UNIVERSITAT POLITÈCNICA DE CATALUNYA

Doctorate Program:

SIGNAL THEORY AND COMMUNICATIONS

Ph. D. Thesis

A BAYESIAN APPROACH TO ROBUST IDENTIFICATION. APPLICATION TO FAULT DETECTION

Rosa Mª Fernández Cantí

Supervisors: **Ph. D. Vicenç Puig Cayuela and Ph. D. Joaquim Blesa Izquierdo**

Tutor: **Ph. D. José A. Lázaro Villa**

December 2012

Abstract

In the Control Engineering field, the so-called Robust Identification techniques deal with the problem of obtaining not only a nominal model of the plant, but also an estimate of the uncertainty associated to the nominal model. Such model of uncertainty is typically characterized as a region in the parameter space or as an uncertainty band around the frequency response of the nominal model.

Uncertainty models have been widely used in the design of robust controllers and, recently, their use in model-based fault detection procedures is increasing. In this later case, consistency between new measurements and the uncertainty region is checked. When an inconsistency is found, the existence of a fault is decided.

There exist two main approaches to the modeling of model uncertainty: the deterministic/worst case methods and the stochastic/probabilistic methods. At present, there are a number of different methods, e.g., model error modeling, set-membership identification and non-stationary stochastic embedding. In this dissertation we summarize the main procedures and illustrate their results by means of several examples of the literature.

As contribution we propose a Bayesian methodology to solve the robust identification problem. The approach is highly unifying since many robust identification techniques can be interpreted as particular cases of the Bayesian framework. Also, the methodology can deal with non-linear structures such as the ones derived from the use of observers. The obtained Bayesian uncertainty models are used to detect faults in a quadruple-tank process and in a three-bladed wind turbine.

♦♦♦

Keywords:
Robust Identification
Bayesian Modeling
Fault Detection
Set-membership Identification

A la memòria del meu pare Josep Fernández López
i del profesor Jaume Herranz Luis

Acknowledgment

Aquesta tesi no hauria estat possible sense l'ajut de diverses persones a les quals, des d'aquí, vull expressar el meu agraïment. En primer lloc vull donar les gràcies als meus directors, Vicenç Puig i Joaquim Blesa, per haver acceptat dirigir-me-la i per haver-me ajudat a enllestir-la. Les reunions de treball han estat molt profitoses (crec que amb les idees que han anat sorgint tindrem feina per a anys!) i els ànims que m'han transmès són els que m'han permès tirar-la endavant. També vull donar les gràcies al meu tutor, Jose Antonio Lázaro, per la confiança que ha dipositat en mi i els bons consells que m'ha donat. Però sobretot vull tenir un record per al meu primer director de tesi, el professor Jaume Herranz Luis. M'agrada pensar que aquesta tesi, la última que va dirigir i que no va arribar a veure acabada, porta la seva emprenta tant com per la redacció com per la manera de treballar i enfocar la temàtica. També vull mencionar al professor José M. Miguel, gràcies al qual vaig poder reemprendre el treball i posar ordre a tota la feina que havia fet fins aleshores i qui em va ajudar a mirar-me la tesi com a contribució pràctica i útil per a la societat. Tots dos professors, Herranz i Miguel, representen un model d'Universitat que tant de bo sobrevisqui a tots els canvis dels nostres dies. També vull donar les gràcies a la secretària de doctorat del meu departament, Beni Vázquez, per la seva amabilitat i excel·lent tracte i les continues facilitats amb la paperassa al llarg d'aquests anys. Finalment, vull donar les gràcies a la meva família, en especial a la meva mare i al Ramón, pel seu suport incondicional i per la seva ajuda. Si hi ha algú que ha patit els inconvenients d'aquest treball interminable han estat ells. Desgraciadament el meu pare ja no podrà veure el resultat de tanta feinada, però m'agrada pensar que allà on sigui ara estarà orgullós de mi.

Moltes gràcies a tots.

Contents

List of Acronyms

AME	average modeling error
AR	auto-regressive
ARMA	auto-regressive moving average
ARMAX	auto-regressive moving average with exogenous input
ARX	auto-regressive with exogenous input
BCFR	Bayesian credible frequency response
BCMS	Bayesian credible model set
BCPS	Bayesian credible parameter set
BCVS	Bayesian credible value set
BIBO	bounded input bounded output
BJ	Box Jenkins model
BLUE	best lineal unbiased estimator
BMA	Bayesian model averaging
CMS	candidate model set
CR	Cramér-Rao
FAST	Fatigue, aerodynamics, structures, and turbulence
FBO	feedback block oriented
FDI	fault detection and isolation
FIR	finite impulse response
FMS	feasible model set
FPS	feasible parameter set
GOB	generalized orthonormal basis
HPD	highest posterior density
i.i.d.	independent identically distributed
IBC	information based complexity
LF	likelihood function
LFT	linear fractional transformation
LHP	left-half plane
LS, LSE	least squares, least squares estimation
LTI	linear time-invariant
MAP	maximum *a posteriori*
MCMC	Markov chain Monte Carlo
MEM	model error modeling
MIMO	multiple input multiple output

MISO	multiple input single output
ML, MLE	maximum likelihood, maximum likelihood estimation
MR	minimum risk
NREL	U.S. National Renewable Energy Laboratory
NSSE	non-stationary stochastic embedding
OE	output error
PDF	probability density function
PE, PEM	prediction error, prediction error methods
PRBS	pseudo random binary signal
RHP	right-half plane
RI	robust identification
SIVIA	set inversion via interval analysis
SM, SMI	set-membership, set-membership identification
SVD	singular value decomposition
UBB	unknown but bounded
WLS, WLSE	weighted least squares, weighted least squares estimation
ZOH	zero order hold

Notation and symbols

General:

$*$	superscript * denotes complex conjugate
δ_{ij}	Kronecker delta function with $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$.
$\mathbb{R}$	field of real numbers
$\mathbb{C}$	field of complex numbers
ℓ_p	linear space of p-norm bounded sequences (ℓ_1: magnitude-summable sequences, ℓ_2: square-summable sequences, ℓ_∞: magnitude-bounded sequences.
$\mathcal{H}_\infty$	Hardy space (of rational stable transfer functions)

Common variables:

n, k	discrete-time variable (k is used for fault detection)
q	forward shift operator, $qu_n = u_{n+1}$
d	dimension of the parameter vector $\boldsymbol{\theta}$
N	data set length
δ	error bound in set-membership techniques

System identification:

$\{u_n\}_{n=0}^{N-1}$, $\mathbf{u} = (u_0, \dots, u_{N-1})^T$	excitation signal (samples, matrix notation)
$\{y_n\}_{n=0}^{N-1}$, $\mathbf{y} = (y_0, \dots, y_{N-1})^T$	response signal (samples, matrix notation)
$\{v_n\}_{n=0}^{N-1}$, $\mathbf{v} = (v_0, \dots, v_{N-1})^T$	additive measurement noise signal (samples, matrix notation)
$\{\varepsilon_n\}_{n=0}^{N-1}$, $\boldsymbol{\varepsilon} = (\varepsilon_0, \dots, \varepsilon_{N-1})^T$	residuals (samples, matrix notation)
σ_v^2, λ	variance of the measurement noise
$\boldsymbol{\theta}$	parameter vector
$\boldsymbol{\theta}_{true}$	"true" parameter vector
$\widehat{\boldsymbol{\theta}}_N$	optimal estimate for the parameter vector from N samples
$\widetilde{\boldsymbol{\theta}}_N$	identification error, $\widetilde{\boldsymbol{\theta}}_N = \widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}$
$\boldsymbol{\varphi}_N$	regression (row) vector, $G(q, \boldsymbol{\theta})u_n = \boldsymbol{\varphi}_n^T\boldsymbol{\theta}$
$\boldsymbol{\Phi}$	design matrix, $\boldsymbol{\Phi}^T = (\boldsymbol{\varphi}_0 \quad \cdots \quad \boldsymbol{\varphi}_{N-1})$

$\mathbf{R_N}$	precision matrix, $\mathbf{R_N} = \mathbf{\Phi^T \Phi}$
$\mathbf{P_N}$	covariance matrix, $\mathbf{P_N} = \lambda \boldsymbol{R}_N^{-1}$
$B_k(q)$	(orthonormal) basis functions, $G(q, \boldsymbol{\theta}) = \sum_{k=0}^{d-1} \theta_k B_k(q)$
$\eta,\ \eta_N$	experiment operator (for N samples)
$l(\boldsymbol{\theta}\|\mathbf{y})$	likelihood function
$L(\boldsymbol{\theta}\|\mathbf{y})$	log-likelihood function

Models and model sets:

$\mathcal{G}$	model class
$G(q, \boldsymbol{\theta})$	model (discrete-time transfer function, parameterized by $\boldsymbol{\theta}$)
G_0	nominal model, $G_0 = G(q, \widehat{\boldsymbol{\theta}}_N)$
G_{true}	"true" model
G_e	error model
$\mathcal{B}$	Bayesian credible model set
$\mathcal{B}_{\boldsymbol{\theta}}$	Bayesian credible parameter set
$\mathcal{B}_{\omega}$	Bayesian credible frequency response region
$\mathcal{B}_{\omega_i}$	Bayesian credible value set

Probability and statistics:

$E[\,.\,]$	expectation
$Var[\,.\,]$	variance
$\mathcal{N}(\mu, \sigma^2)$	normal (Gaussian) probability distribution
$X \sim \mathcal{N}(\mu, \sigma^2)$	random variable X is distributed as $\mathcal{N}(\mu, \sigma^2)$
$\chi^2(d)$	chi-squared probability distribution with d degrees of freedom
$\mathcal{U}(a, b)$	uniform probability distribution
$\Pr(A)$	probability of event A
$p(x)$	probability density function (if the random variable X is continuous) or probability mass function (if X is discrete)

Bayesian decision theory

$p(G\|\mathbf{y})$	posterior probability distribution of model G conditioned to the observations $\mathbf{y}$
$p(G)$	prior probability distribution for model G
$p_v(v)$	prior probability distribution for the measurement noise v
$p(\mathbf{y}\|G)$	likelihood function of the observations $\mathbf{y}$ conditioned to G
$c(\alpha)$	critical level for the posterior distribution, $100(1-\alpha)\%$ is the credibility level
$\mathbf{R}_0, \mathbf{P_0}, \boldsymbol{\theta}_0$	prior precision matrix, prior covariance matrix and prior parameter vector
$\mathbf{R}_N, \mathbf{P}(\mathbf{y}), \widehat{\boldsymbol{\theta}}(\mathbf{y})$	posterior precision matrix, posterior covariance matrix and posterior parameter vector
$U(\eta)$	utility function for experiment design
$H_0\ , H_1$	null hypothesis, alternative hypothesis

CHAPTER 1

Introduction

This thesis presents a Bayesian approach to the robust identification problem. In the present chapter we summarize the main features of this problem and its applications.

