{1}{2}\left(\Delta_{\theta_n+\delta s_n\varphi(u_n)}(u_n)\sqrt{q_{\theta_n+\delta s_n\varphi(u_n)}} - \Delta_{\theta_n}(u)\sqrt{q_{\theta_n}} \right) \sqrt{\bar{\rho}_{\theta_n,n}} \right\|_{\lambda,2} \\
&= o(1),
\end{aligned}$$

where the convergence in the last line is due to (45). □

Lemma A.6. *In the setting of Proposition 3.4.1, it holds that*

$$2\sum_{t=1}^n Z_{n,t} = 2\sum_{t=1}^n W_{n,t} - \tau^4/2 + o_{P^n_{\theta_n}}(1).$$

Proof. Throughout expectations are taken under $P^n_{\theta_n}$. Let $m_n(X_t) := \mathbb{E}[Z_{n,t}|X_t] = \mathbb{E}[Z_{n,t}|\mathcal{F}_{n,t-1}]$ with $\mathcal{F}_{n,t} := \sigma(\epsilon_i : i = 1, \ldots, t)$.[35] Form $U_{n,t} := Z_{n,t} - m_n(X_t) - W_{n,t}$ and note that $(U_{n,t}, \mathcal{F}_{n,t})_{n\in\mathbb{N}, 1\leq t\leq n}$ is a martingale difference array (by (34)). Hence

$$\mathbb{V}\left[\sum_{t=1}^n U_{n,t}\right] = \sum_{t=1}^n \mathbb{E}\left[Z_{n,t} - W_{n,t}\right]^2 + \sum_{t=1}^n \mathbb{E}[m_n(X_t)^2] - 2\sum_{t=1}^n \mathbb{E}\left[(Z_{n,t} - W_{n,t})m_n(X_t)\right].$$

[35] See e.g. Theorem 7.3.1 in Chow and Teicher (1997) for the (almost sure) equality of the conditional expectations.

Observe that

$$\mathbb{E}\left[(Z_{n,t} - W_{n,t})m_n(X_t)\right] = \mathbb{E}\left[\mathbb{E}\left[(Z_{n,t} - W_{n,t})m_n(X_t)|X_t\right]\right] = \mathbb{E}\left[m_n(X_t)\mathbb{E}\left[Z_{n,t}|X_t\right]\right] = \mathbb{E}\left[m_n(X_t)^2\right],$$

and so by Lemma A.5

$$0 \leq \mathbb{V}\left[\sum_{t=1}^{n} U_{n,t}\right] = \sum_{t=1}^{n} \mathbb{E}\left[Z_{n,t} - W_{n,t}\right]^2 - \sum_{t=1}^{n} \mathbb{E}[m_n(X_t)^2] \leq \sum_{t=1}^{n} \mathbb{E}\left[Z_{n,t} - W_{n,t}\right]^2 \to 0,$$

which, in combination with (L3) of Theorem 2.1.1 in Taniguchi and Kakizawa (2000) (which is noted to hold in the proof of Proposition 3.4.1), yields

$$2\sum_{t=1}^{n} Z_{n,t} - 2\sum_{t=1}^{n} W_{n,t} + \sum_{t=1}^{n} \mathbb{E}[Z_{n,t}^2|\mathcal{F}_{n,t-1}] = o_{P_{\theta_n}^n}(1).$$

It therefore suffices to show that $\sum_{t=1}^{n} \mathbb{E}[Z_{n,t}^2|\mathcal{F}_{n,t-1}] \xrightarrow{P_{\theta_n}^n} \tau^2/4$. For this, first observe that by Lemma A.10,

$$\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{2}\Delta_{\theta_n}(u)(Y_t, X_t)\right)^2 \xrightarrow{P_{\theta_n}^n} \frac{\tau^2}{4}.$$

Next, since the $\left(\frac{1}{2}\Delta_{\theta_n}(u)\right)^2$ are UI by Lemma A.9, applying Theorem 2.22 in Hall and Heyde (1980), Jensen's inequality for conditional expectations and the de la Vallée Poussin criterion for uniform integrability (e.g. Bogachev, 2007, Theorem 4.5.9) we have that

$$\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{2}\Delta_{\theta_n}(u)(Y_t, X_t)\right)^2 - \mathbb{E}\left[\left(\frac{1}{2}\Delta_{\theta_n}(u)(Y_t, X_t)\right)^2 |\mathcal{F}_{n,t-1}\right] \xrightarrow{L_1} 0.$$

To complete the proof it therefore suffices show that

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}([\sqrt{n}Z_{n,t}]^2|\mathcal{F}_{n,t-1}) - \mathbb{E}\left[\left(\frac{1}{2}\Delta_{\theta_n}(u)\right)^2 |\mathcal{F}_{n,t-1}\right] \xrightarrow{L_1} 0. \tag{47}$$

Since $\mathbb{E}[\mathscr{U}_{n,t}|\mathcal{F}_{n,t-1}] = \mathbb{E}[\mathscr{U}_{n,t}|X_t]$ for $\mathscr{U}_{n,t} \in \left\{[\sqrt{n}Z_{n,t}]^2, \left(\frac{1}{2}\Delta_{\theta_n}(u)(Y_t, X_t)\right)^2\right\}$,[36] we

[36] Cf. footnote 35.

have

$$\begin{aligned}
&\mathbb{E}\left|\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}([\sqrt{n}Z_{n,t}]^2|\mathcal{F}_{n,t-1}) - \mathbb{E}\left[\left(\frac{1}{2}\Delta_{\theta_n}(u)(Y_t,X_t)\right)^2|\mathcal{F}_{n,t-1}\right]\right| \\
&\quad\leq \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left|\mathbb{E}([\sqrt{n}Z_{n,t}]^2|\mathcal{F}_{n,t-1}) - \mathbb{E}\left[\left(\frac{1}{2}\Delta_{\theta_n}(u)(Y_t,X_t)\right)^2|\mathcal{F}_{n,t-1}\right]\right| \\
&\quad\leq \frac{1}{n}\sum_{t=1}^{n}\int\int\left|\sqrt{n}(q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}) - \frac{1}{2}\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\right|\left|\sqrt{n}(q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}) + \frac{1}{2}\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\right|\,\mathrm{d}y\rho_{\theta_n,t}\,\mathrm{d}x \\
&\quad= \left\langle\left|\sqrt{n}(q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}) - \frac{1}{2}\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\right|\bar{\rho}_{\theta_n,n}^{1/2},\ \left|\sqrt{n}(q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}) + \frac{1}{2}\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\right|\bar{\rho}_{\theta_n,n}^{1/2}\right\rangle_{\lambda} \\
&\quad\leq \left\|\left[\sqrt{n}(q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}) - \frac{1}{2}\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\right]\bar{\rho}_{\theta_n,n}^{1/2}\right\|_{\lambda,2}\left\|\left[\sqrt{n}(q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}) + \frac{1}{2}\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\right]\bar{\rho}_{\theta_n,n}^{1/2}\right\|_{\lambda,2},
\end{aligned}$$

by Cauchy-Schwarz. The proof of (47) (and hence the Lemma) is completed by applying Lemmas A.5, A.9 and noting

$$\begin{aligned}
&\left\|\left[\sqrt{n}(q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}) + \frac{1}{2}\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\right]\bar{\rho}_{\theta_n,n}^{1/2}\right\|_{\lambda,2} \\
&\qquad\leq \left(\left\|\left[\sqrt{n}(q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}) - \frac{1}{2}\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\right]\bar{\rho}_{\theta_n,n}^{1/2}\right\|_{\lambda,2} + \left\|\Delta_{\theta_n}(u)q_{\theta_n}^{1/2}\bar{\rho}_{\theta_n,n}\right\|_{\lambda,2}\right).
\end{aligned}$$

□

Lemma A.7. *Suppose that assumption 3.3.1 holds. Then* (42) *in the proof of Lemma A.5 holds.*

Proof. The finiteness of the integral on the right hand side follows by direct evaluation using the moment bounds in assumption 3.3.1 along with the fact that under π_θ, $\mathbb{E}\|X_t\|^{4+\delta} < \infty$ which can be seen on combining Lemma A.3 with the fact that π_θ is the law of a stationary solution to the defining VAR equation (see e.g. Kallenberg, 2021, Theorem 11.11).

For the convergence, by Lemma A.11 and Corollary 2.9 in Feinberg et al. (2016) it is enough to prove the uniform $G_{\vartheta_{v_n,n},n}$ – integrability of $(g^2_{\vartheta_{v_n,n}})_{n\in\mathbb{N}}$ for an arbitrary $(v_n)_{n\in\mathbb{N}} \subset [0,1]$. As each h_k is bounded, it suffices to show $\sup_{n\in\mathbb{N}}\int\left|c'\dot{\ell}_{\vartheta_{v_n,n}}\right|^{2+\delta/2}\mathrm{d}G_{\vartheta_{v_n,n},n} < \infty$ for some $\delta > 0$. The form of $\dot{\ell}_{\vartheta_{v_n,n}}$ is the same as that given in equations (3.7) – (3.9) once each ϕ_k is replaced by

$$\tilde{\phi}_{k,n} := \phi_k + \frac{v_n h_k/\sqrt{n}}{1 + v_n h_k/\sqrt{n}},$$

where, since each h_k is bounded, the second term is bounded for large enough n. Since $\vartheta_{v_n,n} \to \theta$, the continuity and continuous differentiability conditions in assumption 3.3.1 ensure that all non-random terms in the expressions (3.7) – (3.9) converge and are thus bounded.[37] The required bound then follows as, under $G_{\vartheta_{v_n,n},n}$ we have that $V_{\vartheta_{v_n,n},t} \sim \epsilon_t$, with independent components and also independent of X_t, and $\sup_{n\in\mathbb{N}}\mathbb{E}[|\epsilon_t|^{4+\delta}] < \infty$, $\sup_{n\in\mathbb{N}}\mathbb{E}[|\phi_k(\epsilon_t)|^{4+\delta}] < \infty$ and $\sup_{n\in\mathbb{N}}\mathbb{E}\|X_t\|^{4+\delta} < \infty$. The first two moment

[37] Cf. footnote 40.

bounds are immediate from assumption 3.3.1. The latter follows since under each $\rho_{\theta,t}$, $\sup_{n\in\mathbb{N},1\leq t\leq n}\mathbb{E}\|X_t\|^{4+\delta}<\infty$ which follows as in the proof of Lemma A.9 and hence

$$\sup_{n\in\mathbb{N}}\int\|x\|^{4+\delta}\frac{1}{n}\sum_{t=1}^{n}\rho_{\theta,t}(x)\,\mathrm{d}\lambda\leq\sup_{n\in\mathbb{N},1\leq t\leq n}\int\|x\|^{4+\delta}\rho_{\theta,t}(x)\,\mathrm{d}\lambda<\infty.$$

□

Lemma A.8 (Cf. Lemma A.10 in van der Vaart (1988a)). *Suppose that Assumption 3.3.1 holds. Then for any $\tilde{\theta}_n$ which takes the form $\tilde{\theta}_n=(\alpha_n,\beta_n+b_n/\sqrt{n},\eta)$ with $b_n\to b\in\mathbb{R}^{L_\beta}$ a convergent sequence,*

$$R_n:=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\tilde{\ell}_{\tilde{\theta}_n}(Y_t,X_t)-\tilde{\ell}_{\theta_n}(Y_t,X_t)\right]+\tilde{I}_{n,\theta_n}(0',b_n')'\xrightarrow{P^n_{\theta_n}}0.$$

Proof. Let q_θ be as defined in Lemma A.5 and note that the sequence which is to be shown to converge to zero (in probability) can be written as the sum of the following two terms

$$\begin{aligned}R_{1,n}&:=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\tilde{\ell}_{\tilde{\theta}_n}(Y_t,X_t)\left(1-\frac{q_{\tilde{\theta}_n}(Y_t,X_t)^{1/2}}{q_{\theta_n}(Y_t,X_t)^{1/2}}\right)\right]+\frac{1}{2}\tilde{I}_{n,\theta_n}(0',b_n')'\\R_{2,n}&:=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\tilde{\ell}_{\tilde{\theta}_n}(Y_t,X_t)\frac{q_{\tilde{\theta}_n}(Y_t,X_t)^{1/2}}{q_{\theta_n}(Y_t,X_t)^{1/2}}-\tilde{\ell}_{\theta_n}(Y_t,X_t)\right]+\frac{1}{2}\tilde{I}_{n,\theta_n}(0',b_n')'\end{aligned}$$

To simplify notation, let $Z_{n,t,1}:=\tilde{\ell}_{\tilde{\theta}_n}(Y_t,X_t)\frac{q_{\tilde{\theta}_n}(Y_t,X_t)^{1/2}}{q_{\theta_n}(Y_t,X_t)^{1/2}}$ and $Z_{n,t,2}:=\tilde{\ell}_{\theta_n}(Y_t,X_t)$. Define $m_n(x):=\int\tilde{\ell}_{\tilde{\theta}_n}(y,x)q_{\tilde{\theta}_n}(y,x)^{1/2}q_{\theta_n}(y,x)^{1/2}\,\mathrm{d}y$. Evaluated at X_t, this is the conditional (on X_t) expectation of $Z_{n,t,1}$. Observe that since $\mathbb{E}[\tilde{\ell}_{\theta_n}(Y_t,X_t)|X_t]=0$ under $P^n_{\tilde{\theta}_n}$,

$$\begin{aligned}m_n(X_t)&=\int\tilde{\ell}_{\tilde{\theta}_n}(y,X_t)q_{\tilde{\theta}_n}(y,X_t)^{1/2}q_{\theta_n}(y,X_t)^{1/2}\,\mathrm{d}y\\&=\int\tilde{\ell}_{\tilde{\theta}_n}(y,X_t)q_{\tilde{\theta}_n}(y,X_t)^{1/2}\left[q_{\theta_n}(y,X_t)^{1/2}-q_{\tilde{\theta}_n}(y,X_t)^{1/2}\right]\,\mathrm{d}y.\end{aligned}$$

Let $\rho_{\theta_n,t}$ be the density of (the non-deterministic parts of) X_t under $P^n_{\theta_n}$, $\bar{\rho}_{\theta_n,n}:=\frac{1}{n}\sum_{t=1}^{n}\rho_{\theta_n,t}$ and $G_{\theta_n,n}$ be the measure corresponding to $\bar{\rho}_{\theta_n,n}$. By Lemma A.5,

$$\lim_{n\to\infty}\int\left[\sqrt{n}\left(q_{\theta_n}^{1/2}-q_{\tilde{\theta}_n}^{1/2}\right)\bar{\rho}_{\theta_n,n}^{1/2}+\frac{1}{2}b_n'\dot{\ell}_{\theta_n}q_{\theta_n}^{1/2}\bar{\rho}_{\theta_n,n}^{1/2}\right]^2\,\mathrm{d}\lambda=0.\tag{48}$$

Additionally,

$$\lim_{n\to\infty}\int\left\|\tilde{\ell}_{\theta_n}q_{\theta_n}^{1/2}\bar{\rho}_{\theta_n,n}^{1/2}-\tilde{\ell}_{\tilde{\theta}_n}q_{\tilde{\theta}_n}^{1/2}\bar{\rho}_{\theta_n,n}^{1/2}\right\|^2\,\mathrm{d}\lambda=0.\tag{49}$$

To demonstrate this we first note that by inspection of their forms, it is clear that for ϑ_n equal to either θ, θ_n or $\tilde{\theta}_n$ and any $x_n\to x$, $\tilde{\ell}_{\vartheta_n}(y,x_n)q_{\vartheta_n}(y,x_n)^{1/2}\to\tilde{\ell}_\theta(y,x)q_\theta(y,x)^{1/2}$ (pointwise in y). Moreover, noting the fact that these integrals are expectations conditional on X and using the forms given in Lemma 3.4.3 along with the continuity given by

Assumption 3.3.1 we have that

$$\lim_{n\to\infty} \int \left\| \tilde{\ell}_{\vartheta_n}(y, x_n) q_{\vartheta_n}^{1/2}(y, x_n) \right\|^2 \, \mathrm{d}y = \int \left\| \tilde{\ell}_{\theta}(y, x) q_{\theta}^{1/2}(y, x) \right\|^2 \, \mathrm{d}y < \infty. \tag{50}$$

Hence by Proposition 2.29 in van der Vaart (1998) we have that

$$\lim_{n\to\infty} \int \left\| \tilde{\ell}_{\vartheta_n}(y, x_n) q_{\vartheta_n}^{1/2}(y, x_n) - \tilde{\ell}_{\theta}(y, x) q_{\theta}^{1/2}(y, x) \right\|^2 \, \mathrm{d}y = 0. \tag{51}$$

Since this also applies with $\vartheta_n = \theta$ we may conclude that

$$\mathscr{Q}_n(x) := \int \left\| \tilde{\ell}_{\vartheta_n}(y, x) q_{\vartheta_n}^{1/2}(y, x) - \tilde{\ell}_{\theta}(y, x) q_{\theta}^{1/2}(y, x) \right\|^2 \, \mathrm{d}y \tag{52}$$

converges continuously to the zero function. By Corollary A.4 and the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1) it follows that $\mathscr{Q}_n(\tilde{X}_n) \rightsquigarrow 0$ where $\tilde{X}_n$ has law $G_{\vartheta_n,n}$. We next show that $\mathscr{Q}_n(\tilde{X}_n)_{n\in\mathbb{N}}$ is uniformly integrable. Dominate the n-th term by

$$\begin{aligned} \mathscr{Q}_n(\tilde{X}_n) &\le 2 \left[\int \left\| \tilde{\ell}_{\vartheta_n}(y, \tilde{X}_n) \right\|^2 q_{\vartheta_n}(y, \tilde{X}_n) \, \mathrm{d}y + \int \left\| \tilde{\ell}_{\theta}(y, \tilde{X}_n) \right\|^2 q_{\theta}(y, \tilde{X}_n) \, \mathrm{d}y \right] \\ &\le 2 \left[\mathbb{E} \left[\left\| \tilde{\ell}_{\vartheta_n}(\tilde{Y}_n, \tilde{X}_n) \right\|^2 \middle| \tilde{X}_n \right] + \mathbb{E} \left[\left\| \tilde{\ell}_{\theta}(\tilde{Y}, \tilde{X}_n) \right\|^2 \middle| \tilde{X}_n \right] \right], \end{aligned}$$

where $\tilde{Y}_n$ and $\tilde{Y}$ have laws such that their conditional density given $\tilde{X}_n$ is q_{ϑ_n} and q_{θ} respectively. Under Assumption 3.3.12 and using Lemma A.3 and the forms given in Lemma 3.4.3 it is easily seen that each of $\left(\left\| \tilde{\ell}_{\vartheta_n}(\tilde{Y}_n, \tilde{X}_n) \right\|^2 \right)_{n\in\mathbb{N}}$ and $\left(\left\| \tilde{\ell}_{\theta}(\tilde{Y}, \tilde{X}_n) \right\|^2 \right)_{n\in\mathbb{N}}$ are uniformly integrable. The uniform integrability of the corresponding conditional expectations above then follows from Jensen's inequality for conditional expectations and the de la Vallée - Poussin criterion for uniform integrability (e.g. Bogachev, 2007, Theorem 4.5.9). We may now conclude that

$$\lim_{n\to\infty} \int \left\| \tilde{\ell}_{\vartheta_n} q_{\vartheta_n}^{1/2} \bar{\rho}_{\vartheta_n,n}^{-1/2} - \tilde{\ell}_{\theta} q_{\theta}^{1/2} \bar{\rho}_{\vartheta_n,n}^{-1/2} \right\|^2 \, \mathrm{d}\lambda = 0,.$$

Using this result twice (once with $\vartheta_n = \theta_n$ and once with $\vartheta_n = \tilde{\theta}_n$) we obtain (49). Combination of (48) and (49) with the continuity of the inner product yields

$$\lim_{n\to\infty} \left\langle \tilde{\ell}_{\tilde{\theta}_n} q_{\tilde{\theta}_n}^{1/2} \bar{\rho}_{\theta_n,n}^{-1/2}, \sqrt{n} \left(q_{\theta_n}^{1/2} - q_{\tilde{\theta}_n}^{1/2} \right) \bar{\rho}_{\theta_n,n}^{-1/2} \right\rangle_{\lambda} - \left\langle \tilde{\ell}_{\theta_n} q_{\theta_n}^{1/2} \bar{\rho}_{\theta_n,n}^{-1/2}, -\frac{1}{2} b_n' \dot{\ell}_{\theta_n} q_{\theta_n}^{1/2} \bar{\rho}_{\theta_n,n}^{-1/2} \right\rangle_{\lambda} = 0.$$

Since

$$\int \sqrt{n} m_n \bar{\rho}_{\theta_n,n} \, \mathrm{d}\lambda = \left\langle \tilde{\ell}_{\tilde{\theta}_n} q_{\tilde{\theta}_n}^{1/2} \bar{\rho}_{\theta_n,n}^{-1/2}, \sqrt{n} \left(q_{\theta_n}^{1/2} - q_{\tilde{\theta}_n}^{1/2} \right) \bar{\rho}_{\theta_n,n}^{-1/2} \right\rangle_{\lambda}$$

and

$$\left\langle \tilde{\ell}_{\theta_n} q_{\theta_n}^{1/2} \bar{\rho}_{\theta_n,n}^{-1/2}, -\frac{1}{2} b_n' \dot{\ell}_{\theta_n} q_{\theta_n}^{1/2} \bar{\rho}_{\theta_n,n}^{-1/2} \right\rangle_{\lambda} = -\frac{1}{2} \tilde{I}_{n,\theta_n} (0', b_n')'.$$

Combining these displays allows us to conclude that to establish that $R_{2,n} \to 0$ in $P_{\theta_n}^n$-probability it will suffice to show the same for $R_{2,n}' := \frac{1}{\sqrt{n}} \sum_{t=1}^n Z_{n,t,1} - m_n(X_t) - Z_{n,t,2}$.

As is easy to verify, $(Z_{n,t,1}-m_n(X_t)-Z_{n,t,2}, \mathcal{F}_{n,t})_{n\in\mathbb{N},1\leq t\leq n}$ forms a martingale difference array with $\mathcal{F}_{n,t} := \sigma(\epsilon_1,\dots,\epsilon_t)$. It follows that it suffices to show that (under $P_{\theta_n}^n$)

$$\mathbb{V}\left[\frac{1}{\sqrt{n}}\sum_{t=1}^n Z_{n,t,1} - m_n(X_t) - Z_{n,t,2}\right] = \frac{1}{n}\sum_{t=1}^n \mathbb{V}\left[Z_{n,t,1} - m_n(X_t) - Z_{n,t,2}\right] \to 0.$$

In view of (49) for this, it suffices to show that

$$\frac{1}{n}\sum_{t=1}^n \mathbb{E}[\|m_n(X_t)\|^2] = \int \|m_n(x)\|^2 \bar{\rho}_{\theta_n,n}(x)\,\mathrm{d}x \to 0.$$

For this, define $\mathsf{m}_n(y,x) := \tilde{\ell}_{\tilde{\theta}_n}(y,x) q_{\tilde{\theta}_n}(y,x)^{1/2}\left[q_{\theta_n}(y,x)^{1/2} - q_{\tilde{\theta}_n}(y,x)^{1/2}\right]$ and note that $\mathsf{m}_n(y,x_n) \to 0$ pointwise for any convergent $x_n \to x$. We additionally have by Cauchy-Schwarz that

$$\left\|\int \mathsf{m}_n(y,x_n)\,\mathrm{d}y\right\| \leq \left[\int \|\tilde{\ell}_{\tilde{\theta}_n}\|^2 q_{\tilde{\theta}_n}(y,x_n)\,\mathrm{d}y\right]^{1/2}\left[\int \left(q_{\theta_n}(y,x_n)^{1/2} - q_{\tilde{\theta}_n}(y,x_n)^{1/2}\right)^2\mathrm{d}y\right]^{1/2}.$$

As can be easily verified, $\int \|\tilde{\ell}_{\tilde{\theta}_n}\|^2 q_{\tilde{\theta}_n}(y,x_n)\,\mathrm{d}y$ is upper bounded by a term of the form $M_1 + M_2\|x_n\|^2$ (with M_1, M_2 finite positive constants which do not depend on n). Additionally $(q_{\theta_n}(y,x_n)^{1/2} - q_{\tilde{\theta}_n}(y,x_n)^{1/2})^2 \to 0$ pointwise in y and is upper bounded by $2q_{\theta_n}(y,x_n) + 2q_{\tilde{\theta}_n}(y,x_n)$ which satisisfies $\int 2q_{\theta_n}(y,x_n) + 2q_{\tilde{\theta}_n}(y,x_n)\,\mathrm{d}y = 4 = \int 4q_\theta(y,x)\,\mathrm{d}y$ for each $n \in \mathbb{N}$. Therefore, by the generalised Lebesgue dominated convergence theorem

$$\int \left(q_{\theta_n}(y,x_n)^{1/2} - q_{\tilde{\theta}_n}(y,x_n)^{1/2}\right)^2\mathrm{d}y \to 0.$$

It follows that $\|m_n(x_n)\|^2 \to 0$ pointwise for any $x_n \to x$. Re-using the bound from above, we note that

$$\|m_n(X_t)\|^2 \leq 4(M_1 + M_2\|X_t\|^2)$$

and hence $\|m_n(X_t)\|^2$ is $G_{\theta_n,n}$-uniformly integrable by Lemma A.3.[38] Moreover, by corollary A.4, $G_{\theta_n,n} \rightsquigarrow \pi_\theta$ and hence by Theorem 3.5 in Serfozo (1982)

$$\frac{1}{n}\sum_{t=1}^n \mathbb{E}[\|m_n(X_t)\|^2] = \int \|m_n\|^2\,\mathrm{d}G_{\theta_n,n} \to \int 0\,\mathrm{d}\pi_\theta = 0.$$

This establishes that $R_{2,n} \xrightarrow{P_{\theta_n}^n} 0$. For $R_{1,n}$, define $f_n(y,x) := c_n q_{\theta_n}(y,x)^{1/2} q_{\tilde{\theta}_n}(y,x)^{1/2}\bar{\rho}_{\theta_n,n}(x)$, where

$$c_n^{-1} := \int q_{\theta_n}^{1/2} q_{\tilde{\theta}_n}^{1/2}\bar{\rho}_{\theta_n,n}\,\mathrm{d}\lambda = 1 - \frac{1}{2}\int (q_{\theta_n}^{1/2} - q_{\tilde{\theta}_n}^{1/2})^2\bar{\rho}_{\theta_n,n}\,\mathrm{d}\lambda.$$

[38] Lemma A.3 bounds the moments of the (demeaned) stationary solution; it is easy to see that this provides a uniform (in t, n) upper bound for our process (conditional on the initial values).

We have

$$-n\left(q_{\theta_n}^{1/2}-q_{\tilde{\theta}_n}^{1/2}\right)^2=\left(\sqrt{n}\left[q_{\tilde{\theta}_n}^{1/2}-q_{\theta_n}^{1/2}\right]-\frac{1}{2}b_n'\dot{\ell}_{\theta_n}q_{\theta_n}^{1/2}\right)^2+\left(\frac{1}{2}b_n'\dot{\ell}_{\theta_n}q_{\theta_n}^{1/2}\right)^2$$
$$-b_n'\dot{\ell}_{\theta_n}q_{\theta_n}^{1/2}\sqrt{n}\left(q_{\tilde{\theta}_n}^{1/2}-q_{\theta_n}^{1/2}\right),$$

and so by Lemma A.5 and the continuity of the inner product

$$\begin{aligned}\int(q_{\theta_n}^{1/2}-q_{\tilde{\theta}_n}^{1/2})^2\bar{\rho}_{\theta_n,n}\,\mathrm{d}\lambda&=\frac{1}{n}\int b_n'\dot{\ell}_{\theta_n}q_{\theta_n}^{1/2}\bar{\rho}_{\theta_n,n}^{1/2}\sqrt{n}\left(q_{\tilde{\theta}_n}^{1/2}-q_{\theta_n}^{1/2}\right)\bar{\rho}_{\theta_n,n}^{1/2}\,\mathrm{d}\lambda\\&\quad-\frac{1}{n}\int\left(\frac{1}{2}b_n'\dot{\ell}_{\theta_n}q_{\theta_n}^{1/2}\right)^2\bar{\rho}_{\theta_n,n}\,\mathrm{d}\lambda+o(n^{-1})\\&=\frac{1}{2}(n^{-1/2}b_n)'\dot{I}_{\theta_n}(n^{-1/2}b_n)+o(n^{-1}),\end{aligned}$$

where $\dot{I}_{\theta_n}:=\int\dot{\ell}_{\theta_n}\dot{\ell}_{\theta_n}'q_{\theta_n}\bar{\rho}_{\theta_n,n}\,\mathrm{d}\lambda$.[39] It follows that $c_n^{-1}=1-a_n$ with $a_n\to 0$ and $na_n=\frac{1}{4}b_n'\dot{I}_{\theta_n}b_n+o(1)$. By Taylor's theorem $\log(1-a_n)=-a_n+R(1-a_n)a_n^2$ with $R_n(1-x)\to 0$ as $x\to 0$. Hence $\log c_n^n=-n\log(1-a_n)=na_n-nR(1-a_n)a_n^2=\frac{1}{4}b_n'\dot{I}_{\theta_n}b_n+o(1)$. $P_{\theta_n}^n$ is the measure corresponding to the product density $\prod_{t=1}^n q_{\theta_n}\rho_{\theta_n,t}$. Let Q_n^n be the measure corresponding to the product density $\prod_{t=1}^n c_nq_{\theta_n}^{1/2}q_{\tilde{\theta}_n}^{1/2}\bar{\rho}_{\theta_n,n}$. Writing $\Lambda_n:=\Lambda_n(Q_n^n,P_{\theta_n}^n)$ for their log-likelihood ratio and using notation from the proof of Proposition 3.4.1, by (35)