1.1 Motivation

1.1.1 Model uncertainty

This thesis deals with the problem of modeling model uncertainty. All models are uncertain since they are "only" models, that is, partial representations of reality.

Causes of uncertainty: The causes of the model uncertainty are twofold: *practical* and *theoretical*. The practical issues include the quality of measurement instruments, the effect of operating points, aging, tolerance of components, and so on. The theoretical causes include the lack of knowledge, the difficulty of modeling and the model simplification. The latter is very common in the Control Engineering field since, even if we can produce a good and detailed model of the plant behavior, we always prefer to use simplified versions in order to get controllers of low complexity. This way, all the "undesirable" characteristics (as high frequency poles, smooth nonlinearities, and so on) are treated as uncertainty of the former nominal model.

Characterization of model uncertainty: An uncertain model can be represented by means of a *model set*. Given a physical plant, we can obtain several models of its dynamical behavior. The model set contains all these models. In particular, it includes

the nominal model that is to be used in the controller design. But it can include much more complicated models containing the dynamics that have been neglected in the nominal model. And, in the hypothetical case that a "true", perfect model exists, this model has to be included in the model set as well.

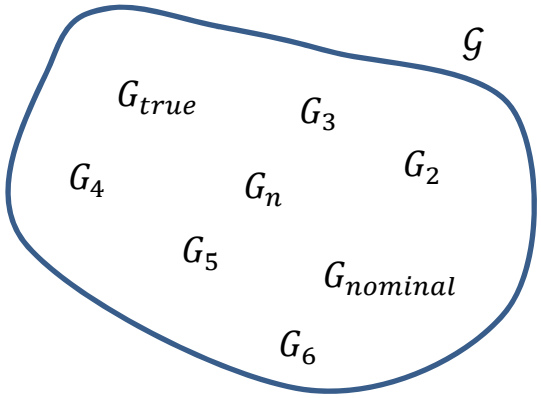


Fig. 1.1. Model set

Types of uncertainty models: There exist two main approaches to analytically describe the model set. We can speak of parametric (structured) uncertainty and dynamic (unstructured) uncertainty.

In the *parametric* case we assume that the model structure is correct and that the only source of error is in the values of the parameters, e. g.,

$$\mathcal{G} = \left\{ G(s): \quad G(s) = \frac{1}{s^2 + as + 1} \;,\; a_{min} \le a \le a_{max} \right\} \tag{1}$$

This type of uncertainty leads to uncertainty regions in the parameter space.

By contrast, in the *dynamic* case we assume that the nominal model $G_0(s)$ is not able to completely describe the plant dynamics and hence a (dynamical) model error term has to be included.

If the error model is added to the nominal model, we obtain an additive (absolute) description of the uncertainty,

$$\mathcal{G} = \{ G(s): \quad G(s) = G_0(s) + \Delta_a(s) \} \tag{2}$$

And if the error term is multiplying the nominal model, we speak of multiplicative (relative) uncertainty,

$$\mathcal{G} = \{ G(s): \quad G(s) = G_0(s)[1 + \Delta_m(s)] \} \tag{3}$$

The dynamic uncertainty leads to uncertainty regions in the frequency domain (uncertainty bands around the Bode plots, or frequency-to-frequency uncertainty regions on the Nyquist and Nichols planes).

Application of uncertainty models: Uncertainty models have been widely used in the design of *robust controllers* but, recently, their use in model-based *fault detection* techniques is increasing. Let us summarize the main features of these two application fields.

1.1.2 Application to robust control

Robust control techniques: A robust controller is a controller that provides stability and performance to all the models that are inside the model set, and not only to the nominal model. Actually, the "robustness" property refers to robustness in front to the model uncertainties.

In the last decades, most works on robust control theory have placed the emphasis on the controller synthesis procedures, and as a result of these efforts current algorithms are quite efficient and reliable (Skogestad and Postlethwaite, 1996), (Sánchez Peña and Sznaier, 1998), (Zhou with Doyle, 1998), (Houpis, *et al.*, 2006), (Chiang *et al*., 2007).

Robust control techniques consist of two stages: The *formulation stage*, which includes the selection of the control system specifications and the plant modeling (nominal model with uncertainty bounds); and the *solution stage*, which is the controller synthesis procedure/algorithm.

Importance of the formulation stage: It turns out that the formulation stage is more critical than the solution stage. In fact, even the best synthesis algorithm may fail, or lead to a useless design, for hard design trade-offs or for plants that are not well characterized.

An example of this is the so-called "spill-over effect" which consists in the degradation of the controller performance due to the excitation of unmodeled dynamics. This is a phenomenon typical of lightly damped flexible structures which are distributed parameters systems and thus have infinite dimensional analytic models (Balas, 1982). In these applications, it is very important to derive not only a good, reduced order nominal model but also good uncertainty bands.

Not so dramatic, a much more common situation is that too pessimistic quantifications of model uncertainty yield to designs where the control system performance is penalized in order to attain a large robustness degree that is actually not necessary.

Finally, the appropriate uncertainty characterization is also an important issue in the formulation stage since it allows establishing high performance yet realistic specifications on the basis of the design trade-offs and performance limits (Seron, Braslavsky, and Goodwin, 1997).

Robust identification problem: The problem of obtaining uncertainty models is known as the "robust identification problem". This is a short version of the original name "robust control-oriented identification", which indicates that this research field was initiated for use in the robust control techniques ($\mathcal{H}_\infty$ and ℓ_1). Several seminal works are (Helmicki, Jacobson, and Nett,1991), (Milanese and Vicino, 1991) and (Gu and Khargonekar, 1992).

1.1.3 Application to fault detection

Fault detection: In Control Engineering, a *fault* is an undesirable deviation from the normal operation of at least one system property or parameter. The consequence of a fault is the degradation of the system performance and in some cases it may be catastrophic for the system or human operators. The purpose of the two fields known as Fault Detection and Isolation (FDI) and Fault Diagnosis is to detect, isolate and identify the faults affecting the system.

Model-free and model-based approaches: FDI can be accomplished by a *model-free* approach or by a *model-based* approach. The model-free approach includes techniques such as the introduction of sensor redundancy, the use of special sensors, and the application of spectrum and statistical analysis tools (Zanardelli *et al*., 2007), (Tharrault *et al*., 2009).

On the other hand, the model-based approach relies on the concept of analytical redundancy, i.e., the consistency between the measurements of the physical system and the information contained in a model is checked. The resulting differences are called the *residuals*. A fault is detected/decided when a residual is greater than a given threshold or when an estimated parameter abnormally deviates.

Uncertainty and false alarms: To avoid false alarms, the model-based fault detection system must be *robust*, i.e., it must be sensitive *only* to faults, even in the presence of model uncertainty. However, since a model is only an approximate representation of reality, residuals may be nonzero even in the absence of faults. These modeling errors should not be detected as faults. To solve this problem *active* and *passive* methods have been developed.

Active and passive methods: Active methods aim to generate residuals that are insensitive to uncertainty but not to faults. Main methods include the use of unknown input observers, eigenstructure assignment and structured parity equations. See the books of (Chen and Patton, 1999), (Blanke *et al*., 2003) and (Ding, 2008) for a survey.

Passive methods use robust identification techniques to describe the fault-free uncertain system. The uncertainty is characterized by bounded regions, in the parameter space or in the state space, that are consistent with the measurements. When a new measurement

is inconsistent with the uncertainty set, a fault is decided. The major drawback of this approach is that the fault will be not detected if it enters inside the bounded region, thus the importance to derive tight uncertainty regions.

Passive robust model based methods: The passive robust model-based approach has received a lot of attention in the last years. Main methods include the use of the bounding approach in the parity space (Ploix and Adrot, 2006), the development of new set-membership techniques (Blesa, 2011a), and the use of diagnostic interval observers (Puig *et al*., 2008), (Raïssi *et al*., 2010). Some of these methods can deal with nonlinear systems, e.g., on the basis of subpaving algorithms or multimodel approaches (Letellier *et al*., 2011). Most passive methods use deterministic regions, but recently probabilistic credible regions are receiving attention; see e.g. (Jaulin, 2010).

1.1.4 Main approaches to robust identification

Deterministic/worst case methods and *stochastic/probabilistic methods* constitute the main solutions to the robust identification problem. Current research in both approaches is mainly focused in improving the performance of the identification algorithms and in obtaining tighter uncertainty bands. Chapter 2 explains in detail the deterministic and stochastic approaches, but here we list the main techniques.

Stochastic methods: Stochastic methods, such as the *non-stationary stochastic embedding* (NSSE) (Goodwin *et al*., 2002) and the stochastic versions of the *model error modeling* (MEM) approach such as the ones based in *prediction error methods* (PEM) (Reinelt *et al*., 2002), enjoy a low computational load compared to deterministic methods. However, they make little use of possible prior information about the system to be modeled. As a result, the obtained nominal model can be too *biased* and the associated uncertainty bands may result too *pessimistic*.

Deterministic methods: By contrast, deterministic methods such as the worst case system identification in $\mathcal{H}_\infty$ (Chen and Gu, 2000) and other methods based on the *set-membership identification* (SMI) paradigm (Milanese and Taragna, 2005) are computationally intensive but they do consider explicitly any possible prior information about the plant and measurement noise by means the definition of the so-called *feasible model set* (FMS).

1.1.5 Shortcomings of current robust identification methods

Apart from the controversy between the defenders of the deterministic viewpoint and the defenders of the stochastic viewpoint, we have no knowledge about the existence of a *conclusive work* in favor of an approach or particular method over the others.

Also, since robust identification embraces a wide variety of methods and techniques, it is difficult to point out common drawbacks. Moreover many times the identification

procedure is tailored to the particular application. There exist very few works *comparing* the performance of the different methods and the resulting robust controllers. A relevant one is (Reinelt, Garulli, and Ljung, 2002), where suboptimal SMI, PEM-based MEM, and NSSE methods are compared. The author's conclusion is that all three methods present very similar performance and results. Other related works are (Esmaeilsabzali *et al*., 2006), (Herrero, 2006), and (Raafat *et al*., 2009).