$$\Lambda_n=\log c_n^n+2\sum_{t=1}^n\log(Z_{n,t}+1)\overset{P_{\theta_n}^n}{\rightsquigarrow}Z,$$

where Z has a normal distribution. By Example 6.5 in van der Vaart (1998) $P_{\theta_n}^n\lhd Q_n^n$ and so by Le Cam's first lemma (e.g. van der Vaart, 1998, Lemma 6.4) it suffices to show that $R_{n,1}\xrightarrow{Q_n^n}0$. For this we first note that if

$$\frac{1}{n}\sum_{t=1}^n\tilde{\ell}_{\theta_n}\ell_{\theta_n}'-\tilde{I}_{n,\theta_n}\xrightarrow{Q_n^n}0,\tag{53}$$

then we have $R_{n,1}\xrightarrow{Q_n^n}0$ as

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^n\int\left\|\tilde{\ell}_{\tilde{\theta}_n}\sqrt{n}\left(1-\frac{q_{\tilde{\theta}_n}^{1/2}}{q_{\theta_n}^{1/2}}\right)+\frac{1}{2}\sqrt{n}\tilde{\ell}_{\theta_n}\dot{\ell}_{\theta_n}'(b_n/\sqrt{n})\right\|c_nq_{\theta_n}^{1/2}q_{\tilde{\theta}_n}^{1/2}\bar{\rho}_{\theta_n,n}\,\mathrm{d}\lambda\\&\quad\le c_n\int\left\|\tilde{\ell}_{\tilde{\theta}_n}q_{\tilde{\theta}_n}^{1/2}\right\|\sqrt{n}\left|q_{\tilde{\theta}_n}^{1/2}-q_{\theta_n}^{1/2}-\frac{1}{2\sqrt{n}}b_n'\dot{\ell}_{\theta_n}q_{\theta_n}^{1/2}\right|\bar{\rho}_{\theta_n,n}\,\mathrm{d}\lambda\\&\quad=o(1),\end{aligned}$$

where the convergence follows from Lemma A.5 and the continuity of the inner product. It

[39] This sequence of matrices is bounded (see e.g. intermediate results used in the proof of Proposition 3.4.1).

remains to prove (53). For this it suffices to observe that

$$Q_n^n \left\| \frac{1}{n} \sum_{t=1}^{n} \tilde{\ell}_{\theta_n} \dot{\ell}_{\theta_n}' - \tilde{I}_{n,\theta_n} \right\| \leq |c_n - 1| \int \left\| \tilde{\ell}_{\theta_n} \dot{\ell}_{\theta_n}' \right\| q_{\theta_n}^{1/2} q_{\tilde{\theta}_n}^{1/2} \bar{\rho}_{\theta_n,n} \, \mathrm{d}\lambda + \int \left\| \tilde{\ell}_{\theta_n} \dot{\ell}_{\theta_n}' \right\| q_{\theta_n}^{1/2} |q_{\tilde{\theta}_n}^{1/2} - q_{\theta_n}^{1/2}| \bar{\rho}_{\theta_n,n} \, \mathrm{d}\lambda$$
$$= o(1),$$

by Cauchy-Schwarz and the facts that $\sup_{n\in\mathbb{N}} \| \|\tilde{\ell}_{\theta_n} \dot{\ell}_{\theta_n}' \| q_{\theta_n}^{1/2} \bar{\rho}_{\theta_n,n}^{1/2} \|_{\lambda,2} < \infty$ (under assumption 3.3.1), $\|q_{\tilde{\theta}_n}^{1/2} \bar{\rho}_{\theta_n,n}^{1/2}\|_{\lambda,2} = 1$, $c_n \to 1$ and $\| |q_{\theta_n}^{1/2} - q_{\tilde{\theta}_n}^{1/2}| \bar{\rho}_{\theta_n,n}^{1/2} \|_{\lambda,2} \to 0$ by Lemma A.5. □

Lemma A.9. *Suppose assumption 3.3.1 holds and let $\Delta_\theta(u)$ be as defined as in Lemma A.5. If $\theta_n = (\gamma_n, \eta) \to (\gamma, \eta) = \theta$, the sequence $(\Delta_{\theta_n}(u))_{n\in\mathbb{N}}$ has uniformly bounded $4+\delta$ moments under $P_{\theta_n}^n$, i.e.*

$$\sup_{n\in\mathbb{N}, 1\leq t\leq n} \int |\Delta_{\theta_n}(u)|^{4+\delta} \, \mathrm{d}P_{\theta_n}^n < \infty.$$

In consequence, it is uniformly square $P_{\theta_n}^n$-integrable.

Proof. The continuity and continuous differentiability conditions in assumption 3.3.1 ensure that all non-random terms in the expressions (3.7) – (3.9) converge and are thus bounded.[40] Note that under $P_{\theta_n}^n$, $A_{k\bullet} V_{\theta_n,t} \sim \eta_k$ and is independent of both X_t and $A_{j\bullet} V_{\theta_n,t}$ for $j \neq k$. Given this independence and the forms given in (3.7) – (3.9) it suffices to show that

$$\mathbb{E}[|\phi_k(\epsilon_k)|^{4+\delta}] < \infty, \quad \mathbb{E}[|\epsilon_k|^{4+\delta}] < \infty, \quad \sup_{n\in\mathbb{N}, 1\leq t\leq n} \mathbb{E}\|X_t\|^{4+\delta} < \infty,$$

where the expectation is taken under $P_{\theta_n}^n$ and we note that η_k does not depend on n. The first two of these follow immediately from the moment assumptions in part 2 of assumption 3.3.1. For the last term, by recursing backwards we obtain

$$Z_t = \sum_{j=0}^{t-1} \mathsf{B}_\theta^j \mathsf{C}_\theta + \sum_{j=0}^{t-1} \mathsf{B}_\theta^j \mathsf{D}_\theta \epsilon_{t-j} + \mathsf{B}_\theta^t Z_0. \tag{54}$$

Assumption 3.3.11 ensures that the matrices B_θ^j are absolutely summable and $\sum_{j=0}^{\infty} \mathsf{B}_\theta^j = (I - \mathsf{B}_\theta)^{-1}$ exists (e.g. Lütkepohl, 2005, Section A.9.1). Moreover, $\mathsf{T} := \{\theta_n : n \in \mathbb{N}\} \cup \{\theta\}$ is compact, and the spectral radius $M \mapsto \rho(M)$ is a continuous function, then $\{\rho(\mathsf{B}_\vartheta) : \vartheta \in \mathsf{T}\}$ is compact, which ensures that this set is bounded above by some $\upsilon < 1$. Let $m_1, m_2 \in \mathbb{N}$ with $m_2 \geq m_1$ and let $\mathsf{B}_\vartheta = U_\vartheta^* J_\vartheta U_\vartheta$ be a Schur decomposition of B_ϑ (see e.g. Horn and Johnson, 2013, Theorem 2.3.1). Let $\|\cdot\|$ be the spectral norm and note that we have $\|U_\vartheta\| = 1$ and hence by Lemma 5.6.10 in Horn and Johnson (2013), for any $\varepsilon > 0$ with $\upsilon + \varepsilon < 1$ we have

$$\left\| \sum_{j=0}^{m_2} \mathsf{B}_\vartheta^j - \sum_{j=0}^{m_1} \mathsf{B}_\vartheta^j \right\| \leq \sum_{j=m_1}^{m_2} \|\mathsf{B}_\vartheta^j\| \leq \sum_{j=m_1}^{m_2} \|J_\vartheta\|^j \leq \sum_{j=m_1}^{m_2} (\upsilon + \varepsilon)^j.$$

[40] These terms are of the form 1, $A(\alpha_n, \sigma_n) D_{b_l}$ (with l an integer) or $[D_{x_l}(\alpha_n, \sigma_n)]_{k\bullet} [A(\alpha_n, \sigma_n)^{-1}]_{\bullet j}$ for $k, j \in \{1, \ldots, K\}$ and $x \in \{\alpha, \sigma\}$ (with l an integer).

Since $\sum_{j=0}^{\infty}(v+\varepsilon)^j = \frac{1}{1-v-\varepsilon} < \infty$, in view of the preceding display, the convergence $\sum_{j=0}^{\infty} \mathsf{B}_\theta^j = (I - \mathsf{B}_\theta)^{-1}$ is uniform in $\theta \in \mathsf{T}$. Since $\theta \mapsto \mathsf{C}_\theta$ is continuous, this immediately implies that $\sup_{n\in\mathbb{N},1\leq t\leq n} \left\| \sum_{j=0}^{t-1} \mathsf{B}_{\theta_n}^j \mathsf{C}_{\theta_n} \right\|^{4+\delta} < \infty$. Similarly using the same uniform bound, that $(\epsilon_t)_{t\geq 1}$ are i.i.d. and since $\theta \mapsto \|\mathsf{D}_\theta\|$ is continuous we have that

$$\sup_{n\in\mathbb{N},1\leq t\leq n} \mathbb{E} \left\| \sum_{j=0}^{t-1} \mathsf{B}_{\theta_n}^j \mathsf{D}_{\theta_n} \epsilon_{t-j} \right\|^{4+\delta} \leq \sup_{n\in\mathbb{N}} \|\mathsf{D}_{\theta_n}\|^{4+\delta} \mathbb{E}\|\epsilon_t\|^{4+\delta} \sup_{n\in\mathbb{N},1\leq t\leq n} \sum_{j=0}^{t-1} \|B_{\theta_n}^j\|^{4+\delta} < \infty.$$

Hence by Minkowski's inequality we have that $\sup_{n\in\mathbb{N},1\leq t\leq n} \mathbb{E}\|X_t\|^{4+\delta} < \infty$, where the expectation is taken under $P_{\theta_n}^n$. □

Lemma A.10. *Suppose assumption 3.3.1 holds and let $\Delta_\theta(u)$ be as defined as in Lemma A.5. If $\theta_n = (\gamma_n, \eta) \to (\gamma, \eta) = \theta$ and G_θ is defined as in Lemma A.11, then under $P_{\theta_n}^n$,*

$$\lim_{n\to\infty} \mathbb{E} \left| \frac{1}{n} \sum_{t=1}^{n} \Delta_{\theta_n}(u)(Y_t, X_t)^2 - \tau^2 \right|^2 = 0, \qquad \textit{with } \tau^2 := G_\theta \Delta_\theta(u)^2 < \infty.$$

Proof. Let ϑ_n indicate either θ_n or θ. By inspection of their forms it is clear that for any $x_n \to x$, $[\Delta_{\vartheta_n}(u)(y, x_n)] q_{\vartheta_n}(y, x_n)^{1/2} \to [\Delta_\theta(u)(y,x)] q_\theta(y,x)^{1/2}$ pointwise in y. By inspection of their form, the continuity given by Assumption 3.3.1 we have

$$\lim_{n\to\infty} \int [\Delta_{\vartheta_n}(u)(y, x_n)]^2 q_{\vartheta_n}(y, x_n)\,\mathrm{d}y = \int [\Delta_\theta(u)(y,x)]^2 q_\theta(y,x)\,\mathrm{d}y < \infty,$$

i.e. $\mathscr{Q}_n$ converges continuously to $\mathscr{Q}$ where

$$\mathscr{Q}_n(x) := \int [\Delta_{\vartheta_n}(u)(y,x)]^2 q_{\vartheta_n}(y,x)\,\mathrm{d}y, \quad \mathscr{Q}(x) := \int [\Delta_\theta(u)(y,x)]^2 q_\theta(y,x)\,\mathrm{d}y.$$

We can bound

$$\mathscr{Q}_n(x) \leq 2\left[\mathbb{E}\left[\Delta_{\theta_n}(u)(\tilde{Y}_n, \tilde{X}_n)^2 \middle| \tilde{X}_n\right] + \mathbb{E}\left[\Delta_\theta(u)(\tilde{Y}, \tilde{X}_n)^2 \middle| \tilde{X}_n\right]\right],$$

where $\tilde{Y}_n$ and $\tilde{Y}$ have laws such that their conditional density given $\tilde{X}_n$ is q_{ϑ_n} and q_θ respectively. Under Assumption 3.3.12 and an argument similar to that of Lemma A.9 it is easily seen that each of $\left(\Delta_{\theta_n}(u)(\tilde{Y}_n, \tilde{X}_n)^2\right)_{n\in\mathbb{N}}$ and $\left(\Delta_\theta(u)(\tilde{Y}, \tilde{X}_n)^2\right)_{n\in\mathbb{N}}$ are uniformly integrable. The uniform integrability of the corresponding conditional expectations above then follows from Jensen's inequality for conditional expectations and the de la Vallée - Poussin criterion for uniform integrability (e.g. Bogachev, 2007, Theorem 4.5.9). Corollary A.4 in conjunction with Theorem 3.5 of Serfozo (1982) then yields that $G_{\theta_n,\theta_n,n}\Delta_{\theta_n}(u)^2 \to G_\theta \Delta_\theta(u)^2 < \infty$, where $G_{\vartheta,\theta,n}(A) := \int\int \mathbf{1}_A(y,x) q_\vartheta(y,x)\,\mathrm{d}y\,\mathrm{d}\bar{\rho}_{\theta,n}(x)$ and the finiteness follows from the form of $\Delta_\theta(u)$, assumption 3.3.12 and Lemma A.3.[41] The convergence follows on combining Lemma A.9, Proposition A.2 and Corollary 19.3(ii) of Davidson (1994). □

Lemma A.11. *Suppose that assumption 3.3.1 holds. Let $\rho_{\theta,t}$ be the density of $\mathsf{X}_t :=$ $\mathrm{vec}(Y_{t-1}, \ldots, Y_{t-p})$ (i.e. the non-deterministic parts of X_t) under θ. Let $G_{\vartheta,n}$ be the*

[41] $\bar{\rho}_{\theta,n} := \frac{1}{n}\sum_{t=1}^{n} \rho_{\theta,t}$.

measure defined by $G_{\vartheta,n}(A) := \int_A q_\vartheta \frac{1}{n}\sum_{t=1}^n \rho_{\theta,t}\,\mathrm{d}\lambda$ and G_θ the measure defined by $G_\theta(A) := \int_A q_\theta(y,x)\,\mathrm{d}(\lambda(y)\otimes\pi_\theta(x))$ for q_ϑ as defined as in (40). *Let $(\vartheta_n)_{n\in\mathbb{N}} \subset \Theta$ be an sequence with $\vartheta_n = (\gamma_n,\eta) \to (\gamma,\eta) = \theta$. Then, $G_{\vartheta_n,n} \xrightarrow{TV} G_\theta$.*