In this dissertation, after a study of the existing robust identification literature we have identified the following weak points:

Computational load: In general, deterministic methods are computationally intensive compared to stochastic approaches. This is justified by the necessity of considering all possible plant perturbations and it is especially unavoidable when nonlinear structures are considered. However, in many practical cases the trade-off between the computational burden and the final uncertainty region obtained is somehow deceiving. A comparison of the computational cost of several worst case $\mathcal{H}_\infty$ identification algorithms can be found in (Milanese and Taragna, 2005).

Size of uncertainty regions: It is clear that the uncertainty bands must be kept small while retaining all relevant plant perturbations. The problem is that the size of the uncertainty regions is very *sensitive* to the assumptions taken during the modeling procedure. In general, stochastic methods yield smaller bands, which size depends on the chosen probability level (Goodwin, Braslavsky, and Seron, 2002).

In SMI approaches uncertainty regions can be tightened by means the introduction of prior knowledge about the plant and noise. In fact, SMI approaches make a more *efficient* use of prior knowledge than MEM methods, which usually limit their use to the selection of the nominal model order.

Reliability of prior knowledge: However, in order to be useful, prior knowledge must be *reliable*, since the size of the uncertainty regions is very *sensitive* to it. For instance, a pessimistic choice of the noise bound δ may produce too much large uncertainty regions. This particular sensitivity problem is considered in (Ninness and Goodwin, 1995). Another example is the selection of the basis functions that are mostly used to define the model structure. Poles of such bases are usually selected after a spectral analysis of the data. If the plant presents resonant modes, the selection of (the number and value) of basis poles is easy. However the selection is not so clear if the modes are real. Methods for pole selection do exist, see e.g. the *average modeling error* (AME), but they are computationally intensive since they imply computing several models and analyze which one presents better performance (see (Reinelt, Garulli, and Ljung, 2002)).

Unfortunately, current methods do not allow knowing if the prior assumptions used are erroneous or not. It has to be said that prior knowledge "checking" is implicitly included in the FMS unfalsification stage, but only grossly erroneous prior assumptions can be detected.

Control purposes: Another criticism to current methods is related to the control purposes. We feel that many times the obtained models are not as oriented to robust control as they could be. In fact, no information regarding the final control system is considered in the modeling procedure. Most of times, the only requirement is that the nominal model must be of restricted complexity to produce low order robust controllers. It has to be said though that closed loop identification schemes (Van den Hof and Schrama, 1995) and integrated identification-control strategies (Cooley and Lee, 1998) exist to overcome this problem.

To our knowledge, none of SMI and MEM existing methods considers the cost of "wrong modeling" when selecting a nominal model nor when obtaining the uncertainty bounds. It is known that an educated selection of the nominal model leads to smaller uncertainty bounds (see e.g. (Skogestad and Postlethwaite, 1996)), therefore it would probably lead to better robust high-performance designs. Since it is clear that in practice many frequencies are more critical than others, it seems reasonable to impose some kind of model penalty at least at such frequencies.

An interesting example of the consequences of a "blind" uncertainty modeling can be found in (Onatski and Williams, 2002). Also, some critical papers have appeared (Douma and Van den Hof, 2005), which evidence that the usual models for robust control are much more models for *a posteriori* robustness *analysis* (once the control system is designed) than for robust controllers *design*.

1.1.6 The Bayesian viewpoint

The Bayesian solution: To overcome the shortcomings discussed in the previous section, in this thesis we propose a Bayesian methodology for solving the robust identification problem. We think that the Bayesian framework is adequate for the following reasons.

- From a general viewpoint, Bayesian Confirmation Theory[1] is concerned precisely to *model building*, with strong relations to concepts such as induction/deduction, statistical inference, meaning of probability, and validity of scientific theories.

- From a particular (robust identification) viewpoint, it allows the formal description of the prior information (or lack of it –prior ignorance), it allows the efficient combination of prior information about the model with experimental information in order to obtain a posterior distribution of how likely the different models are, and it allows the selection of a nominal model on the basis of some minimum risk criteria.

[1] Bayesian Science Theory is known as *Bayesian Confirmation Theory* and it concerns the validation of scientific theories from a Bayesian viewpoint. In the science context, the subject is how to assign probabilities to theories or hypotheses h in the light of the evidence e. Bayes formula tells us how to modify the probability of one hypothesis $\Pr(h)$ in order to attain a new and revised probability on the light of any specified evidence $\Pr(h|e)$, $\Pr(h|e) = \Pr(h)\frac{\Pr(e|h)}{\Pr(e)}$.

To enforce our viewpoint, note that modeling model uncertainty for robust control and/or fault detection is only one particular application of the general problem of uncertainty modeling that arises in any scientific or technical discipline. So, it seems reasonable use the tools that are already developed for other areas, in particular, statistical tools. Quoting (Berger, 2000), who is a convinced Bayesian:

> "Statistics is about measuring uncertainty, and over 50 years of efforts to prove otherwise have convincingly demonstrated that the only coherent language in which to discuss uncertainty is the Bayesian language"

Works regarding Bayesian uncertainty modeling: Recently, there has been a renewed interest for the Bayesian point of view in system identification (Ninness and Henriksen, 2010), (Schön *et al*., 2011). The topic is not new since early works in system identification already considered the Bayesian parameter estimation (Eykhoff, 1974) and model classification (Peterka, 1981). The Bayesian ideas, although appealing, have largely not been implemented due to the difficulty of computing the integrals involved in the posterior distributions. Recent advances in simulation techniques such as Monte Carlo Markov chains (MCMC) have overcome this situation (Robert and Casella, 1999), (Chen, Shao, and Ibrahim, 2000), (Bolstad, 2010).

If fact, we are surprised of the little number of works relating Bayesian modeling with Robust Control, even more when the terminology used in robust identification (*a priori* information, *a posteriori* information) directly points out to the Bayesian terminology. This is not the case in the field of Fault Detection and Diagnosis where some recent Bayesian references are (Lee, 2008), (Pernestål, 2009), and (Dearden, 2010).

To finish this section, we list some of the Bayesian works in other Engineering areas. For an overview of the activity in the field of Bayesian analysis, see (Berger, 2000) and the references therein. In (Hoeting *et al*, 1999) it can be found a survey about Bayesian Model Averaging (BMA), which is a technique to reduce the uncertainty inherent in the model selection process. In the field of structures engineering we can refer to the use of Bayesian conjugate distributions (Igusa *et al*, 2002), model updating (Papadimitriou, Beck, and Katafygiotis, 2001), and model identifiability, (Katafygiotis and Beck, 1998). In the field of reliability analysis, hierchical uncertainty models are used in (Utkin, 2003), risk analysis is treated in (Apeland *et al.*, 2002), and a discussion about evidence *vs*. Bayes can be found in (Soundappan *et al*., 2004). In Econometrics, Bayes factors are used in (Cairns, 2000), and an application of the Bayesian estimation is presented in (Onatski and Williams, 2002). Finally, in the Ecology field we can refer to the uncertainty analysis of (Borsuk *et al*., 2004), the insight about Monte Carlo methods in (Qian *et al*., 2003), and the use of the Bayesian state space modeling for hydrology in (Cornford, 2004).

1.2 Objectives and scope

In this thesis we propose a Bayesian methodology to solve the robust identification problem. The particular objectives are the following:

Bayesian Credible Model Set: Characterization of a stochastic Bayesian Credible Model Set $\mathcal{B}$ inspired in the Feasible Model Set (FMS) of deterministic methods. Instead of some norm of the residuals, $\mathcal{B}$ will be expressed in terms of the posterior probability distributions of the model G conditioned to the measurement data $\mathbf{y}$, $p(G|\mathbf{y})$. The robust identification problem will be formulated in terms of $\mathcal{B}$. It will be shown the relationship between $\mathcal{B}$ and the existing deterministic and stochastic methods, and the results will be compared by means of several examples.

Case of parametric uncertainty: Characterization of $\mathcal{B}$ when the support for the probability distributions is the parameter space. The Bayes' rule will be used to derive *analytical* expressions for the model posterior distribution in the case of linear regression models and Gaussian probability distributions (for both the parameters and the measurement noise). For high order models and arbitrary non-conjugate probability distributions, *simulation* methods based on Markov Chain Monte Carlo (MCMC) integration will be used.

Computation of the Highest Posterior Density (HPD) credible regions that constitute the uncertainty description in the Bayesian framework. These credible regions will be compared to the ones obtained by means of classical confidence regions. For the Gaussian case, exact credible regions will be derived, assuming that the noise variance is known and is unknown. For the case where the number of parameters increases and/or the distributions are non-Gaussian, the credible regions will be obtained by means MCMC techniques.

Case of dynamic uncertainty: Definition of $\mathcal{B}$ for frequency domain data. In this case the support for the probability distributions will be the complex plane. Since several sources of uncertainty may be present (i.e. uncertainty in the structure and parameters) hierarchical priors and sets of competing models will be used.

Mixture prior distributions in the Nyquist plane will be derived for the case of linear models expressed by means a set of basis functions and Gaussian distributions. Bayes's rule and the law of total probability will be used to compute the mixture posterior distributions frequency to frequency. Finally, HPD credible regions in the Nyquist plane will be obtained.

Application to fault detection: The iterative computation of the likelihood function assuming uniform noise will be used to detect faults in a quadruple-tank process. Multiple Input Single Output (MISO) case, MISO case with observer and Multiple Input Multiple Output (MIMO) case will be considered and compared to set-membership techniques.

For the case of a three-bladed wind turbine, deterministic uncertainty regions will be obtained, and sensor and actuator faults will be detected assuming both uniform noise and Gaussian noise.

1.3 Outline

The outline of this dissertation is as follows.

Chapter 2 summarizes the state of the art of robust identification. In particular, conventional system identification, stochastic methods and deterministic methods are presented. Several examples of the literature are provided in order to compare these techniques to the Bayesian technique explained in Chapter 3.