Proof. By the form of $\theta \mapsto q_\theta$ we have that $q_{\vartheta_n} \to q_\theta$ (pointwise) as $n \to \infty$. Hence, for any x, $q_{\vartheta_n}(\cdot,x) \to q_\theta(\cdot,x)$ pointwise and since each $q_{\vartheta_n}(\cdot,x)$ and $q_\theta(\cdot,x)$ is a probability density with respect to Lebesgue measure, by Proposition 2.29 in van der Vaart (1998),

$$\mathscr{Q}_n(x) := \int |q_{\vartheta_n}(y,x) - q_\theta(y,x)|\,\mathrm{d}y \to 0,$$

pointwise in x. Let $(\psi_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions on $\mathbb{R}^{Kp}$ with $\psi_n \in [0,1]$ and $\pi_{\theta,n}$ the probability measure corresponding to the density $\frac{1}{n}\sum_{t=1}^n \rho_{\theta,t}$. Then

$$\left|\int\int \psi_n(q_{\vartheta_n}(y,x) - q_\theta(y,x))\,\mathrm{d}y\,\mathrm{d}\pi_{\theta,n}(x)\right| \leq \int \mathscr{Q}_n(x)\,\mathrm{d}\pi_{\theta,n}(x).$$

Since $\mathscr{Q}_n(x) \leq \int q_{\vartheta_n}(y,x)\,\mathrm{d}y + \int q_\theta(y,x)\,\mathrm{d}y$ and $\int\left[\int q_{\vartheta_n}(y,x)\,\mathrm{d}y + \int q_\theta(y,x)\,\mathrm{d}y\right]\,\mathrm{d}\pi_{\theta,n}(x) = 2 = \int 2\int q_\theta(y,x)\,\mathrm{d}y\,\mathrm{d}\pi_\theta(x)$, the $\mathscr{Q}_n(x)$ are uniformly $\pi_{\theta,n}$ – integrable.[42] Hence by Proposition A.2 and Corollary 2.9 of Feinberg et al. (2016), $\int \mathscr{Q}_n(x)\,\mathrm{d}\pi_{\theta,n}(x) \to 0$. □

Lemma A.12. *Let $\gamma_n = (\alpha_n,\beta) \to (\alpha,\beta) = \gamma$ and $\theta_n = (\gamma_n,\eta) \to (\gamma,\eta) = \theta$ for $\gamma_n,\gamma \in \Gamma$. Let $\tilde\gamma_n = (\alpha_n,\beta_n) \to (\alpha,\beta) = \gamma$, $\tilde\theta_n := (\tilde\gamma_n,\eta) \to (\gamma,\eta) = \theta$ with $b_n := \sqrt{n}(\beta_n - \beta) \to b$ and $\check\theta_n := (\tilde\gamma_n,\tilde\eta_n) \to \theta$ with $\tilde\eta_n := \eta(1 + h_n/\sqrt{n})$ for $h_n \to h$ in $\mathscr{H}$. Then, under the conditions of Theorem 3.5.1,*

1. *If $Z_{n,1} := \frac{1}{\sqrt{n}}\sum_{t=1}^n \tilde\ell_{\theta_n}(Y_t,X_t)$ and $Z_{n,2} := \Lambda^n_{\check\theta_n\backslash\theta_n}(Y^n)$, then under $P^n_{\theta_n}$,*

$$Z_n \rightsquigarrow Z \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\sigma^2_{b,h}\end{pmatrix}, \begin{pmatrix} \tilde I_\theta & \tilde I_\theta(0',b')' \\ (0',b')\tilde I_\theta & \sigma^2_{b,h}\end{pmatrix}\right).$$

2. *We have that*

$$\frac{1}{n}\sum_{t=1}^n \left(\hat\ell_{\tilde\theta_n}(Y_t,X_t) - \tilde\ell_{\tilde\theta_n}(Y_t,X_t)\right) = o_{P^n_{\tilde\theta_n}}(n^{-1/2})$$

3. *$\tilde I_{n,\theta_n} \to \tilde I_\theta := G_\theta\dot\ell_\theta\dot\ell_\theta'$ and $P^n_{\tilde\theta_n}\left(\|\hat I_{n,\tilde\theta_n} - \tilde I_\theta\|_2 < \nu_n\right) \to 1$ where ν_n is defined in Assumption 3.3.2 and G_θ in Lemma A.11.*

4. *We have that*

$$R_n := \frac{1}{\sqrt{n}}\sum_{t=1}^n \left[\tilde\ell_{\tilde\theta_n}(Y_t,X_t) - \tilde\ell_{\theta_n}(Y_t,X_t)\right] + \tilde I_{n,\theta_n}(0',\sqrt{n}(\beta_n - \beta)')' \xrightarrow{P^n_{\theta_n}} 0.$$

[42] The uniform integrability follows since, by the integral equality, Proposition A.2 and Proposition A.16,

$$\int\left|\int q_{\vartheta_n}(y,x)\,\mathrm{d}y + \int q_\theta(y,x)\,\mathrm{d}y - 2\int q_\theta(y,x)\,\mathrm{d}y\right|\,\mathrm{d}\pi_{\theta,n}(x) \to 0.$$

Proof. For part (i), let $z_{n,t}$ be

$$z_{n,t} := \left(\tilde{\ell}_{\theta_n}(Y_t, X_t)', c'\dot{\ell}_{\theta_n}(Y_t, X_t) + \sum_{k=1}^{K} h_k(A_{k\bullet}V_{\theta_n,t}) \right)',$$

and $\mathcal{F}_{n,t} := \sigma(\epsilon_1, \dots, \epsilon_t)$. Under assumption 3.3.12, $\mathbb{E}\|z_{n,t}\|_2^2 < \infty$ and $\{z_{n,t}, \mathcal{F}_{n,t} : 1 \leq t \leq n, n \in \mathbb{N}\}$ is a martingale difference array (all under $P^n_{\theta_n}$) such that

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[z_{n,t}z_{n,t}'\right] = \begin{bmatrix} \tilde{I}_{n,\theta_n} & \tilde{I}_{n,\theta_n}(0', b_n')' \\ (0', b_n')\tilde{I}_{n,\theta_n} & \sigma^2_{n,b,h} \end{bmatrix} \to \begin{bmatrix} \tilde{I}_{\theta} & \tilde{I}_{\theta}(0', b')' \\ (0', b')\tilde{I}_{\theta} & \sigma^2_{b,h} \end{bmatrix},$$

noting Lemma 3.4.3 and Theorem 12.14 of Rudin (1991). That $\sigma^2_{n,b,h}$ converges to a $\sigma^2_{b,h}$ is part of the conclusion of Proposition 3.4.1. That $\tilde{I}_{n,\theta_n} \to \tilde{I}_{\theta}$ follows from Lemma A.10. Moreover, the Lindeberg condition in (67) is satisfied since $\{\|z_{n,t}\|^2 : 1 \leq t \leq n, n \in \mathbb{N}\}$ is uniformly $P^n_{\theta_n}$-integrable. That this is true for $\|z_{n,t,2}\|^2$ follows from A.9. That it is also true for $\|z_{n,t,1}\|^2$ can be shown by an analogous argument. Part (i) then follows from Propositions 3.4.1, A.17 and Lemma A.10.

Next, define $A_n := A_{\tilde{\theta}_n}$ and $B_n := B_{\tilde{\theta}_n}$ and note that each $A_{n,k}(Y_t - c_n - B_nX_t) \approx \epsilon_{k,t} \sim \eta_k$ under $P^n_{\tilde{\theta}_n}$. Hence we can compute certain properties of the efficient score using the equality in distribution:

$$\tilde{\ell}_{\tilde{\theta}_n,\alpha_l}(Y_t, X_t) \approx \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\alpha}_{l,k,j}\phi_k(\epsilon_{k,t})\epsilon_{j,t} + \sum_{k=1}^{K}\zeta^{\alpha}_{n,l,k,k}\left[\tau_{k,1}\epsilon_{k,t} + \tau_{k,2}\kappa(\epsilon_{k,t})\right] \tag{55}$$

$$\tilde{\ell}_{\tilde{\theta}_n,\sigma_l}(Y_t, X_t) \approx \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta^{\sigma}_{n,l,k,j}\phi_k(\epsilon_{k,t})\epsilon_{j,t} + \sum_{k=1}^{K}\zeta^{\sigma}_{l,k,k}\left[\tau_{k,1}\epsilon_{k,t} + \tau_{k,2}\kappa(\epsilon_{k,t})\right] \tag{56}$$

$$\tilde{\ell}_{\tilde{\theta}_n,b_l}(Y_t, X_t) \approx \sum_{k=1}^{K} -A_{n,k\bullet}D_{b,l}\left[\phi_k(\epsilon_{k,t})(X_t - \mathbb{E}X_t) - \mathbb{E}X_t\left(\varsigma_{k,1}\epsilon_{k,t} + \varsigma_{k,2}\kappa(\epsilon_{k,t})\right)\right] \tag{57}$$

where we note that the same observation implies that $\tau_{k,n} = \tau_k$ and $\varsigma_{k,n} = \varsigma_k$ for each n.[43] By our assumptions on the map $(\alpha, \sigma) \mapsto A(\alpha, \sigma)$, we have $\zeta^{\alpha}_{n,l,k,j} \to \zeta^{\alpha}_{\infty,l,k,j} := [D_{\alpha_l}(\alpha_0, \sigma)]_{k\bullet}A(\alpha, \sigma)^{-1}_{\bullet j}$ and $\zeta^{\sigma}_{n,l,k,j} \to \zeta^{\sigma}_{\infty,l,k,j} := [D_{\sigma_l}(\alpha, \sigma)]_{k\bullet}A(\alpha, \sigma)^{-1}_{\bullet j}$. Note that the entries of $D_{b,l}$ are all zero except for entry l (corresponding to b_l) which is equal to one.

We verify (ii) for each component of the efficient score (55)-(57). Components (55) and (56) follow similarly and we focus on (55). We define

$$\varphi_{1,n,t} := \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta_{l,k,j,n}\phi_k(A_{n,k\bullet}V_{n,t})A_{n,j\bullet}V_{n,t},$$

and

$$\hat{\varphi}_{1,n,t} := \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K} \zeta_{l,k,j,n}\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t})A_{n,j\bullet}V_{n,t},$$

[43] In the preceding display we have written $\zeta^{\alpha}_{n,l,k,k}$ and $\zeta^{\sigma}_{n,l,k,k}$ rather than $\zeta^{\alpha}_{l,k,k}$ and $\zeta^{\sigma}_{l,k,k}$ to indicate their dependence on $\tilde{\theta}_n$.

with $V_{n,t} = Y_t - B_n X_t$, and let $\overline{\zeta}_n := \max_{l\in[L],j\in[K],k\in[K]} |\zeta^{\alpha}_{l,j,k,n}|$ which converges to $\overline{\zeta} := \max_{l\in[L],j\in[K],k\in[K]} |\zeta^{\alpha}_{l,j,k,\infty}| < \infty$. We have that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\hat{\varphi}_{1,n,t}-\varphi_{1,n,t}) \leq \sqrt{n}\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\overline{\zeta}_n \left|\frac{1}{n}\sum_{t=1}^{n}\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t})A_{n,j\bullet}V_{n,t} - \phi_k(A_{n,k\bullet}V_{n,t})A_{n,j\bullet}V_{n,t}\right|,$$

Since each $\left|\frac{1}{n}\sum_{t=1}^{n}\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t})A_{n,j\bullet}V_{n,t} - \phi_k(A_{n,k\bullet}V_{n,t})A_{n,j\bullet}V_{n,t}\right| = o_{P_{\theta_n}}(n^{-1/2})$ by applying the Lemma A.1 with $W_{n,t} = A_{n,j\bullet}V_{n,t}$ (noting that $A_{n,k\bullet}V_{n,t} \simeq \epsilon_{k,t}$ and $A_{n,j\bullet}V_{n,t} \simeq \epsilon_{j,t}$ are independent with $\mathbb{E}_{\theta_n}(A_{n,j\bullet}V_{n,t})^2 = 1$ by Assumption 3.3.12, hence the WLLN implies the required convergence) and the outside summations are finite, it follows that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\hat{\varphi}_{1,n,t} - \varphi_{1,n,t}) = o_{P^n_{\tilde{\theta}_n}}(1)\,. \tag{58}$$

That $\hat{\tau}_{k,n} \xrightarrow{P^n_{\tilde{\theta}_n}} \tau_k$ follows from Lemma A.14. Now, consider $\varphi_{2,\tau,n,t}$ defined by

$$\varphi_{2,\tau,n,t} := \sum_{k=1}^{K}\zeta^{\alpha}_{n,l,k,k}\left[\tau_{k,1}A_{n,k\bullet}V_{n,t} + \tau_{k,2}\kappa(A_{n,k\bullet}V_{n,t})\right].$$

Since sum is finite and each $|\zeta^{\alpha}_{n,l,k,k}| \to |\zeta^{\alpha}_{\infty,l,k,k}| < \infty$ it is sufficient to consider the convergence of the summands. In particular we have that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\hat{\tau}_{k,n,1} - \tau_{k,1}\right]A_{n,k\bullet}V_{n,t} = \left[\hat{\tau}_{k,n,1} - \tau_{k,1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_{n,k\bullet}V_{n,t} \to 0,$$

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\hat{\tau}_{k,n,2} - \tau_{k,2}\right]\kappa(A_{n,k\bullet}V_{n,t}) = \left[\hat{\tau}_{k,n,2} - \tau_{k,2}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa(A_{n,k\bullet}V_{n,t}) \to 0,$$

in probability, since $A_{n,k\bullet}V_{n,t} \eqsim \epsilon_{k,t} \sim \eta_k$ and $(\epsilon_{k,t})_{t\geq 1}$ and $(\kappa(\epsilon_{k,t}))_{t\geq 1}$ are i.i.d. mean-zero sequences with finite second moments such that the CLT holds.

Together these yield that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\varphi_{2,\hat{\tau}_n,n,t} - \varphi_{2,\tau,n,t}) \to 0 \quad \text{in } P^n_{\tilde{\theta}_n}\text{-probability.} \tag{59}$$

Putting (58) and (59) together yields the required convergence for components of the type (55). We note that the required convergence for components of type (56) follows using identical steps.

For components (57) let $a_{n,k,l} := -A_{n,k\bullet}D_{b_l}$ and note for $\tilde{\varsigma}_{k,n,1} := \hat{\varsigma}_{k,n} - \varsigma_k$,

$$
\begin{aligned}
\frac{1}{\sqrt{n}} & \sum_{t=1}^{n}(\hat{\ell}_{\tilde{\theta}_n,b_l}(Y_t,X_t) - \tilde{\ell}_{\tilde{\theta}_n,b_l}(Y_t,X_t)) \\
&= \sum_{k=1}^{K} a_{n,k,l}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[(X_t - \mathbb{E}X_t)\left[\hat{\phi}_k(A_{n,k\bullet}V_{n,t}) - \phi_k(A_{n,k\bullet}V_{n,t})\right]\right] \\
&\quad + \sum_{k=1}^{K} a_{n,k,l}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[(\mathbb{E}X_t - \bar{X}_n)\left(\phi_k(A_{n,k\bullet}V_{n,t}) + \hat{\varsigma}_{k,n,1}A_{n,k\bullet}V_{n,t} + \hat{\varsigma}_{k,n,2}\kappa(A_{n,k\bullet}V_{n,t})\right)\right] \\
&\quad - \sum_{k=1}^{K} a_{n,k,l}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\mathbb{E}X_t\left(\tilde{\varsigma}_{k,n,1}A_{n,k\bullet}V_{n,t} + \tilde{\varsigma}_{k,n,2}\kappa(A_{n,k\bullet}V_{n,t})\right)\right].
\end{aligned}
$$

Noting first that $a_{n,k,l} \to a_{\infty,k,l} := A_{k\bullet}D_{b_l}$, each of the terms on the right hand side converges to zero in probability. For the first term, this follows from Lemma A.1 applied with $W_{n,t} := a_{n,k,l}(X_t - \mathbb{E}X_t)$, noting that this is independent of $A_{n,k\bullet}V_{n,t}$ by Assumption 2.[44] For the second term, this follows from A.14, the CLT applied to $A_{n,k\bullet}V_{n,t} \simeq \epsilon_{k,t}$, $\kappa(A_{n,k\bullet}V_{n,t}) \simeq \kappa(\epsilon_{k,t})$ and $\phi_k(A_{n,k\bullet}V_{n,t}) \simeq \phi_k(\epsilon_{k,t})$ and the fact that $\bar{X}_n - \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}X_t$ converges to zero in probability by e.g. Corollary 19.3 in Davidson (1994), Lemma A.3 (which provides a uniform upper bound for the $4+\delta$ moments of the X_t) and Proposition A.2. For the third term, this follows from A.14 and the CLT applied to $A_{n,k\bullet}V_{n,t} \simeq \epsilon_{k,t}$ and $\kappa(A_{n,k\bullet}V_{n,t}) \simeq \kappa(\epsilon_{k,t})$.

The first part of (iii) follows from Lemma A.10. To verify the second part of (iii) we will show that

$$
\left\|\hat{I}_{n,\tilde{\theta}_n} - \tilde{I}_{\theta}\right\|_2 \leq \left\|\hat{I}_{n,\tilde{\theta}_n} - \breve{I}_{n,\tilde{\theta}_n}\right\|_2 + \left\|\breve{I}_{n,\tilde{\theta}_n} - \tilde{I}_{\theta}\right\|_2 = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}). \tag{60}
$$

where $\tilde{I}_\theta := \mathbb{E}[\tilde{\ell}_\theta(Y_t,X_t)\tilde{\ell}_\theta(Y_t,X_t)'] = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[\tilde{\ell}_\theta(Y_t,X_t)\tilde{\ell}_\theta(Y_t,X_t)']$ with the expectation taken under G_θ, $\hat{I}_{n,\theta} := \frac{1}{n}\sum_{t=1}^{n}\hat{\ell}_\theta(Y_t,X_t)\hat{\ell}_\theta(Y_t,X_t)'$ and $\breve{I}_{n,\theta} := \frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_\theta(Y_t,X_t)\tilde{\ell}_\theta(Y_t,X_t)'$.

To obtain the rates we start with $\|\tilde{I}_{\theta_n} - \tilde{I}_\theta\|_2$, for which we show that each component satisfies the required rate. To set this up, let $Q^{r,s}_{l,m,t,n} = \tilde{\ell}_{\tilde{\theta}_n,r_l}(Y_t,X_t)\tilde{\ell}_{\tilde{\theta}_n,s_m}(Y_t,X_t)$, where $r,s \in \{\alpha,\sigma,b\}$ and l,m denote the indices of the components of the efficient scores. Fix any r, s and l,m and note that it suffices to show

$$
\frac{1}{n}\sum_{t=1}^{n}Q^{r,s}_{l,m,t,n} - \mathbb{E}_{\tilde{\theta}_n}Q^{r,s}_{l,m,t,n} + \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\tilde{\theta}_n}[Q^{r,s}_{l,m,t,n}] - \mathbb{E}_{\theta}[Q^{r,s}_{l,m,t,\infty}] = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}).
$$

For the first term, by the fact that $\tilde{\ell}_{\tilde{\theta}_n}$ has uniformly bounded $4+\delta$ moments,[45] Proposition A.2 and Theorem 1 of Kanaya (2017) we obtain

$$
\frac{1}{n}\sum_{t=1}^{n}Q^{r,s}_{l,m,t,n} - \mathbb{E}_{\tilde{\theta}_n}Q^{r,s}_{l,m,t,n} = O_{P^n_{\tilde{\theta}_n}}\left(n^{(1/p-1)/2}\right) = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}), \quad p \in (1, 1+\delta/4].
$$

[44] The convergence condition follows by combining Proposition A.2, Lemma A.3 (which provides a uniform upper bound for the $4+\delta$ moments of the X_t) and Corollary 19.3 of Davidson (1994).

[45] Argue as in Lemma A.9.

That the second term is $o(\nu_n^{1/2})$ follows by the assumed Lipschitz continuity of the map defining the ζ's, that of each $\beta \mapsto A(\alpha, \sigma)_{k\bullet}$ (which holds locally at θ) and Lemma A.13.

For the other component of the sum, let $r \in \{\alpha, \sigma, b\}$ and let l denote an index, we write $\hat{U}_{n,t,r_l} := \hat{\ell}_{\tilde{\theta}_n, r_l}(Y_t, X_t)$, $\tilde{U}_{t,r_l} := \tilde{\ell}_{\tilde{\theta}_n, r_l}(Y_t, X_t)$ and $D_{n,t,r_l} := \hat{\ell}_{\tilde{\theta}_n, r_l}(Y_t, X_t) - \tilde{\ell}_{\tilde{\theta}_n, r_l}(Y_t, X_t)$.

Since it is the absolute value of the $(r,l)-(s,m)$ component of $\hat{I}_{n,\tilde{\theta}_n} - \breve{I}_{n,\tilde{\theta}_n}$, it is sufficient to show that $\left|\frac{1}{n}\sum_{t=1}^n \hat{U}_{n,t,r_l} D_{n,t,s_m} + \frac{1}{n}\sum_{t=1}^n D_{n,t,r_l}\tilde{U}_{t,s_m}\right| = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2})$ as $n \to \infty$ for any $r, s \in \{\alpha, \sigma, b\}$ and l, m. By Cauchy-Schwarz and lemma A.15

$$\left|\frac{1}{n}\sum_{t=1}^n D_{n,t,r_l}\tilde{U}_{t,s_m}\right| \leq \left(\frac{1}{n}\sum_{t=1}^n \tilde{U}_{t,s_m}^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=1}^n D_{n,t,r_l}^2\right)^{1/2} = O_{P^n_{\tilde{\theta}_n}}(1) \times o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}) = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}),$$

$$\left|\frac{1}{n}\sum_{t=1}^n \hat{U}_{n,t,r_l} D_{n,t,s_m}\right| \leq \left(\frac{1}{n}\sum_{t=1}^n \hat{U}_{n,t,r_l}^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=1}^n D_{n,t,s_m}^2\right)^{1/2} = O_{P^n_{\tilde{\theta}_n}}(1) \times o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}) = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}),$$

for any $(r,l)-(s,m)$. It follows that

$$\left[\frac{1}{n}\sum_{t=1}^n \hat{U}_{n,t,r_l} D_{n,t,s_m} + D_{n,t,r_l}\tilde{U}_{t,s_m}\right]^2 \leq 2\left[\frac{1}{n}\sum_{t=1}^n \hat{U}_{n,t,r_l} D_{n,t,s_m}\right]^2 + 2\left[\frac{1}{n}\sum_{t=1}^n D_{n,t,r_l}\tilde{U}_{t,s_m}\right]^2 = o_{P^n_{\tilde{\theta}_n}}(\nu_n)$$

and hence $\|\hat{I}_{n,\tilde{\theta}_n} - \breve{I}_{n,\tilde{\theta}_n}\|_2 \leq \|\hat{I}_{n,\tilde{\theta}_n} - \breve{I}_{n,\tilde{\theta}_n}\|_F = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2})$. We can combine these results to obtain:

$$\|\hat{I}_{n,\tilde{\theta}_n} - \tilde{I}_\theta\|_2 \leq \|\hat{I}_{n,\tilde{\theta}_n} - \breve{I}_{n,\tilde{\theta}_n}\|_2 + \|\breve{I}_{n,\tilde{\theta}_n} - \tilde{I}_\theta\|_2 = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}) + o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}) = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2}).$$

Part (iv) follows directly from Lemma A.8. □

Lemma A.13. *In the setting of Lemma A.12*

1. $\frac{1}{n}\sum_{t=1}^n \mathbb{E}_{\tilde{\theta}_n} X_t - \mathbb{E}_\theta X_t = o(\nu_n^{1/2})$,
2. $\frac{1}{n}\sum_{t=1}^n [\mathbb{E}_{\tilde{\theta}_n} X_t][\mathbb{E}_{\tilde{\theta}_n} X_t]' - [\mathbb{E}_\theta X_t][\mathbb{E}_\theta X_t]' = o(\nu_n^{1/2})$.
3. $\frac{1}{n}\sum_{t=1}^n \mathbb{E}_{\tilde{\theta}_n}[X_t - \mathbb{E}_{\tilde{\theta}_n} X_t][X_t - \mathbb{E}_{\tilde{\theta}_n} X_t]' - \mathbb{E}_\theta[X_t - \mathbb{E}_\theta X_t][X_t - \mathbb{E}_\theta X_t]' = o(\nu_n^{1/2})$.

Proof. For (i) we decompose as

$$\mathbb{E}_{\tilde{\theta}_n} X_t - \mathbb{E}_\theta X_t = [\mathbb{E}_{\tilde{\theta}_n} X_t - \mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t] + [\mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t - \mathbb{E}_\theta \tilde{X}_t] + [\mathbb{E}_\theta \tilde{X}_t - \mathbb{E}_\theta X_t]$$

where $\tilde{X}_t$ denotes a stationary solution to the VAR equation. Note that for all $\vartheta \in \{\tilde{\theta}_n :$

$n \in \mathbb{N}\} \cup \{\theta\}$ and some $\rho_\star < 1$

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_\vartheta X_t - \mathbb{E}_\vartheta \tilde{X}_t\|^2 &= \frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_\vartheta Z_{t-1} - \mathbb{E}_\vartheta \tilde{Z}_{t-1}\|^2 \\
&\leq \|\mathsf{C}_\vartheta\|^2 \times \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{j=t}^{\infty} \|\mathsf{B}_\vartheta^j\|\right)^2 \\
&\leq \|\mathsf{C}_\vartheta\|^2 \times \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{j=t}^{\infty} \rho_\star^j\right)^2 \\
&= \frac{\|\mathsf{C}_\vartheta\|^2}{(1-\rho_\star)^2} \times \frac{1}{n}\sum_{t=1}^{n} \rho_\star^{2t} \\
&= \frac{\|\mathsf{C}_\vartheta\|^2\left(1-\rho_\star^{2(n+1)}\right)}{(1-\rho_\star)^2(1-\rho_\star^2)} \times \frac{1}{n} \\
&= O(n^{-1}),
\end{aligned} \tag{61}$$

and hence by Jensen's inequality $\frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_\vartheta X_t - \mathbb{E}_\vartheta \tilde{X}_t\| = O(n^{-1/2})$. The middle term satisfies

$$\frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t - \mathbb{E}_\theta \tilde{X}_t\| = \|\mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t - \mathbb{E}_\theta \tilde{X}_t\| = (I-\mathsf{B}_{\tilde{\theta}_n})^{-1}\mathsf{C}_{\tilde{\theta}_n} - (I-\mathsf{B}_\theta)^{-1}\mathsf{C}_\theta = O(n^{-1/2}),$$

since $\beta \mapsto (I - \mathsf{B}_\theta)^{-1}\mathsf{C}_\theta$ is locally Lipschitz at θ.

For (ii), note that combination of the preceding displays yields that

$$\frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_{\tilde{\theta}_n} X_t - \mathbb{E}_\theta X_t\|^2 = O(n^{-1}),$$

which, in conjunction with the Cauchy-Schwarz inequality and Lemma A.3 yields (ii).

For (iii) let $U_{\vartheta,t} \coloneqq X_t - \mathbb{E}_\vartheta X_t$ and $\tilde{U}_{\vartheta,t} \coloneqq \tilde{X}_t - \mathbb{E}_\vartheta \tilde{X}_t$. We note that for all $\vartheta \in \{\tilde{\theta}_n : n \in \mathbb{N}\} \cup \{\theta\}$, some $\rho_\star < 1$ and some finite, positive M

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_\vartheta\left(U_{\vartheta,t}U_{\vartheta,t}'\right) - \mathbb{E}_\vartheta\left(\tilde{U}_{\vartheta,t}\tilde{U}_{\vartheta,t}'\right) &= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=t}^{\infty} \mathsf{B}_\vartheta^j \mathsf{D}_\vartheta \mathsf{D}_\vartheta' (\mathsf{B}_\vartheta^j)' \\
&\leq M\frac{1}{n}\sum_{t=1}^{n}\sum_{j=t}^{\infty} \rho_\star^{2j} \\
&= \frac{M}{1-\rho_\star^2}\frac{1}{n}\sum_{t=1}^{n} \rho_\star^{2t} \\
&= \frac{M\left(1-\rho_\star^{2(n+1)}\right)}{(1-\rho_\star^2)^2}\frac{1}{n} \\
&= O(n^{-1}).
\end{aligned}$$

Additionally, we can write $\operatorname{vec}(\mathbb{E}_\vartheta \tilde{U}_{\vartheta,t}\tilde{U}_{\vartheta,t}') = (I - \mathsf{B}_\vartheta \otimes \mathsf{B}_\vartheta)^{-1} \operatorname{vec}(\mathsf{D}_\vartheta \mathsf{D}_\vartheta')$, which is locally Lipschitz in β at θ under our assumptions. This implies that

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\tilde{\theta}_n}\tilde{U}_{\tilde{\theta}_n,t}\tilde{U}_{\tilde{\theta}_n,t}' - \mathbb{E}_{\theta}\tilde{U}_{\theta,t}\tilde{U}_{\theta,t}' = O(n^{-1/2}).$$

By using a similar decomposition as in (i), the previous two displays suffice for (iii). □

Lemma A.14. *If assumption 3.3.1 holds, then* $\|\hat{\varrho}_{k,n} - \varrho_{k,n}\|_2 = o_{P^n_{\tilde{\theta}_n}}(\nu_{n,p}) = o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2})$, *where* $\tilde{\theta}_n$ *is as in Lemma A.12 and* $\varrho \in \{\tau, \varsigma\}$.