Chapter 3 is focused on the proposed Bayesian methodology to solve the robust identification problem. The Bayesian Credible Model Set $\mathcal{B}$ is defined and characterized. The construction of $\mathcal{B}$ in the parametric case and in the frequency domain is illustrated. It is explained how to obtain the credible regions that constitute the uncertainty regions, and several interesting features, such as the iterative computation of the regions or the effect of the prior distributions, are illustrated.

Chapter 4 illustrates the application of the credible regions to the fault detection problem. Two case studies are considered: a quadruple tank process and a three-bladed wind turbine. In the first application, MISO case, MISO case with observer and MIMO case are considered and compared to set-membership techniques. In the second application, deterministic uncertainty regions are obtained and used for fault detection assuming uniform noise and Gaussian noise.

Finally, Chapter 5 draws the conclusions of this work and point out several lines for future research.

Additionally to the previous chapters, several appendices are provided to introduce complementary material:

In Appendix A, we summarize some concepts of the Optimal Estimation Theory that are used in this thesis. The point estimation problem and the set estimation problem are presented. The maximum likelihood estimation technique is treated in detail, and a comparison between the main point estimators is presented.

Appendix B is focused to the study of the orthonormal basis functions that are used in system identification. Laguerre, Kautz, and generalized functions are presented for linear systems, and bases for the Wiener and Hammerstein models are presented for the case of nonlinear systems.

Appendix C summarizes the simulation techniques known as Markov Chain Monte Carlo (MCMC). In particular, the Metropolis Hastings algorithm, the Gibbs sampler, and the reversible jump algorithm are explained.

Finally, Appendix D contains many definitions of the Bayesian decision theory, and presents the fundamentals of the Bayesian modeling, including the concept of subjective priors.

CHAPTER 2

State of the Art of Robust Identification

The present chapter summarizes the main current methods for the identification of model uncertainty. As a preliminary result, in Section 2.1, classical system identification is presented. Specific Robust Identification methods, stochastic and deterministic, are treated in Sections 2.1 and 2.2 respectively.

2.1 Classical system identification

The so-called Prediction Error Methods (PEM) constitute the classical solution to the problem of system identification (Eykhoff, 1974), (Goodwin and Payne, 1977), (Söderström and Stoica, 1989), (Schoukens and Pintelon, 1991), (Ljung, 1999a).

In this section we illustrate how this approach obtains a nominal model and characterizes the uncertainty around it by means of the computation of the confidence regions. For simplicity, we focus on the Output Error (OE) linear model case and quadratic cost function.

2.1.1 Nominal model

The experiment: Let us assume that we have collected N input/output measurement data obtained by applying an excitation sequence $\{u_n\}_{n=0}^{N-1}$ to an unknown system G_{true} and collecting the response samples $\{y_n\}_{n=0}^{N-1}$ corrupted by additive measurement noise $\{v_n\}_{n=0}^{N-1}$,

$$y_n = G_{true}(q)u_n + v_n \quad , \quad n = 0, \dots, N-1 \tag{4}$$

where q is the forward shift operator, $qu_n = u_{n+1}$. We assume that $\{v_n\}_{n=0}^{N-1}$ is a sequence of i.i.d. (independent identically distributed) noise with variance σ_v^2 and that it is independent to the excitation $\{u_n\}_{n=0}^{N-1}$. To simplify the notation, we define $\mathbf{y} = (y_0, \dots, y_{N-1})^T$, $\mathbf{u} = (u_0, \dots, u_{N-1})^T$, and $\mathbf{v} = (v_0, \dots, v_{N-1})^T$.

Cost function: The objective is to get an estimate $\hat{G}$ of G_{true} from the experimental data $\{u_n, y_n\}_{n=0}^{N-1}$. The simplest approach is to compute a model $G(q, \boldsymbol{\theta})$ parameterized by means of a d-dimension parameter vector $\boldsymbol{\theta}$, which fits the experimental data by minimizing the Euclidean norm of the prediction error, $\varepsilon_n \equiv y_n - G(q, \boldsymbol{\theta})u_n$. If we define such a cost function,

$$V_N = \frac{1}{N}\sum_{n=0}^{N-1} \varepsilon_n^2 \tag{5}$$

the optimal estimate, $\hat{\boldsymbol{\theta}}_N$, which will be selected as the *nominal* model, is the Least Squares Estimate (LSE):

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}} V_N \tag{6}$$

This solution is a particular case of the Maximum Likelihood Estimation (MLE). See Appendix A for details.

Model structure: The computation of $\hat{\boldsymbol{\theta}}_N$ depends on the selected structure for the model. The simplest case is when the parameter vector $\boldsymbol{\theta}$ parameterizes $G(q, \boldsymbol{\theta})$ linearly, $G(q, \boldsymbol{\theta})u_n = \boldsymbol{\varphi}_n^T\boldsymbol{\theta}$, where the row vector $\boldsymbol{\varphi}_n^T$ is the regression vector.

In AR (Auto Regressive) models, the parameter vector $\boldsymbol{\theta}$ contains the coefficients of both the model numerator and denominator polynomials, $B(q)$ and $A(q)$ respectively, and the regression vectors $\boldsymbol{\varphi}_n^T$ are built using the previous samples of the input and output, for instance, $\boldsymbol{\varphi}_n^T = (u_{n-1} \;\; u_n \;\; -y_{n-2} \;\; -y_{n-1})$. See Appendix B.

Another common linear structure is

$$G(q, \boldsymbol{\theta}) = \sum_{k=0}^{d-1} \theta_k B_k(q) \tag{7}$$

where $B_k(q)$ are fixed functions. These functions contain any prior information that we already have about the system to be identified and that we do not want to estimate from the measurement data (for instance, the poles position). In the Robust Identification field, (7) is the most used structure and $B_k(q)$ are often selected as the orthonormal basis functions of some series expansion (trigonometric, Laguerre, Kautz, generalized). See Appendix B.

If we use the basis functions $B_k(q)$, the regression vectors $\boldsymbol{\varphi}_n^T$ are deterministic and given by

$$\boldsymbol{\varphi}_n^T = (B_0(q)u_n \quad \dots \quad B_{d-1}(q)u_n) \tag{8}$$

The (noiseless) *n*-th sample of the model output is then

$$G(q,\boldsymbol{\theta})u_n = (B_0(q)u_n \quad \dots \quad B_{d-1}(q)u_n)\begin{pmatrix}\theta_0\\ \vdots \\ \theta_{d-1}\end{pmatrix} = \boldsymbol{\varphi}_n^T\boldsymbol{\theta} \tag{9}$$

and the all *N* samples of the model output are

$$G(q,\boldsymbol{\theta})\mathbf{u} = \begin{bmatrix} B_0(q)u_0 & \cdots & B_{d-1}(q)u_0 \\ \vdots & & \vdots \\ B_0(q)u_{N-1} & \cdots & B_{d-1}(q)u_{N-1}\end{bmatrix}\begin{pmatrix}\theta_0\\ \vdots \\ \theta_{d-1}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\varphi}_0^T\\ \vdots \\ \boldsymbol{\varphi}_{N-1}^T\end{pmatrix}\begin{pmatrix}\theta_0\\ \vdots \\ \theta_{d-1}\end{pmatrix} = \boldsymbol{\Phi}\boldsymbol{\theta} \tag{10}$$

where matrix $\boldsymbol{\Phi}$ is addressed as *design matrix*. The experiment (4) in matrix notation is then expressed as $\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \mathbf{v}$.

Estimation of the nominal parameter vector: In the linear case, the solution presents a closed expression that can be easily obtained. Since the cost function $V_N = \frac{1}{N}\sum_{n=0}^{N-1}[y_n - \boldsymbol{\varphi}_n^T\boldsymbol{\theta}]^2$ is quadratic in $\boldsymbol{\theta}$, we can obtain its minimum value by cancelling its derivative with respect to $\boldsymbol{\theta}$:

$$0 = \frac{d}{d\boldsymbol{\theta}}V_N\Big|_{\widehat{\boldsymbol{\theta}}_N} = \frac{1}{N}\sum_{n=0}^{N-1}2(-\boldsymbol{\varphi}_n)[y_n - \boldsymbol{\varphi}_n^T\boldsymbol{\theta}]\Big|_{\widehat{\boldsymbol{\theta}}_N}$$

Remark: To obtain the expression above, the result $\frac{d\boldsymbol{\varphi}_n^T\boldsymbol{\theta}}{d\boldsymbol{\theta}} = \boldsymbol{\varphi}_n$ has been used.

The last equation can be expressed as

$$\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n y_n = \sum_{n=0}^{N-1}\boldsymbol{\varphi}_n\boldsymbol{\varphi}_n^T\widehat{\boldsymbol{\theta}}_N$$

where, by isolating the parameter vector, we have

$$\widehat{\boldsymbol{\theta}}_N = \left[\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n\boldsymbol{\varphi}_n^T\right]^{-1}\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n y_n \tag{11}$$

In matrix notation, the resulting nominal parameter vector that characterizes the *nominal model* is

$$\widehat{\boldsymbol{\theta}}_N = \mathbf{R}_N^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} \quad , \quad \mathbf{R}_N = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} \tag{12}$$

where $\mathbf{R}_N$ is called the *precision matrix* and $\boldsymbol{\Phi}^{\mathrm{T}} = (\boldsymbol{\varphi}_0 \quad \cdots \quad \boldsymbol{\varphi}_{N-1})$.

2.1.2 Uncertainty characterization

Mean and variance of the estimation error: To characterize the uncertainty around the nominal model, it is worth noting that this approach assumes that the true system can be totally described by a d-dimension parameter vector $\boldsymbol{\theta}_{true}$. Therefore, the identification error $\widetilde{\boldsymbol{\theta}}_N = \widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}$ depends only on the measurement noise and the data length. To determine the quality of the estimate one can obtain the mean value and covariance matrix of this error.

The response of the true plant is $y_n = \boldsymbol{\varphi}_n^T\boldsymbol{\theta}_{true} + v_n, \;\; n = 0, \ldots, N-1$. Substituting this value y_n in Equation (11) we have

$$\widehat{\boldsymbol{\theta}}_N = \left[\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n\boldsymbol{\varphi}_n^T\right]^{-1}\left[\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n\boldsymbol{\varphi}_n^T\boldsymbol{\theta}_{true} + \sum_{n=0}^{N-1}\boldsymbol{\varphi}_n v_n\right]$$

and thus

$$\widetilde{\boldsymbol{\theta}}_N = \widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true} = \left[\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n\boldsymbol{\varphi}_n^T\right]^{-1}\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n v_n$$

In matrix notation, the parametric error is $\widetilde{\boldsymbol{\theta}}_N = \mathbf{R}_N^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{v}$.