Proof. Under $P^n_{\tilde{\theta}_n}$, $A_{n,k\bullet}V_{n,t} \eqsim \epsilon_{k,t} \sim \eta_k$, for $V_{n,t} := Y_t - c_n - B_nX_t$. Let $w \in \{(0,-2)', (1,0)'\}$ By the fact that the map $M \mapsto M^{-1}$ is Lipschitz at a positive definite matrix M_0 we have that for a positive constant C then for large enough n, with probability approaching one

$$\|\hat{\varrho}_{k,n} - \varrho_{k,n}\|_2 = \|(\hat{M}_{k,n}^{-1} - M_k^{-1})w\|_2 \leq 2\|\hat{M}_{k,n}^{-1} - M_k^{-1}\|_2 \leq 2C\|\hat{M}_{k,n} - M_k\|_2. \tag{62}$$

By Theorem 2.5.11 in Durrett (2019)

$$\frac{1}{n}\sum_{t=1}^{n}[(A_{n,k\bullet}V_{n,t})^3 - \mathbb{E}(A_{n,k\bullet}V_{n,t})^3] = o_{P^n_{\tilde{\theta}_n}}\left(n^{\frac{1-p}{p}}\right)$$
$$\frac{1}{n}\sum_{t=1}^{n}[(A_{n,k\bullet}V_{n,t})^4 - \mathbb{E}(A_{n,k\bullet}V_{n,t})^4] = o_{P^n_{\tilde{\theta}_n}}\left(n^{\frac{1-p}{p}}\right).$$

These together imply that

$$\|\hat{M}_{k,n} - M_k\|_2 \leq \|\hat{M}_{k,n} - M_k\|_F = o_{P^n_{\tilde{\theta}_n}}\left(n^{\frac{1-p}{p}}\right) = o_{P^n_{\tilde{\theta}_n}}(\nu_{n,p}).$$

Combining these convergence rates with equation (62) yields the result. □

Lemma A.15. *Suppose assumptions 3.3.1 and 3.3.2 hold and* $\tilde{\theta}_n = (\alpha_n, \beta_n, \eta)$ *where* $\sqrt{n}(\beta_n - \beta) = O(1)$ *is a deterministic sequence. Then for each* $r \in \{\alpha, \sigma, b\}$ *and* l

$$\frac{1}{n}\sum_{t=1}^{n}\left(\hat{\ell}_{\tilde{\theta}_n,r_l}(Y_t, X_t) - \tilde{\ell}_{\tilde{\theta}_n,r_l}(Y_t, X_t)\right)^2 = o_{P^n_{\tilde{\theta}_n}}(\nu_n).$$

Proof. We start by considering elements in $\frac{1}{n}\sum_{t=1}^{n}\left(\hat{\ell}_{\tilde{\theta}_n,\alpha_l}(Y_t, X_t) - \tilde{\ell}_{\tilde{\theta}_n,\alpha_l}(Y_t, X_t)\right)^2$. We define $\tilde{\tau}_{k,n,q} := \hat{\tau}_{k,n,q} - \tau_{k,q}$ and $V_{n,t} = Y_t - c_n - B_nX_t$. Since each $|\zeta^{\alpha}_{n,l,k,j}| < \infty$ and the sums over k, j are finite, it is sufficient to demonstrate that for every $k, j, m, s \in [K]$, with $k \neq j$ and $s \neq m$,

$$\frac{1}{n}\sum_{t=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t}) - \phi_k(A_{n,k\bullet}V_{n,t})\right]\left[\hat{\phi}_{s,n}(A_{n,s\bullet}V_{n,t}) - \phi_s(A_{n,s\bullet}V_{n,t})\right]A_{n,j\bullet}V_{t,n}A_{n,m\bullet}V_{n,t} = o_{P^n_{\tilde{\theta}_n}}(\nu_n), \tag{63}$$

$$\frac{1}{n}\sum_{t=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t}) - \phi_k(A_{n,k\bullet}V_{n,t})\right]A_{n,j\bullet}V_{n,t}\left[\tilde{\tau}_{s,n,1}A_{n,s\bullet}V_{n,t} + \tilde{\tau}_{s,n,2}\kappa(A_{n,s\bullet}V_{n,t})\right] = o_{P^n_{\tilde{\theta}_n}}(\nu_n), \tag{64}$$

$$\frac{1}{n}\sum_{t=1}^{n}\left[\tilde{\tau}_{s,n,1}A_{n,s\bullet}V_{n,t}+\tilde{\tau}_{s,n,2}\kappa(A_{n,s\bullet}V_{n,t})\right]\left[\tilde{\tau}_{k,n,1}A_{n,k\bullet}V_{n,t}+\tilde{\tau}_{k,n,2}\kappa(A_{n,k\bullet}V_{n,t})\right]=o_{P^n_{\tilde{\theta}_n}}(\nu_n). \tag{65}$$

For (65), let $\xi_1(x)=x$ and $\xi_2(x)=\kappa(x)$. Then, we can split the sum into 4 parts, each of which has the following form for some $q,w\in\{1,2\}$

$$\frac{1}{n}\sum_{t=1}^{n}\tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w}\xi_q(A_{n,s\bullet}V_{n,t})\xi_w(A_{n,k\bullet}V_{n,t})=\tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w}\frac{1}{n}\sum_{t=1}^{n}\xi_q(A_{n,s\bullet}V_{n,t})\xi_w(A_{n,k\bullet}V_{n,t})=o_{P^n_{\tilde{\theta}_n}}(\nu_n),$$

since we have that each $\tilde{\tau}_{s,n,q}\tilde{\tau}_{k,n,w}=o_{P^n_{\tilde{\theta}_n}}(\nu_n)$ by lemma A.14.[46] For (64) we can argue similarly. Again let $\xi_1(x)=x$ and $\xi_2(x)=\kappa(x)$. Then, we can split the sum into 2 parts, each of which has the following form for some $q\in\{1,2\}$

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t})-\phi_k(A_{n,k\bullet}V_{n,t})\right]A_{n,j\bullet}V_{n,t}\tilde{\tau}_{s,n,q}\xi_q(A_{n,s\bullet}V_{n,t})\\
&\quad\le\tilde{\tau}_{s,n,q}\left(\frac{1}{n}\sum_{t=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t})-\phi_k(A_{n,k\bullet}V_{n,t})\right]^2(A_{n,j\bullet}V_{n,t})^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=1}^{n}\xi_q(A_{n,s\bullet}V_{n,t})^2\right)^{1/2}\\
&\quad=o_{P^n_{\tilde{\theta}_n}}(\nu_n).
\end{aligned}$$

by Lemma A.1 applied with $W_{n,t}=A_{n,j\bullet}V_{n,t}$ and $\tilde{\tau}_{s,n,q}=o_{P^n_{\tilde{\theta}_n}}(\nu_n^{1/2})$.[47] For (63) use Cauchy-Schwarz with Lemma A.1

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t})-\phi_k(A_{n,k\bullet}V_{n,t})\right]\left[\hat{\phi}_{s,n}(A_{n,s\bullet}V_{n,t})-\phi_s(A_{n,s\bullet}V_{n,t})\right]A_{n,j\bullet}V_{n,t}A_{n,m\bullet}V_{n,t}\\
&\quad\le\left(\frac{1}{n}\sum_{t=1}^{n}\left[\hat{\phi}_{k,n}(A_{n,k\bullet}V_{n,t})-\phi_k(A_{n,k\bullet}V_{n,t})\right]^2(A_{n,j\bullet}V_{n,t})^2\right)^{1/2}\\
&\qquad\quad\times\left(\frac{1}{n}\sum_{t=1}^{n}\left[\hat{\phi}_{s,n}(A_{n,s\bullet}V_{n,t})-\phi_s(A_{n,s\bullet}V_{n,t})\right]^2(A_{n,m\bullet}V_{n,t})^2\right)^{1/2}\\
&\quad=o_{P^n_{\tilde{\theta}_n}}(\nu_n).
\end{aligned}$$

This completes the proof for the components corresponding to α_l. We note that the components corresponding to σ_l follow identically.

Finally, we consider the elements in $\frac{1}{n}\sum_{t=1}^{n}\left(\hat{\ell}_{\theta_n,b_l}(Y_t,X_t)-\tilde{\ell}_{\theta_n,b_l}(Y_t,X_t)\right)^2$, where we

[46] The fact that $\frac{1}{n}\sum_{t=1}^{n}\xi_q(A_{n,s\bullet}V_{n,t})\xi_w(A_{n,k\bullet}V_{n,t})=O_{P^n_{\tilde{\theta}_n}}(1)$ can be seem to hold using the moment and i.i.d. assumptions from assumption 3.3.1 and Markov's inequality, noting once more that $A_{n,k\bullet}V_{n,t}\simeq\epsilon_{k,t}$ under $P^n_{\tilde{\theta}_n}$.

[47] See footnote 46.

note that with $\tilde{\varsigma}_{k,n} := \hat{\varsigma}_{k,n} - \varsigma_k$,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left(\hat{\ell}_{\theta_n,b_l}(Y_t,X_t) - \tilde{\ell}_{\theta_n,b_l}(Y_t,X_t)\right)^2 \\
&\quad\lesssim \sum_{k=1}^{K}\frac{1}{n}\sum_{t=1}^{n}\left[[a_{n,k,l}(X_t - \mathbb{E}X_t)]^2\left[\hat{\phi}_k(A_{n,k\bullet}V_{n,t}) - \phi_k(A_{n,k\bullet}V_{n,t})\right]^2\right] \\
&\qquad + \sum_{k=1}^{K}\frac{1}{n}\sum_{t=1}^{n}\left[[a_{n,k,l}(\mathbb{E}X_t - \bar{X}_n)]^2\left(\phi_k(A_{n,k\bullet}V_{n,t}) + \hat{\varsigma}_{k,n,1}A_{n,k\bullet}V_{n,t} + \hat{\varsigma}_{k,n,2}\kappa(A_{n,k\bullet}V_{n,t})\right)^2\right] \\
&\qquad + \sum_{k=1}^{K}\frac{1}{n}\sum_{t=1}^{n}\left[[a_{n,k,l}\mathbb{E}X_t]^2\left(\tilde{\varsigma}_{k,n,1}A_{n,k\bullet}V_{n,t} + \tilde{\varsigma}_{k,n,2}\kappa(A_{n,k\bullet}V_{n,t})\right)^2\right]
\end{aligned}$$

That the first right hand side term is $o_{P^n_{\tilde{\theta}_n}}(\nu_n)$ follows by Lemma A.1.[48] and the Cauchy-Schwarz inequality. The third follows from Lemma A.14 since $[a_{n,k,l}\mathbb{E}X_t]^2$ is uniformly (in t) bounded (cf. Lemma A.3).

For the second, let $\tilde{X}_t$ denote a random vector which has the stationary distribution of X_t and note that by equation (61) we have

$$\frac{1}{n}\sum_{t=1}^{n}\|\mathbb{E}_{\tilde{\theta}_n}X_t - \mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t\|^2 = O(n^{-1}).$$

Now let

$$\begin{aligned}
U_{n,t} &:= \left(\phi_k(A_{n,k\bullet}V_{n,t}) + \varsigma_{k,1}A_{n,k\bullet}V_{n,t} + \varsigma_{k,2}\kappa(A_{n,k\bullet}V_{n,t})\right)^2 \\
\tilde{U}_{n,t} &:= \left(\tilde{\varsigma}_{k,n,1}A_{n,k\bullet}V_{n,t} + \tilde{\varsigma}_{k,n,2}\kappa(A_{n,k\bullet}V_{n,t})\right)^2.
\end{aligned}$$

By Theorem 1 in Arnold (1985) and Markov's inequality, we have that $\max_{1\leq t\leq n}U_{n,t} = O_{P^n_{\tilde{\theta}_n}}(n^{1/p})$. Then

$$\frac{1}{n}\sum_{t=1}^{n}\|\mathbb{E}_{\tilde{\theta}_n}X_t - \mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t\|^2U_{n,t} \leq \max_{1\leq t\leq n}U_{n,t}\frac{1}{n}\sum_{t=1}^{n}\|\mathbb{E}_{\tilde{\theta}_n}X_t - \mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t\|^2 = O_{P^n_{\tilde{\theta}_n}}(n^{-1+1/p}) = o_{P^n_{\tilde{\theta}_n}}(\nu_n).$$

Additionally, by equation (61), Jensen's inequality, Lemma A.3 and Theorem 2 of Kanaya (2017)

$$\|\mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t - \bar{X}_n\|^2 \leq 2\left[\left\|\frac{1}{n}\sum_{t=1}^{n}(X_t - \mathbb{E}_{\tilde{\theta}_n}X_t)\right\|^2 + \left\|\frac{1}{n}\sum_{t=1}^{n}(\mathbb{E}_{\tilde{\theta}_n}X_t - \mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t)\right\|^2\right] = O_{P^n_{\tilde{\theta}_n}}(n^{-1}) + O(n^{-1}),$$

hence

$$\frac{1}{n}\sum_{t=1}^{n}\|\mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t - \bar{X}_n\|^2U_{n,t} = \|\mathbb{E}_{\tilde{\theta}_n}\tilde{X}_t - \bar{X}_n\|^2\frac{1}{n}\sum_{t=1}^{n}U_{n,t} = O_{P^n_{\tilde{\theta}_n}}(n^{-1}) = o_{P^n_{\tilde{\theta}_n}}(\nu_n).$$

To complete the proof, it suffices to combine the above results with the observation that by

[48] Cf. footnote 44.

Lemma A.14 and Theorem 1 of Arnold (1985)

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_{\tilde{\theta}_n} X_t - \bar{X}_n\|^2 \tilde{U}_{n,t} \\
&\quad \lesssim \tilde{\tau}_{k,n,1}^2 \frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_{\tilde{\theta}_n} \tilde{X}_t - \bar{X}_n\|^2 (A_{n,k\bullet}V_{n,t})^2 + \tilde{\tau}_{k,n,1}^2 \max_{1\leq t\leq n} (A_{n,k\bullet}V_{n,t})^2 \frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_{\tilde{\theta}_n} \tilde{X}_t - \mathbb{E}_{\tilde{\theta}_n} X_t\|^2 \\
&\qquad + \tilde{\tau}_{k,n,2}^2 \frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_{\tilde{\theta}_n} \tilde{X}_t - \bar{X}_n\|^2 \kappa(A_{n,k\bullet}V_{n,t})^2 + \tilde{\tau}_{k,n,2}^2 \max_{1\leq t\leq n} \kappa(A_{n,k\bullet}V_{n,t})^2 \frac{1}{n}\sum_{t=1}^{n} \|\mathbb{E}_{\tilde{\theta}_n} \tilde{X}_t - \mathbb{E}_{\tilde{\theta}_n} X_t\|^2 \\
&\quad = o_{P^n_{\tilde{\theta}_n}}(\nu_n).
\end{aligned}$$

□

A.4. Miscellaneous results

The results in this subsection are general results, which are useful in establishing the main results of the paper, but are not specific to the model under study.

Proposition A.16 (Cf. Proposition 2.29 in van der Vaart, 1998)**.** *Suppose that on a measureable space* $(S, \mathcal{S})$*,* $(\mu_n)_{n\in\mathbb{N}}$ *is a sequence of finite measures such that* $\mu_n \xrightarrow{TV} \mu$ *(with* μ *a finite measure on* $(S, \mathcal{S})$*. If* $(f_n)_{n\in\mathbb{N}}$ *and* f *are (real-valued) measurable functions such that* $f_n \to f$ *in* μ*-measure and* $\limsup_{n\to\infty} \int |f_n|^p \,\mathrm{d}\mu_n \leq \int |f|^p \,\mathrm{d}\mu < \infty$ *for some* $p \geq 1$*, then* $\int |f_n - f|^p \,\mathrm{d}\mu_n \to 0$*.*

Proof. $(a+b)^p \leq 2^p(a^p + b^p)$ for any $a, b \geq 0$ and hence, under our hypotheses,

$$0 \leq 2^p|f_n|^p + 2^p|f|^p - |f_n - f|^p \to 2^{p+1}|f|^p \quad \text{in } \mu \text{ - measure.}$$

By Corollary 2.3 of Feinberg et al. (2016) and the hypothesis that $\limsup_{n\to\infty} \int |f_n|^p \,\mathrm{d}\mu_n \leq \int |f|^p \,\mathrm{d}\mu < \infty$,

$$\begin{aligned}
\int 2^{p+1}|f|^p \,\mathrm{d}\mu &\leq \liminf_{n\to\infty} \int 2^p|f_n|^p + 2^p|f|^p - |f_n - f|^p \,\mathrm{d}\mu_n \\
&\leq 2^{p+1} \int |f|^p \,\mathrm{d}\mu - \limsup_{n\to\infty} \int |f_n - f|^p \,\mathrm{d}\mu_n.
\end{aligned}$$

□

Proposition A.17. *Let* $\{Z_{n,k}, \mathcal{F}_{n,k} : k \leq n, n \in \mathbb{N}\}$ *be a martingale difference array of* $L-$*dimensional random vectors, such that* $\Sigma_{n,k} := \mathbb{E}\left[Z_{n,k}Z_{n,k}'\right]$ *exists. Suppose that*

$$\frac{1}{n}\sum_{k=1}^{n} \Sigma_{n,k} \to \Sigma_\star, \tag{66}$$

with $\Sigma_\star$ *positive semi-definite (and finite) and that for each* $\varepsilon > 0$

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\|Z_{n,k}\|^2 \mathbf{1}\{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}\}\right] \to 0. \tag{67}$$

Then

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} Z_{n,k} \rightsquigarrow \mathcal{N}(0, \Sigma_\star).$$

Proof. Put $\xi_{n,k} := Z_{n,k}/\sqrt{n}$ for $k \leq n$ and $\xi_{n,k} := 0$ otherwise. Fix $a \in \mathbb{R}^L$. The adapted sequence $(a'\xi_{n,k}, \mathcal{F}_{n,k})_{k\in\mathbb{N}}$ is clearly a martingale difference sequence under our hypotheses. Moreover, the sums $\sum_{k=1}^{\infty} a'\xi_{n,k} = \sum_{k=1}^{n} a'\xi_{n,k}$ and $\sum_{k=1}^{\infty} \mathbb{E}[(a'\xi_{n,k})^2] = \sum_{k=1}^{n} \mathbb{E}[(a'\xi_{n,k})^2]$ trivially converge with probability 1 for each $n \in \mathbb{N}$. By linearity and continuity we have that

$$\sum_{k=1}^{\infty} \mathbb{E}[(a'\xi_{n,k})^2] = \sum_{k=1}^{n} \mathbb{E}[(a'\xi_{n,k})^2] = a' \left[\frac{1}{n} \sum_{k=1}^{n} \Sigma_{n,k} \right] a \to a'\Sigma_\star a \geq 0.$$

Next, suppose that $a \neq 0$ and let $\varepsilon > 0$. We have that $\{|a'Z_{n,k}| \geq \varepsilon\sqrt{n}\} \subset \{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}/\|a\|\}$ and therefore

$$\sum_{k=1}^{\infty} \mathbb{E}\left[(a'\xi_{n,k})^2 \mathbf{1}\{|a'\xi_{n,k}| \geq \varepsilon\} \right] \leq \|a\|^2 \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[\|Z_{n,k}\|^2 \mathbf{1}\{\|Z_{n,k}\| \geq \varepsilon\sqrt{n}/\|a\|\} \right] \to 0,$$

by assumption.[49] This establishes that the conditions of Theorem 18.1 of Billingsley (1999) are satisfied and hence

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} a'Z_{n,k} = \sum_{k=1}^{\infty} a'\xi_{n,k} \rightsquigarrow \mathcal{N}(0, a'\Sigma_\star a).$$

The claimed result then follows by an application of the Cramér-Wold theorem. □

Remark A.1. Proposition A.17 is, of course, completely standard. It is recorded here because we have been unable to find a reference for a multivariate CLT for martingale difference arrays which permits a positive *semi*-definite limiting variance matrix.

Theorem A.18 (Extended uniformly equicontinuous mapping). *Let (X, d_X) and (Y, d_Y) be separable metric spaces and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions from $X \to Y$ and $(g_n)_{n\in\mathbb{N}}$ a uniformly equicontinuous sequence of functions from $X \to Y$. Suppose that $x \mapsto d_Y(f_n(x), g_n(x))$ converges compactly to 0. If $(P_n)_{n\in\mathbb{N}}$ and $(Q_n)_{n\in\mathbb{N}}$ are sequences of laws on X such that (i) $(P_n)_{n\in\mathbb{N}}$ is uniformly tight and (ii) $d_{BL}(P_n, Q_n) \to 0$, then $d_{BL}(\tilde{P}_n, \tilde{Q}_n) \to 0$ for $\tilde{P}_n := P_n \circ f_n^{-1}$ and $\tilde{Q}_n := Q_n \circ g_n^{-1}$.*

Proof. By Theorem 11.7.1 in Dudley (2002), there exist on some probability space X-valued random variables X_n and Y_n such that $X_n \sim P_n$ and $Y_n \sim Q_n$ and $d_X(X_n, Y_n) \to 0$ in probability. By the triangle inequality

$$d_Y(f_n(X_n), g_n(Y_n)) \leq d_Y(f_n(X_n), g_n(X_n)) + d_Y(g_n(X_n), g_n(Y_n)).$$

By uniform equicontinuity of $(g_n)_{n\in\mathbb{N}}$, $d_Y(g_n(X_n), g_n(Y_n)) \to 0$ in probability. Let δ, $\varepsilon > 0$ be given and choose a compact K such that (each) $P_n K > 1 - \varepsilon$. The compact convergence ensures that for all sufficiently large n, $\sup_{x\in K} d_Y(f_n(x), g_n(x)) < \delta$. Hence,

[49] In the case that $a = 0$ this limit trivially holds.

for all such n,

$$\mathrm{P}\left(d_Y(f_n(X_n), g_n(X_n)) > \delta\right) \leq \mathrm{P}\left(X_n \notin K\right) = P_n K^{\complement} \leq \epsilon.$$

It follows that $d_Y(f_n(X_n), g_n(Y_n)) \to 0$ in probability and the conclusion follows by applying Theorem 11.7.1 in Dudley (2002) once more. □

Theorem A.19 (Uniform Delta-method)**.** *Let U and V be normed linear spaces and $\phi : U_\phi \to V$ (with $U_\phi \subset U$). Let $(r_n)_{n\in\mathbb{N}}$ be a sequence of constants with $r_n \to \infty$, $(X_n)_{n\in\mathbb{N}}$ a sequence of U_ϕ-valued random variables, $(\theta_n)_{n\in\mathbb{N}} \subset U_\phi$ and $(P_n)_{n\in\mathbb{N}}$, $(Q_n)_{n\in\mathbb{N}}$ sequences of laws on U with (each) $P_n U_0 = 1$ for a separable $U_0 \subset U$. Suppose that (i) ϕ is Hadamard differentiable tangentially to U_0, uniformly along $(\theta_n)_{n\in\mathbb{N}}$, with derivative ϕ'_θ, (ii) $T_n := r_n(X_n - \theta_n) \sim P_n$ where $(P_n)_{n\in\mathbb{N}}$ is uniformly tight, (iii) $d_{BL}(P_n, Q_n) \to 0$ and (iv) $(\phi'_{\theta_n})_{n\in\mathbb{N}}$ is uniformly equicontinuous. Then,*

$$d_{BL}\left(\mathscr{L}\left(r_n\left(\phi(X_n) - \phi(\theta_n)\right)\right),\ Q_n \circ [\phi'_{\theta_n}]^{-1}\right) \to 0. \tag{68}$$

Proof. Define $f_n(h) := r_n\left(\phi(\theta_n + r_n^{-1}h) - \phi(\theta_n)\right)$ and $g_n(h) := \phi'_{\theta_n}(h)$. By our uniform differentiability assumption, for any compact $K \subset U_0$ we have

$$\sup_{h\in K}\left\|r_n\left(\phi(\theta_n + r_n^{-1}h) - \phi(\theta_n)\right) - \phi'_{\theta_n}(h)\right\| \to 0,$$

and so $h \mapsto \|f_n(g) - g_n(h)\|$ converges compactly to 0 on U_0.[50] This fact and (ii) - (iv) allows the application of Theorem A.18 to conclude (68).[51] □

Remark A.2. Since Hadamard derivatives are bounded linear maps by definition, a sufficient condition for the uniform equicontinuity of $(\phi'_{\theta_n})_{n\in\mathbb{N}}$ is that $\sup_{n\in\mathbb{N}}\|\phi'_{\theta_n}\| < \infty$, i.e. their operator norms are uniformly bounded. This ensures that each ϕ'_{θ_n} is Lipschitz with Lipschitz constant $\sup_{n\in\mathbb{N}}\|\phi'_{\theta_n}\|$ and hence the collection is uniformly equicontinuous.

[50] Cf. e.g. pp. 453 – 455 in Bickel et al. (1998)

[51] The image of a separable space under a continuous function is separable, cf. Theorem 16.4(a) in Willard (1970).

Bibliography

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

Amari, S.-I. and Cardoso, J.-F. (1997). Blind source separation-semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700.

Amengual, D., Fiorentini, G., and Sentana, E. (2021). Moment tests of independent components.

Anderson, T. W. and Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, 20(1):46 – 63.

Andrews, D. W., Cheng, X., and Guggenberger, P. (2020). Generic results for establishing the asymptotic size of confidence sets and tests. *Journal of Econometrics*, 218(2):496–531.

Andrews, D. W. K. (1987). Asymptotic results for generalized wald tests. *Econometric Theory*, 3(3):348–358.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405.

Andrews, D. W. K. (2001). Testing When a Parameter is on the Boundary of the Maintained Hypothesis. *Econometrica*, 69(3):683–734.

Andrews, D. W. K. and Cheng, X. (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica*, 80(5):2153–2211.

Andrews, D. W. K. and Cheng, X. (2013). Maximum likelihood estimation and uniform inference with sporadic identification failure. *Journal of Econometrics*, 173(1):36–56.

Andrews, D. W. K. and Guggenberger, P. (2009). Hybrid and size-corrected subsampling methods. *Econometrica*, 77(3):721–762.

Andrews, D. W. K. and Guggenberger, P. (2010a). Applications of subsampling, hybrid, and size-correction methods. *Journal of Econometrics*, 158(2):285–305.

Andrews, D. W. K. and Guggenberger, P. (2010b). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, 26(2):426–468.

Andrews, D. W. K. and Guggenberger, P. (2019). Identification- and singularity-robust inference for moment condition models. *Quantitative Economics*, 10(4):1703–1746.

Andrews, I. and Mikusheva, A. (2015). Maximum likelihood inference in weakly identified dynamic stochastic general equilibrium models. *Quantitative Economics*, 6(1):123–152.

Andrews, I. and Mikusheva, A. (2016a). Conditional inference with a functional nuisance parameter. *Econometrica*, 84(4):1571–1612.

Andrews, I. and Mikusheva, A. (2016b). A geometric approach to nonlinear econometric models. *Econometrica*, 84(3):1249–1264.

Andrews, I. and Mikusheva, A. (2022). Optimal decision rules for weak gmm. *Econometrica*, 90(2):715–748.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753.

Antoine, B. and Renault, E. (2009). Efficient gmm with nearly-weak instruments. *The Econometrics Journal*, 12(s1):S135–S171.

Antoine, B. and Renault, E. (2011). Efficient inference with poor instruments: A general framework. In Ullah, A. and Giles, D. E. A., editors, *Handbook of Empirical Economics and Finance*. CRC Press.

Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, 58(2):277–297.

Arnold, B. C. (1985). p-norm bounds on the expectation of the maximum of a possibly dependent sample. *Journal of Multivariate Analysis*, 17(3):316–332.

Baumeister, C. and Hamilton, J. D. (2015). Sign restrictions, structural vector autoregressions, and useful prior information. *Econometrica*, 83(5):1963–1999.

Baumeister, C. and Hamilton, J. D. (2019). Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks. *American Economic Review*, 109(5):1873–1910.

Bekaert, G., Engstrom, E., and Ermolov, A. (2019). Macro Risks and the Term Structure of Interest Rates. *Working paper*.

Bekaert, G., Engstrom, E., and Ermolov, A. (2020). Aggregate Demand and Aggregate Supply Effects of COVID-19: A Real-time Analysis. *Working paper*.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Ben-Israel, A. and Greville, T. N. E. (2003). *Generalized Inverses: Theory and Applications*. Springer, New York, NY, USA.

Ben-Moshe, D. (2020). Identification of linear regressions with errors in all variables. *Econometric Theory*, page 1–31.

Bhatia, R. (1997). *Matrix Analysis*. Springer, New York, NY, USA.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, NY, USA.

Bickel, P. J. and Ritov, Y. (1987). Efficient Estimation in the Errors in Variables Model. *The Annals of Statistics*, 15(2):513 – 540.

Bickel, P. J., Ritov, Y., and Stoker, T. M. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *Ann. Statist.*, 34(2):721–741.

Billingsley, P. (1995). *Probability and Measure*. Wiley.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley.

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution Free Tests of Independence Based on the Sample Distribution Function. *The Annals of Mathematical Statistics*, 32(2):485 – 498.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1):115–143.

Bogachev, V. I. (2007). *Measure Theory*. Springer Berlin Heidelberg.

Braun, R. (2021). The importance of supply and demand for oil prices: evidence from a svar identified by non-gaussianity. *working paper*.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, New York, NY, USA.

Brunnermeier, M., Palia, D., Sastry, K. A., and Sims, C. A. (2021). Feedbacks: Financial markets and economic activity. *American Economic Review*, 111(6):1845–79.