Since the measurement noise and the excitation are independent, the sequences $\{v_n\}_{n=0}^{N-1}$ and $\{\boldsymbol{\varphi}_n\}_{n=0}^{N-1}$ are independent too. Moreover $\{\boldsymbol{\varphi}_n\}_{n=0}^{N-1}$ is a deterministic sequence if the structure (7) is selected. Hence, the expected value of the error $\widetilde{\boldsymbol{\theta}}_N$ is $E\left[\widetilde{\boldsymbol{\theta}}_N\right] = \mathbf{R}_N^{-1}\boldsymbol{\Phi}^{\mathrm{T}}E[\mathbf{v}]$, where $E[\mathbf{v}]$ is the mean value of the measurement noise. If $E[\mathbf{v}]$=0 (which is the usual case), the estimate $\widehat{\boldsymbol{\theta}}_N$ is unbiased, $E\left[\widehat{\boldsymbol{\theta}}_N\right] = \boldsymbol{\theta}_{true}$.

On the other hand, the covariance matrix of the error $\widetilde{\boldsymbol{\theta}}_N$, $\mathbf{P}_N = E\left[\left(\widetilde{\boldsymbol{\theta}}_N - E\left[\widetilde{\boldsymbol{\theta}}_N\right]\right)\left(\widetilde{\boldsymbol{\theta}}_N - E\left[\widetilde{\boldsymbol{\theta}}_N\right]\right)^T\right]$, is

$$\mathbf{P}_N = E\left[\widetilde{\boldsymbol{\theta}}_N\widetilde{\boldsymbol{\theta}}_N^T\right] = \mathbf{R}_N^{-1}\boldsymbol{\Phi}^T E[\mathbf{v}\mathbf{v}^T]\boldsymbol{\Phi}(\mathbf{R}_N^{-1}) = \sigma_v^2\mathbf{R}_N^{-1}.$$

Note that $\mathbf{P}_N$ depends on the value of the measurement error variance, σ_v^2, which is unknown, but it can be estimated from the experimental data. Lemma II.1 in (Ljung, 1999a, p554) establishes that an unbiased estimate for σ_v^2 is the following:

$$\hat{\sigma}_v^2 = \frac{1}{N-d}\sum_{n=0}^{N-1}\left[y_n - \boldsymbol{\varphi}_n^T\widehat{\boldsymbol{\theta}}_N\right]^2 \tag{13}$$

Probability distribution of the estimation error: Even though the probability distribution of the measurements is not normal, the usual case is that the probability distribution of $\widehat{\boldsymbol{\theta}}_N$ tends to be normal as the number of samples N tends to infinity. This is a consequence of the application of the Central Limit Theorem to the sum of random variables $\{y_n\}_{n=0}^{N-1}$ that constitutes the estimate, see (Ljung, 1999a, p556).

Thus, if we assume that a $\boldsymbol{\theta}_{true}$ exists, the probability distribution of the parametric error is normal too, $\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true} \sim \mathcal{N}(0, \mathbf{P}_N)$. For the i-th component we have $\widehat{\boldsymbol{\theta}}_N(i) - \boldsymbol{\theta}_{true}(i) \sim \mathcal{N}(0, \mathbf{P}_N(i,i))$. The standard normal distribution is obtained by making

$$\frac{\widehat{\boldsymbol{\theta}}_N(i) - \boldsymbol{\theta}_{true}(i)}{\sqrt{\mathbf{P}_N(i,i)}} \sim \mathcal{N}(0,1)$$

Then, by direct application of the definition of the χ^2 probability distribution, we can write $\left(\widehat{\boldsymbol{\theta}}_N(i) - \boldsymbol{\theta}_{true}(i)\right)^T \mathbf{P}_N^{-1} \left(\widehat{\boldsymbol{\theta}}_N(i) - \boldsymbol{\theta}_{true}(i)\right) \sim \chi^2(1)$ for one component.

> **Remark:** If a random variable X is distributed as $\mathcal{N}(0,1)$, the random variable $Y = X^2$ is distributed as $\chi^2(1)$. (Casella and Berger, 2002, p53).

Finally, for the d components of the parameter vector the result is

$$(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}) \mathbf{P}_N^{-1} (\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}) \sim \chi^2(d)$$

where $\chi^2(d)$ denotes the χ^2 distribution with d degrees of freedom.

The last expression allows defining the confidence regions for the estimate. The probability that

$$(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}) \mathbf{P}_N^{-1} (\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}) \geq \alpha \qquad \textbf{(14)}$$

is $\chi_\alpha^2(d)$, being α the probability level of the distribution $\chi^2(d)$. The resulting regions are ellipsoids in the $\mathbb{R}^d$ space, their shape is determined by $\mathbf{P}_N$ and their size by the probability level α. For the normal distribution case, i.e. $\{v_n\}_{n=0}^{N-1}$ are normal distributed, the confidence regions are exact, otherwise they are only valid asymptotically, for $N \to \infty$.

Example 2.1. PEM uncertainty regions in the parameter space

Let us illustrate the computation of the uncertainty regions in the parameter space. Consider the plant in (Ninness and Goodwin, 1995),

$$G_{true}(s) = \frac{e^{-2s}}{(s+1)(10s+1)}.$$

We collect $N = 2000$ input/output samples obtained by exciting the plant with a square signal of frequency 0.02Hz and d.c. (direct current) level of 0.2V. The measurement

noise is uncorrelated to the excitation and it is a Gaussian process with zero mean and variance 0.005. The sampling time is $T_s = 1s$.

The nominal model, of order 2, is a model based in discrete Laguerre functions

$$B_i(q) = \left(\frac{\sqrt{1-\xi^2}}{q-\xi}\right)\left(\frac{1-\xi q}{q-\xi}\right)^{i-1} \ , \ |\xi| < 1 \ , \ i = 1,2$$

where the pole is located at $\xi = \exp(-0.2T_s)$, corresponding to the continuous time pole at 0.2rad/s. The optimal parameter vector which minimizes the squared prediction error is $\widehat{\boldsymbol{\theta}} = (0.1060 \quad 0.1673)^T$ and thus the resulting nominal transfer function is:

$$G_0(q) = \theta_1 B_1(q) + \theta_2 B_2(q) = \frac{-0.018q+0.046}{q^2-1.637q+0.671}.$$

Fig. 2.1 shows the nominal parameter vector and the confidence ellipses around it for the 0.5, 0.8 and 0.9 probability levels.

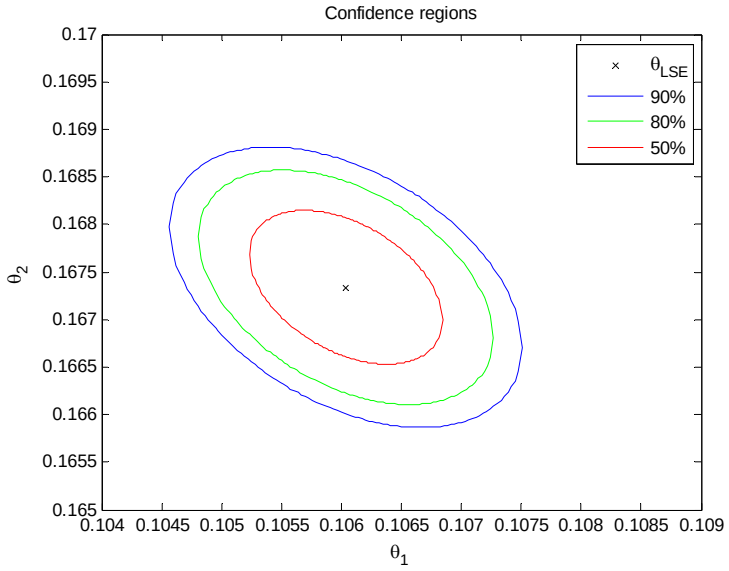


Fig. 2.1. Confidence regions in the parameter space

■

Confidence regions of the model frequency response: The uncertainty region in (14) is expressed in the parameter space. If we want robust control oriented models, we need to translate the confidence region to the frequency domain. Let us illustrate the procedure for the model structure in (7), $G(q, \boldsymbol{\theta}) = \sum_{k=0}^{d-1} \theta_k B_k(q)$.

For each frequency point ω_i, if we define $\mathbf{B}(e^{j\omega_i}) \equiv (B_0(e^{j\omega_i}) \quad \ldots \quad B_{d-1}(e^{j\omega_i}))$, we can write the frequency response of the true system as $G_{true}(e^{j\omega_i}) = \mathbf{B}(e^{j\omega_i})\boldsymbol{\theta}_{true}$. The variance of the estimated frequency response is:

$$E\left[\left|G(e^{j\omega_i}, \widehat{\boldsymbol{\theta}}_N) - G_{true}(e^{j\omega_i})\right|^2\right] = \mathbf{B}(e^{j\omega_i})\mathbf{P}_N\mathbf{B}^*(e^{j\omega_i}) \qquad \textbf{(15)}$$

where the symbol * means conjugate transpose.

Remark: To prove (15) note that the variance is $E\left[\left(\mathbf{B}\widehat{\boldsymbol{\theta}}_N - \mathbf{B}\boldsymbol{\theta}_{true}\right)\left(\mathbf{B}\widehat{\boldsymbol{\theta}}_N - \mathbf{B}\boldsymbol{\theta}_{true}\right)^T\right] = \mathbf{B}E\left[\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}\right)\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}\right)^T\right]\mathbf{B}^*$.

In order to obtain the equivalent of (14) in the Nyquist plane, we will use the following lemma of (Wahlberg and Ljung, 1992):

Lemma 2.1. Let $\mathbf{x} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} > 0 \in \mathbb{R}^{d\times d}$ and $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} \leq 1$. Then, for $\mathbf{w} = \mathbf{A}\mathbf{x} \in \mathbb{R}^n$, $n \leq d$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$ is full rank, the following result is satisfied: $\mathbf{w}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{w} \leq 1$. ■

In our case, $\boldsymbol{\Sigma} = \mathbf{P}_N$, and

$$\underbrace{\begin{pmatrix} \mathrm{Re}\left(G\left(e^{j\omega_i},\widehat{\boldsymbol{\theta}}_N\right) - G_{true}\left(e^{j\omega_i}\right)\right) \\ \mathrm{Im}\left(G\left(e^{j\omega_i},\widehat{\boldsymbol{\theta}}_N\right) - G_{true}\left(e^{j\omega_i}\right)\right) \end{pmatrix}}_{\mathbf{w}} = \underbrace{\begin{pmatrix} \mathrm{Re}\left(\mathbf{B}\left(e^{j\omega_i}\right)\right) \\ \mathrm{Im}\left(\mathbf{B}\left(e^{j\omega_i}\right)\right) \end{pmatrix}}_{\mathbf{A}} \underbrace{\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{true}\right)}_{\mathbf{x}}$$

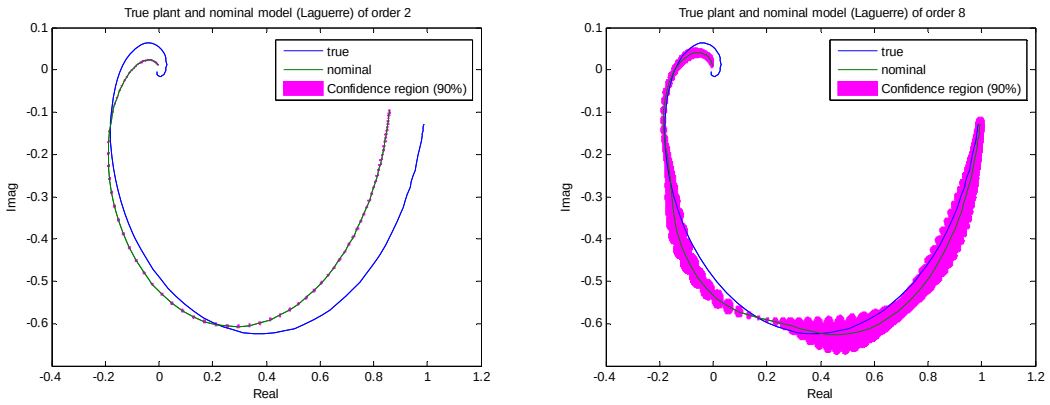
Therefore, the ellipses in the Nyquist plane are defined as:

$$\mathbf{w}^T(\mathbf{A}\mathbf{P}_N\mathbf{A}^T)^{-1}\mathbf{w} \geq \alpha \qquad \textbf{(16)}$$

where the α-level corresponds to the χ^2 distribution with 2 degrees of freedom, $\chi^2(2)$.

Example 2.2. PEM uncertainty regions in the Nyquist plane.

Consider again the Example 2.1 (Ninness and Goodwin, 1995). Now we have obtained the 90% confidence ellipses for the frequency response of the nominal model. Fig. 2.2 shows in the Nyquist plane the results for a Laguerre model of order 2 and for a Laguerre model of order 8 along with the true system frequency response.



(a) **(b)**

Fig. 2.2. True system and nominal model frequency responses. 90% confidence regions for the cases (a) order 2 and (b) order 8

■

In the example above we can see two facts. If the nominal model order is too low, the resulting frequency response is so erroneous (biased) than the confidence ellipses are nonsense. On the other hand, if we increase the model order, the model frequency response approaches the true one, but the size of the uncertainty region (variance) increases. These two effects are a consequence of the so-called bias/variance trade-off explained below.

2.1.3 Bias/variance trade-off

When estimating plant models from a finite set of experimental data, the estimation error can be decomposed in two error terms: the *variance error* and the *bias error*.

Variance error: The variance error is due (1) to the measurement noise corrupting the experimental data and (2) to the finite length N of the sample. In a general case, it is uncorrelated to the excitation signal (if the identification is performed in open loop) and it decreases as the number of samples N increases. The variance error affects uniquely to the model parameter values. If we assume that a model presents only variance error, we are assuming that the model structure is capable of completely describe the plant dynamics. This is what classical PEM assume and, therefore, the confidence regions in Fig. 2.2 are only characterizing the variance error.

In the case of open loop identification and quadratic cost function, the variance error satisfies $\lim_{d\to\infty}\lim_{N\to\infty}\frac{N}{d}Var[G_0]=\frac{\Phi_v(\omega)}{\Phi_u(\omega)}$. This expression is usually substituted by the following approximate result (Ljung, 1999a):

$$Var[G_0]\approx\frac{d}{N}\frac{\Phi_v(\omega)}{\Phi_u(\omega)} \qquad \textbf{(17)}$$

where d is the model order, N is the number of experimental data samples used in the estimation, G_0 is the estimated nominal model and Φ_v and Φ_u are respectively the spectral power densities of the measurement noise v and excitation signal u.

Bias error: On the other hand, the bias error characterizes the under-modeling. This is due (1) to the lack of knowledge about the process to be modeled and (2) to the need of using simple models (linear, time invariant, reduced order) for the control design. The bias error can be interpreted as a model too, and its response magnitude and phase vary with frequency. Unlike the variance error, the bias error does strongly depend on the nominal model and the excitation signal used in the identification experiment (thus, the need to adequately design the experiment).

In a first stage, the bias error can be reduced by increasing the model order. If the model structure is richer, the model will be able to better describe the process to be modeled. Nevertheless, increasing the model order also increases the variance error. In

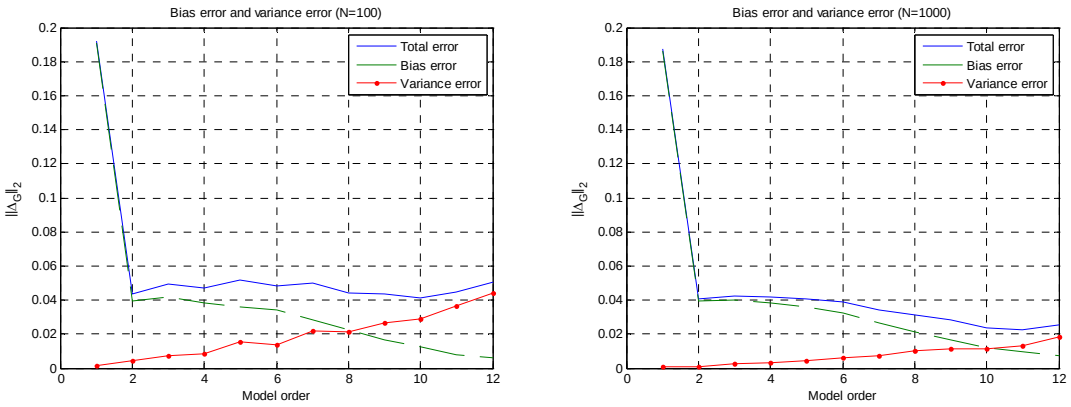
other words, one reaches a point where the SNR (signal to noise ratio) of the data is not large enough to accurately estimate the parameters of a high order model.

This result is known as the bias/variance trade-off and it is a classical result in the field of system identification, see for instance, (Ljung, 1999a), (Ninness and Goodwin, 1995), (Hakvoort and Van den Hof, 1997), (Ninness and Hjalmarsson, 2003a), (Ninness and Hjamarsson, 2003b).

The conclusion is that there exists an optimal model order such as it balances the bias error decrease with the variance error increase and it corresponds to the smaller estimation error. If the data sequence is too short and noisy, the optimal order will be too low (see Example 2.3). As a consequence, the model will be so biased that the confidence intervals corresponding to the variance error will not include the estimation error (this is what happened in the first model in Example 2.2).

Example 2.3. Bias/variance trade-off

Consider again the plant and experiment of Example 2.1 and Example 2.2 (Ninness and Goodwin, 1995). We have obtained the bias and variance errors for different Laguerre models with orders varying between 1 and 12. To quantify the error we have used the Euclidean norm between the true system and the nominal model. For the bias error we have not considered the measurement noise and for the variance error we have computed the average error for 20 different noise realizations. Results are shown in Fig. 2.3.



(a) **(b)**

Fig. 2.3. Bias/variance trade-off. Selection of the order that minimizes the total estimation error

Fig. 2.3 illustrates the effect of model order d in the variance error increase and bias error decrease, but it also shows the effect of the data length N. For a given number of samples N, a large d decreases the bias error but increases the variance error, thus increasing the total error. In Fig. 2.3(a), where the data length is short ($N = 100$) the variance error increases faster than in Fig. 2.3(b), where the data length is larger ($N = 1000$). Thus, a small data length implies a larger total error and a smaller "optimal" order. ■

2.2 Stochastic descriptions for model uncertainty

2.2.1 Model Error Modeling (MEM)

In the previous section, we have seen that the classical solution, although very used in system identification, is not suitable for robust identification. The major drawback is that the PE approach does not consider that residuals are due to both the measurement noise (variance error) and the model structure (bias error). Model Error Modeling (MEM) methods overcome this shortcoming.

> **Remark:** Here we present a stochastic version of MEM methods. However, the MEM concept is also valid in a deterministic framework.

a. Approach

The name MEM (Model Error Modeling) refers to a number of methods that aim to obtain a model G_e of the error between the nominal model $G_0 = G(q, \widehat{\boldsymbol{\theta}}_N)$ and the true system G_{true}. Hence, the experiment of Equation (4) can be expressed as:

$$y_n = G\left(q, \widehat{\boldsymbol{\theta}}_N\right)u_n + G_e(q)u_n + v_n \ , \ n = 1, \dots, N \qquad \textbf{(18)}$$

Here, it is assumed that the component $G_e(q)u_n$ in (18) cannot be well described as a realisation of a stationary stochastic process and that it is too much significant to be neglected. In fact, the "size" of the error model G_e is not negligible in most practical situations, especially those in which the order of the nominal model G_0 must be small (a typical requirement of robust control design techniques).

The uncertainty region for G_0 is then computed on the basis of G_e and its confidence regions. This line of work was initiated by (Ljung, 1997) and some remarkable references are (Garulli and Reinelt, 2000) and (Reinelt *et al*., 2002).

b. Identification of the model error

Once the nominal model $G_0 = G(q, \widehat{\boldsymbol{\theta}}_N)$ has been identified, it is possible to evaluate the size of the unmodeled dynamics by means the residual analysis. The residuals are computed as $\varepsilon_n = y_n - G(q, \widehat{\boldsymbol{\theta}}_N)u_n$ and the error model G_e can be interpreted as a dynamic system where the input is $\{u_n\}_{n=0}^{N-1}$ and the output is $\{\varepsilon_n\}_{n=0}^{N-1}$.