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2012). Optimal inference for instrumental variables regression with non-gaussian errors. *Journal of Econometrics*, 167(1):1 – 15.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020). Large sample properties of partitioning-based series estimators. *The Annals of Statistics*, 48(3):1718 – 1741.

Cavaliere, G., Nielsen, H. B., Pedersen, R. S., and Rahbek, A. (2020). Bootstrap inference on the boundary of the parameter space, with application to conditional volatility models. *Journal of Econometrics*.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics*, 32(2):189–218.

Chaudhuri, S. and Zivot, E. (2011). A new method of projection-based inference in gmm with weakly identified nuisance parameters. *Journal of Econometrics*, 164(2):239–251.

Chen, A. and Bickel, P. J. (2006). Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825 – 2855.

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

Chen, X., Christensen, T. M., and Tamer, E. (2018). Monte carlo confidence sets for identified sets. *Econometrica*, 86(6):1965–2018.

Chen, Y., Ning, J., Ning, Y., Liang, K.-Y., and Bandeen-Roche, K. (2017). On pseudolikelihood inference for semiparametric models with boundary problems. *Biometrika*, 104(1):165–179.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25(3):573–578.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688.

Chetverikov, D., Santos, A., and Shaikh, A. M. (2018). The econometrics of shape restrictions. *Annual Review of Economics*, 10(1):31–63.

Choi, S., Hall, W. J., and Schick, A. (1996). Asymptotically uniformly most powerful tests in parametric and semiparametric models. *Ann. Statist.*, 24(2):841–861.

Chow, Y. S. and Teicher, H. (1997). *Probability Theory*. Springer Texts in Statstics. Springer, third edition.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314.

Conway, J. B. (1985). *A course in functional analysis*. Springer, New York, NY, USA.

Davidson, J. (1994). *Stochastic limit theory*. Oxford University Press.

Davis, R. and Fernandes, L. (2022). Independent Component Analysis with Heavy Tails using Distance Covariance. Working paper.

Davis, R. and Ng, S. (2021). Time series estimation of the dynamic effects of disaster-type shock. arXiv:2107.06663.

Davis, R. and Ng, S. (2022). Time Series Estimation of the Dynamic Effects of Disaster-Type Shocks. Working paper.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer, New York, NY, USA.

Dhrymes, P. J. (1994). *Topics in Advanced Econometrics, Volume II Linear and Nonlinear Simultaneous Equations*. Springer-Verlag New York.

Drabek, P. and Milota, J. (2007). *Methods of Nonlinear Analysis: Applications to Differential Equations*. Birkhäuser Advanced Texts Basler Lehrbücher. Birkhäuser Basel.

Drautzburg, T. and Wright, J. H. (2021). Refining set-identification in vars through independence. Working paper.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge, UK.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 65(6):1365–1387.

Durrett, R. (2019). *Probability Theory and Examples*. Cambridge University Press, Cambridge, UK, fifth edition.

Elliott, G., Müller, U. K., and Watson, M. W. (2015). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica*, 83(2):771–811.

Feinberg, E. A., Kasyanov, P. O., and Zgurovsky, M. Z. (2016). Uniform fatou's lemma. *Journal of Mathematical Analysis and Applications*, 444(1):550–567.

Fernandez, C. and Steel, M. F. J. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Fiorentini, G. and Sentana, E. (2021). Specification tests for non-gaussian maximum likelihood estimators. *Quantiative Economics*, 12.

Fiorentini, G. and Sentana, E. (2022). Discrete mixtures of normals pseudo maximum likelihood estimators of structural vector autoregressions. Working paper.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, Inc., New York, NY, USA.

Frisch, R. (1933). Propagation problems and impulse problems in dynamic economics. In *Economic Essays in Honor of Gustav Cassel*. George Allen and Unwin.

Garoni, C. and Serra-Capizzano, S. (2017). *Generalized Locally Toeplitz Sequences: Theory and Applications*, volume 1. Springer, Cham, Switzerland.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics*, 22(4):1993–2010.

Gouriéroux, C., Monfort, A., and Renne, J.-P. (2017). Statistical inference for independent component analysis: Application to structural VAR models. *Journal of Econometrics*, 196.

Gouriéroux, C., Monfort, A., and Renne, J.-P. (2019). Identification and Estimation in Non-Fundamental Structural VARMA Models. *The Review of Economic Studies*, 87(4):1915–1953.

Granziera, E., Moon, H. R., and Schorfheide, F. (2018). Inference for vars identified with sign restrictions. *Quantitative Economics*, 9(3):1087–1121.

Guay, A. (2020). Identification of structural vector autoregressions through higher unconditional moments. *Journal of Econometrics*. forthcoming.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, 12. Supplement.

Hahn, J. (1994). The efficiency bound of the mixed proportional hazard model. *The Review of Economic Studies*, 61(4):607–629.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Application*. Academic Press, New York, NY, USA.

Hall, W. J. and Mathiason, D. J. (1990). On large-sample estimation and testing in parametric models. *International Statistical Review*, 58(1):77–97.

Hallin, M. and Saidi, A. (2007). Optimal tests of noncorrelation between multivariate time series. *Journal of the American Statistical Association*, 102(479):938–951.

Hallin, M. and Werker, B. J. M. (1999). Optimal testing for semi-parametric autoregressive models: From gaussian lagrange multipliers to regression rank scores and adaptive tests. In Ghosh, S., editor, *Asymptotics, Nonparametrics and Time Series*, pages 295 – 358. Marcel Dekker.

Han, S. and McCloskey, A. (2019). Estimation and inference with a (nearly) singular jacobian. *Quantitative Economics*, 10(3):1019–1068.

Herrera, A. M. and Rangaraju, S. K. (2020). The effect of oil supply shocks on us economic activity: What have we learned? *Journal of Applied Econometrics*, 35(2):141–159.

Herwartz, H. (2019). Long-run neutrality of demand shocks: Revisiting blanchard and quah (1989) with independent structural shocks. *Journal of Applied Econometrics*, 34(5):811–819.

Herwartz, H., Lange, A., and Maxand, S. (2019). Statistical identification in svars - monte carlo experiments and a comparative assessment of the role of economic uncertainties for the us business cycle. Discussion Paper 375, CEGE.

Hoch, I. (1958). Simultaneous equation bias in the context of the cobb-douglas production function. *Econometrica*, 26(4):566–578.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325.

Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis*. Cambridge University Press, second edition.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer-Verlag New York.

Huang, J. Z. and Su, Y. (2021). Asymptotic properties of penalized spline estimators in concave extended linear models: Rates of convergence. arXiv:2105.06367.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley, New York.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120.

Jansson, M. (2008). Semiparametric power envelopes for tests of the unit root hypothesis. *Econometrica*, 76(5):1103–1142.

Jeganathan, P. (1995). Some aspects of asymptotic theory with applications to time series models. *Econometric Theory*, 11(5):818–887.

Jin, K. (1992). Empirical Smoothing Parameter Selection In Adaptive Estimation. *Annals of Statistics*, 20(4).

Jin, Z., Risk, B. B., and Matteson, D. S. (2019). Optimization and testing in linear non-gaussian component analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):141–156.

Kaido, H., Molinari, F., and Stoye, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432.

Kaji, T. (2021). Theory of weak identification in semiparametric models. *Econometrica*, 89(2):733–763.

Kallenberg, O. (2021). *Foundations of Modern Probability*. Springer, third edition.

Kanaya, S. (2017). Convergence rates of sums of α-mixing triangular arrays: With an application to nonparametric drift function estimation of continuous-time processes. *Econometric Theory*, 33(5):1121–1153.

Ketz, P. (2018). Subvector inference when the true parameter vector may be near or at the boundary. *Journal of Econometrics*, 207(2):285–306.

Kilian, L. and Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press.

Kilian, L. and Murphy, D. P. (2012). Why agnostic sign restrictions are not enough: understanding the dynamics of oil market var models. *Journal of the European Economic Association*, 10(5):1166–1188.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803.

Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica*, 73(4):1103–1123.

Kleibergen, F. (2007). Generalizing weak instrument robust iv statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *Journal of Econometrics*, 139(1):181–216.

Kocherlakota, S. and Kocherlakota, K. (1991). Neyman's C(α) test and Rao's efficient score test for composite hypotheses. *Statistics & Probability Letters*, 11(6):491 – 493.

Kuchibhotla, A. K. and Patra, R. K. (2020). Efficient estimation in single index models through smoothing splines. *Bernoulli*, 26(2):1587 – 1618.

Lanne, M. and Luoto, J. (2021). GMM Estimation of Non-Gaussian Structural Vector Autoregression. *Journal of Business & Economic Statistics*, 39(1):69–81.

Lanne, M. and Lütkepohl, H. (2010). Structural vector autoregressions with nonnormal residuals. *Journal of Business & Economic Statistics*, 28(1):159–168.

Lanne, M., Meitz, M., and Saikkonen, P. (2017). Identification and estimation of non-Gaussian structual vector autoregressions. *Journal of Econometrics*, 196.

Le Cam, L. M. (1960). *Locally Asymptotically Normal Families of Distributions: Certain Approximations to Families of Distributions and Their Use in the Theory of Estimation and Testing Hypotheses*. University of California Berkeley, Calif: University of California publications in statistics. University of California Press.

Le Cam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, New York, NY, USA.

Le Cam, L. M. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York, NY, USA, second edition.

Lee, A. (2022). Robust and efficient inferrence for non-regular semiparametric models. Working Paper.

Lee, A. and Mesters, G. (2022a). Robust inference for non-gaussian linear simultaneous equations models. Working Paper.

Lee, A. and Mesters, G. (2022b). Supplement to "robust inference for non-gaussian linear simultaneous equations models". Working Paper.

Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, NY, USA, third edition.

Liebscher, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, 26(5):669–689.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Lütkepohl, H. and Burda, M. M. (1997). Modified wald tests under nonregular conditions. *Journal of Econometrics*, 78(1):315–332.

Ma, Y. and Zhu, L. (2013). Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):305–322.

Magnus, J. R. and Neudecker, H. (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons.

Magnus, J. R., Pijls, H. G. J., and Sentana, E. (2020). The jacobian ofthe exponential function. Working Paper.

Marron, J. S. and Wand, M. P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics*, 20(2):712 – 736.

Marschak, J. and Andrews, W. H. (1944). Random simultaneous equations and the theory of production. *Econometrica*, 12(3/4):143–205.

Matteson, D. S. and Tsay, R. S. (2017). Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112(518):623–637.

Maxand, S. (2018). Identification of independent structural shocks in the presence of multiple gaussian components. *Econometrics and Statistics*.

McCloskey, A. (2017). Bonferroni-based size-correction for nonstandard testing problems. *Journal of Econometrics*, 200(1):17–35.

McCloskey, A. (2020). Asymptotically uniform tests after consistent model selection in the linear regression model. *Journal of Business & Economic Statistics*, 38(4):810–825.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, second edition.

Moneta, A., Entner, D., Hoyer, P. O., and Coad, A. (2013). Causal inference by independent component analysis: Theory and applications*. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730.

Moneta, A. and Pallante, G. (2020). Identification of Structural VAR Models via Independent Component Analysis: A Performance Evaluation Study. LEM Papers Series 2020/24, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy.

Montiel Olea, J. L. (2020). Admissible, similar tests: A characterization. *Econometric Theory*, 36(2):347–366.

Montiel Olea, J. L., Plagborg-Møller, M., and Qian, E. (2022). SVAR Identification from Higher Moments: Has the Simultaneous Causality Problem Been Solved? *AEA Papers and Proceedings*, 112:481–85.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.

Newey, W. K. (1991). Estimation of tobit models under conditional symmetry. In Barnett, W. A., Powell, J., and Tauchen, G. E., editors, *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, International Symposia in Economic Theory and Econometrics. Cambridge University Press.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In Grenander, U., editor, *Probability and Statistics, the Harald Cramér Volume*. Wiley, New York, USA.

Neyman, J. (1979). C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 41(1/2):1–21.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297.

Powell, M. J. D. (1981). *Approximation Theory and Methods*. Cambridge University Press, Cambridge, UK.

Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons, Inc., New York, NY, USA.

Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 18(4):375–389.

Remmert, R. (1991). *Theory of Complex Functions*. Springer.

Rieder, H. (2014). One-sided confidence about functionals over tangent cones. arXiv:1412.1701.

Risk, B. B., Matteson, D. S., and Ruppert, D. (2019). Linear non-gaussian component analysis via maximum likelihood. *Journal of the American Statistical Association*, 114(525):332–343.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20 – 71.

Romano, J. P. and Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, 40(6):2798 – 2822.

Royden, H. and Fitzpatrick, P. (2010). *Real Analysis*. Pearson Prentice Hall.

Rudin, W. (1987). *Real & Complex Analysis*. McGraw Hill.

Rudin, W. (1991). *Functional analysis*. McGraw Hill, Inc., second edition.

Sen, A. (2012). On the Interrelation Between the Sample Mean and the Sample Variance. *The American Statistician*, 66(2).

Serfozo, R. (1982). Convergence of Lebesgue Integrals with Varying Measures. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 44(3):380–402.

Sims, C. A. (2021). Svar identification through heteroskedasticity with misspecified regimes. working paper.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J. H. and Wright, J. H. (2000). GMM with weak identification. *Econometrica*, 68(5):1055–1096.

Strasser, H. (1985). *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*. De Gruyter studies in mathematics. W. de Gruyter.

Swensen, A. R. (1985). The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend. *Journal of Multivariate Analysis*, 16(1):54–70.

Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer.

Tank, A., Fox, E. B., and Shojaie, A. (2019). Identifiability and estimation of structural vector autoregressive models for subsampled and mixed-frequency time series. *Biometrika*, 106(2):433–452.

Tinbergen, J. (1939). *Statistical Testing of Business Cycle Theories: Part I: A Method and Its Application to Investment Activity*.

van den Berg, G. J. (2001). Chapter 55 - duration models: Specification, identification and multiple durations. volume 5 of *Handbook of Econometrics*, pages 3381–3460. Elsevier.

van der Vaart, A. W. (1988a). Estimating a Real Parameter in a Class of Semiparametric Models. *The Annals of Statistics*, 16(4):1450 – 1474.

van der Vaart, A. W. (1988b). *Statistical Estimation in Large Parameter Spaces*. Number 44 in CWI Tracts. Centrum voor Wiskunde en Informatica, Amsterdam.

van der Vaart, A. W. (1991). An asymptotic representation theorem. *International Statistical Review / Revue Internationale de Statistique*, 59(1):97–121.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York, NY, USA.

van der Vaart, A. W. (2002). Semiparametric statistics. In Bernard, P., editor, *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXIX - 1999*. Springer.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag New York, Inc., New York, NY, USA.

Velasco, C. (2020). Identification and estimation of structural varma models using higher order dynamics. arXiv:2009.04428.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Willard, S. (1970). *General Topology*. Addison-Wesley.

Willassen, Y. (1979). Extension of Some Results by Reiersøl to Multivariate Models. *Scandinavian Journal of Statistics*, 6(2):89–91.