The error model $G_e(q)$ can be identified by any system identification method, for instance by classical identification methods.

An important issue is the selection of the error model structure. It must be flexible enough to reveal bias errors in the nominal model and to detect the frequency regions where the uncertainty is significant, but at the same time its confidence regions must remain small. There exist several choices, e.g., (Ljung, 1999b) uses FIR (Finite Impulse Response) models while (Milanese, 1998) proposes the use of non-parametric models. If the order of the error model is high enough, the remaining error will be basically a variance error and the confidence regions could be computed from the covariance matrix of the parameters.

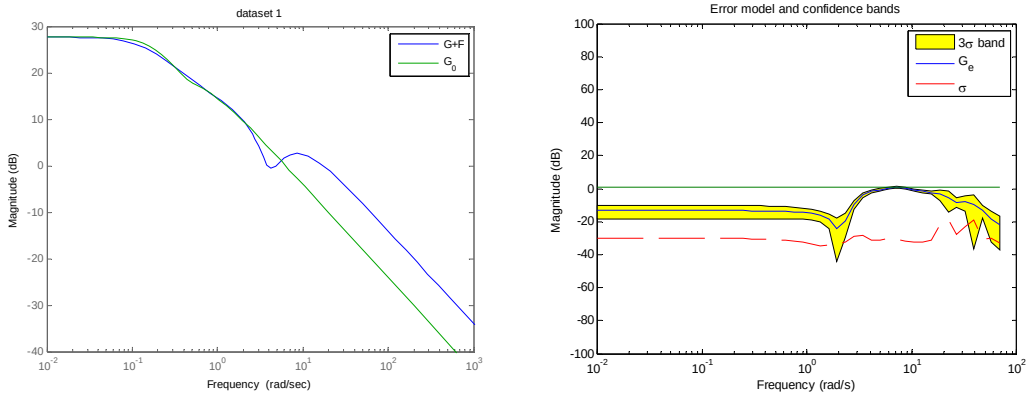
c. Uncertainty bands

Once the error model structure has been selected and G_e has been identified, upper and lower error bounds $G_{e\_up}(\omega)$, $G_{e\_lo}(\omega)$ for the confidence regions of $G_e(\omega)$ are derived. There exist several ways to combine the bounds $G_{e\_up}(\omega)$, $G_{e\_lo}(\omega)$ with $G_0(\omega)$ and $G_e(\omega)$ to obtain the uncertainty band around $G_0(\omega)$. Next example illustrates this point.

Example 2.4. MEM uncertainty regions

We consider the plant and experiments of (Reinelt *et al*., 2002). The nominal model is a continuous time fourth order Laguerre-type with the pole located at $p = -0.2895$ (for the datasets 1 and 2) and at $p = -0.5737$ (for the datasets 3 and 4). Fig. 2.4(a) shows the frequency response magnitude of the nominal model and the linear part of the true model for dataset 1.

The error model is chosen as an Output Error model of the form $\varepsilon_n = \frac{B(q)}{F(q)}u_n + \frac{C(q)}{D(q)}v_n$ where the polynomial orders are $n_b = 20$, $n_c = 10$, $n_d = 10$ and $n_f = 20$. The polynomial coefficients are computed by means the `pem` Matlab™ function. Fig. 2.4(b) shows the magnitude of the error model frequency response along with the upper and lower error bounds $|G_{e\_up}(\omega)|$, $|G_{e\_lo}(\omega)|$. These bounds are obtained by respectively adding and subtracting 3σ to the error model magnitude, where σ is the standard deviation of the model error magnitude.

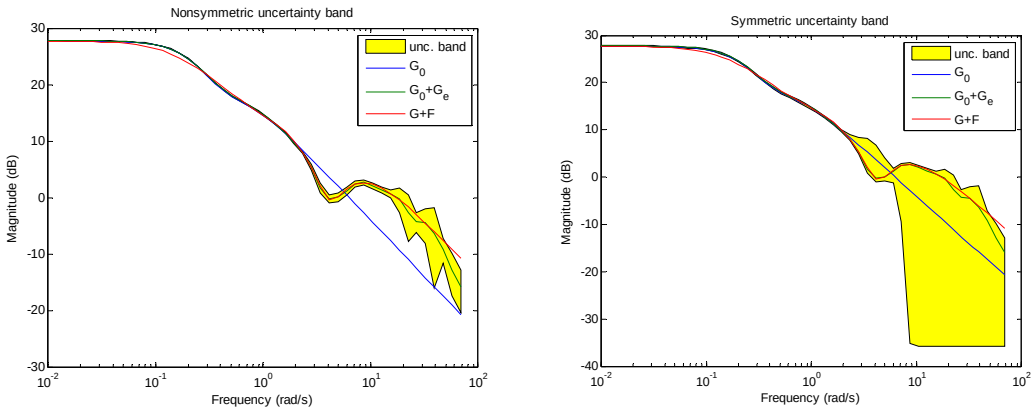


(a) **(b)**

Fig. 2.4. (a) Nominal model and linear part of the true model, (b) error model and confidence region

A direct way to construct the nominal model uncertainty band is to add frequency to frequency the error model to the nominal model, and then simply add and subtract the 3σ confidence regions of the error model. Fig. 2.5(a) shows the resulting non-symmetric region. This solution, although useful for model validation purposes, may lead to the situation where the nominal model is outside its uncertainty band.

An alternative is to construct a symmetric band around the nominal that includes the non-symmetric one. This is shown in Fig. 2.5(b) for the dataset 1.



(a) **(b)**

Fig. 2.5. Uncertainty bands around the nominal model: (a) nonsymmetric, (b) symmetric

2.2.2 Non Stationary Stochastic Embedding (NSSE)

a. Approach

The Non-Stationary Stochastic Embedding (NSSE) method (Goodwin, Braslavsky, and Seron, 2002) can be interpreted as a particular case of MEM, where the error model $G_e = G_0 - G_{true}$ is described as a random variable, in particular, as a realization of a non-stationary stochastic process whose variance grows with frequency.

The simplest selection for such a process is a random walk (also called Brownian motion) in the frequency domain, that is, a zero mean process $\{\lambda(\omega)\}$ of independent, infinitely divisible Gaussian increments,

$$\lambda(\omega) = \int_0^{\omega} d\varepsilon(s) \quad \text{with} \quad E[d\varepsilon(s)d\varepsilon(s)] = \sigma_\varepsilon^2 ds \tag{19}$$

The NSSE method takes a multiplicative description for the uncertainty and thus the frequency response of the true system can be expressed as:

$$G^R(\omega) = G_0^R(\omega, \boldsymbol{\theta}) + G_0^R(\omega, \overline{\boldsymbol{\theta}})\lambda^R(\omega)$$
$$G^I(\omega) = G_0^I(\omega, \boldsymbol{\theta}) + G_0^I(\omega, \overline{\boldsymbol{\theta}})\lambda^I(\omega) \quad \textbf{(20)}$$

where the superscripts R and I refer to the real part and imaginary part respectively, and $\{\lambda^R(\omega)\}$ and $\{\lambda^I(\omega)\}$ are two independent processes with parameter σ_ε^2. In practical situations, the parameter vector of the error model $\overline{\boldsymbol{\theta}}$ can be selected equal to the parameter vector of the nominal model $\boldsymbol{\theta}$.

If we linearly parameterize the nominal model as in (7), i.e., by means of the functions $B_k(q)$, the system frequency response can be expressed as

$$G^R(\omega) = \mathbf{B}^R(\omega)\boldsymbol{\theta} + \mathbf{B}^R(\omega)\overline{\boldsymbol{\theta}}\lambda^R(\omega)$$
$$G^I(\omega) = \mathbf{B}^I(\omega)\boldsymbol{\theta} + \mathbf{B}^I(\omega)\overline{\boldsymbol{\theta}}\lambda^I(\omega) \quad \textbf{(21)}$$

where $\mathbf{B}^R(\omega) = \left(B_1^R(\omega), \dots, B_d^R(\omega)\right)$ and $\mathbf{B}^I(\omega) = \left(B_1^I(\omega), \dots, B_d^I(\omega)\right)$.

As in classical system identification, the uncertainty is described in the Nyquist plane frequency to frequency by ellipses centered at the nominal model and with size and shape determined by the covariance matrix $\boldsymbol{\Sigma}_e(\omega_n) = E[\mathbf{G}_e(\omega_n)\mathbf{G}_e(\omega_n)^T]$, where $\mathbf{G}_e(\omega_n)$ is the resulting total modeling error at frequency ω_n, $\mathbf{G}_e(\omega_n) = (G_e^R(\omega_n) \quad G_e^I(\omega_n))^T$. The difference with classical system identification is that the covariance matrix characterizes both the measurement noise (variance error) and the under-modeling (bias error).

b. Identification of the nominal model

Since the method operates in the frequency domain, the first step is to obtain a point wise estimate $\widehat{\mathbf{G}}$ of the true frequency response. This is then used to compute the nominal model, i.e. to obtain an estimate for $\boldsymbol{\theta}$.

Excitation signal: In order to obtain a frequency to frequency estimate of the true system frequency response, the excitation must be a multi-sinusoid

$$u_n = \sum_{l=1}^{m} A_l \cos(\omega_l n T_s + \varphi_l) \quad \textbf{(22)}$$

consisting of m sinusoids of frequencies $\{\omega_1, \dots, \omega_m\}$, not necessarily uniform spaced. This way, the steady state response samples are given by

$$y_n = y(nT_s) = \sum_{l=1}^{m} A_l g^R(\omega_l)\cos(\omega_l n T_s + \varphi_l) - \sum_{l=1}^{m} A_l g^I(\omega_l)\sin(\omega_l n T_s + \varphi_l) + v_n$$

and the terms $g^R(\omega_l)$ and $g^I(\omega_l)$ can be obtained by correlation methods (they are the coefficients of the trigonometric series):

$$g^R(\omega_l) = \frac{2}{A_l N}\sum_{n=1}^{N} y(nT_s)\cos(\omega_l nT_s + \varphi_l)$$
$$g^I(\omega_l) = -\frac{2}{A_l N}\sum_{n=1}^{N} y(nT_s)\sin(\omega_l nT_s + \varphi_l)$$

True frequency response point estimation: The matrix expression for the point wise estimate of the true system frequency response is then $\widehat{\mathbf{G}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y}$, where $\widehat{\mathbf{G}} = (\hat{g}^R(\omega_1) \quad \hat{g}^I(\omega_1) \quad \dots \quad \hat{g}^R(\omega_m) \quad \hat{g}^I(\omega_m))^T$, $\mathbf{y}$ is the vector containing the time domain steady state response samples and the design matrix $\mathbf{\Phi}$ is

$$\mathbf{\Phi} = \begin{bmatrix} A_1\cos(\omega_1 T_s + \varphi_1) & -A_1\sin(\omega_1 T_s + \varphi_1) & \cdots & -A_m\sin(\omega_m T_s + \varphi_m) \\ A_1\cos(\omega_1 2T_s + \varphi_1) & -A_1\sin(\omega_1 2T_s + \varphi_1) & \cdots & -A_m\sin(\omega_m 2T_s + \varphi_m) \\ \vdots & \vdots & \ddots & \vdots \\ A_1\cos(\omega_1 NT_s + \varphi_1) & -A_1\sin(\omega_1 NT_s + \varphi_1) & \cdots & -A_m\sin(\omega_m NT_s + \varphi_m) \end{bmatrix}$$

The expression of $\mathbf{\Phi}^T\mathbf{\Phi}$ is simplified if the frequencies are selected such that the m sinusoids are orthogonal in any interval of length NT_s. In this case, $\mathbf{\Phi}^T\mathbf{\Phi}$ is a diagonal matrix, $\mathbf{\Phi}^T\mathbf{\Phi} = \frac{N}{2}\mathrm{diag}(A_1^2, A_1^2, \dots, A_m^2, A_m^2)$.

Nominal model estimation: The parameters of the nominal model can be estimated from $\widehat{\mathbf{G}}$ in a least squares sense, $\widehat{\boldsymbol{\theta}} = \mathbf{Q}\widehat{\mathbf{G}}$ where $\mathbf{Q} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ and $\mathbf{B} = \left(\mathbf{B}^R(\omega_1), \mathbf{B}^I(\omega_1), \dots, \mathbf{B}^R(\omega_m), \mathbf{B}^I(\omega_m)\right)^T$.

c. Quantification of the total modeling error

The resulting total modeling error at any frequency ω_n is $\mathbf{G}_e(\omega_n) = \mathbf{B}(\omega_n)\widehat{\boldsymbol{\theta}} - \mathbf{G}(\omega_n)$, where $\mathbf{G}_e(\omega_n) = (G_e^R(\omega_n) \quad G_e^I(\omega_n))^T$, $\mathbf{B}(\omega_n) = (\mathbf{B}^R(\omega_n) \quad \mathbf{B}^I(\omega_n))^T$, and $\mathbf{G}(\omega_n) = (G_{true}^R(\omega_n) \quad G_{true}^I(\omega_n))^T$.

And the covariance matrix $\mathbf{\Sigma}_e(\omega_n) = E[\mathbf{G}_e(\omega_n)\mathbf{G}_e(\omega_n)^T]$ presents two terms, one due to the measurement noise and the other due to the random process that characterizes the under-modeling,

$$\mathbf{\Sigma}_e(\omega_n) = \mathbf{K}_v(\omega_n)\sigma_v^2 + \mathbf{K}_\varepsilon(\omega_n)\sigma_\varepsilon^2 \qquad \textbf{(23)}$$

where

$$\mathbf{K}_v(\omega_n) = \mathbf{B}(\omega_n)\mathbf{Q}\mathbf{A}\mathbf{Q}^T\mathbf{B}^T(\omega_n)$$
$$\mathbf{K}_\varepsilon(\omega_n) = \mathbf{B}(\omega_n)\mathbf{Q}\mathbf{\Omega}\mathbf{Q}^T\mathbf{B}^T(\omega_n) + \left(\mathrm{diag}(\mathbf{B}(\omega_n)\bar{\boldsymbol{\theta}})\right)^2\omega_n - \mathbf{\Psi}(\omega_n) - \mathbf{\Psi}^T(\omega_n)$$

and $\mathbf{A} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$, $\mathbf{\Omega} = \text{diag}(\mathbf{B}\bar{\boldsymbol{\theta}})\left(\begin{bmatrix}\omega_1 & \omega_1 & \cdots & \omega_1\\ \omega_1 & \omega_2 & \cdots & \omega_2\\ \vdots & \vdots & \ddots & \vdots\\ \omega_1 & \omega_2 & \cdots & \omega_m\end{bmatrix}\otimes\mathbf{I}^{2\times 2}\right)\text{diag}(\mathbf{B}\bar{\boldsymbol{\theta}})$,

$\mathbf{\Psi}(\omega_n) = \mathbf{B}(\omega_n)\mathbf{Q}\begin{bmatrix}\text{diag}(\mathbf{B}(\omega_1)\bar{\boldsymbol{\theta}})\omega_1\\ \vdots\\ \text{diag}(\mathbf{B}(\omega_{k-1})\bar{\boldsymbol{\theta}})\omega_{k-1}\\ \text{diag}(\mathbf{B}(\omega_k)\bar{\boldsymbol{\theta}})\omega_n\\ \vdots\\ \text{diag}(\mathbf{B}(\omega_m)\bar{\boldsymbol{\theta}})\omega_n\end{bmatrix}\text{diag}(\mathbf{B}(\omega_n)\bar{\boldsymbol{\theta}})$ and we assume that ω_n is such that $\omega_{k-1} \le \omega_n < \omega_k$, where ω_{k-1} and ω_k are two consecutive frequencies of the excitation signal.

Finally, Equation (23) needs the variance values σ_v^2 and σ_ε^2. These can be estimated from the experimental data. Firstly, in the frequency domain, an unbiased estimate for the measurement noise variance σ_v^2 can be computed as (Goodwin and Payne, 1977)

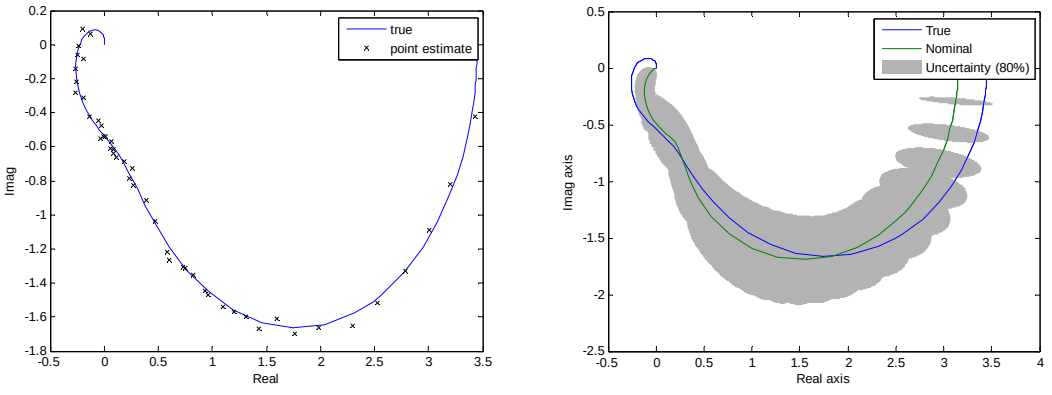
$$\hat{\sigma}_v^2 = \frac{1}{N-2m}\left(\mathbf{y}-\mathbf{\Phi}\hat{\mathbf{G}}\right)^T\left(\mathbf{y}-\mathbf{\Phi}\hat{\mathbf{G}}\right) \tag{24}$$

And secondly in (Goodwin, Braslavsky, and Seron, 2002) an unbiased estimate for the under-modeling random walk variance σ_ε^2 is derived,

$$\hat{\sigma}_\varepsilon^2 = \frac{\left(\hat{\mathbf{G}}-\mathbf{B}\hat{\boldsymbol{\theta}}\right)^T\left(\hat{\mathbf{G}}-\mathbf{B}\hat{\boldsymbol{\theta}}\right)}{\text{trace}[(\mathbf{I}-\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T)\mathbf{\Omega}]} - \frac{\text{trace}[(\mathbf{I}-\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T)\mathbf{A}]}{\text{trace}[(\mathbf{I}-\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T)\mathbf{\Omega}]}\hat{\sigma}_v^2 \tag{25}$$

Example 2.5. NSSE uncertainty regions

Fig. 2.6 illustrates the results for the first example of (Goodwin, Braslavsky, and Seron, 2002). Fig. 2.6(a) shows the true system frequency response and its least squares point estimation for the frequencies of the excitation signal.



(a) **(b)**

Fig. 2.6. NSSE. (a) Point estimate of the true frequency response, (b) Nominal model and uncertainty band

Fig. 2.6(b) shows the identified nominal system along with the uncertainty ellipses computed by (23). The two functions used to parameterize the nominal model are $B_1(s) = \frac{1}{(0.5s+1)^2}$ and $B_2(s) = \frac{1}{(3s+1)^2}$, the resulting optimal parameter vector is $\hat{\boldsymbol{\theta}} = (0.7798 \quad 2.3715)^T$, and the unbiased estimates for the measurement noise and random walk are, respectively, $\hat{\sigma}_v^2 = 0.9807$ and $\hat{\sigma}_\varepsilon^2 = 0.0647$.

■

d. Case of resonant systems

In the case of plants with lightly damped modes at high frequencies, the method may be too conservative at low frequencies. For this reason, in the case of resonant systems, an integrated random walk can be used

$$\mu(\omega) = \int_0^\omega \lambda(s)ds \quad \text{with} \quad E[\mu(\omega_l)\mu(\omega_n)] = \omega_l^2\left(\frac{\omega_n}{2} - \frac{\omega_l}{6}\right)\sigma_\varepsilon^2 \quad \text{for} \quad \omega_l \le \omega_n \tag{26}$$

Equations (23) to (25) are still valid but with the factors $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}(\omega_n)$ slightly modified. See (Goodwin, Braslavsky, and Seron, 2002) for details.

Example 2.6. NSSE uncertainty regions for the case of resonant poles

Fig. 2.7 illustrates the results for the second example in (Goodwin, Braslavsky, and Seron, 2002). The plant presents resonant poles at $-0.5 \pm j05$. Fig. 2.7(a) shows the uncertainty band obtained using the simple random walk of (19) while the uncertainty band in Fig. 2.7(b) has been obtained by using the integrated random walk of (26). Clearly, the conservativeness degree in the second case is lower.

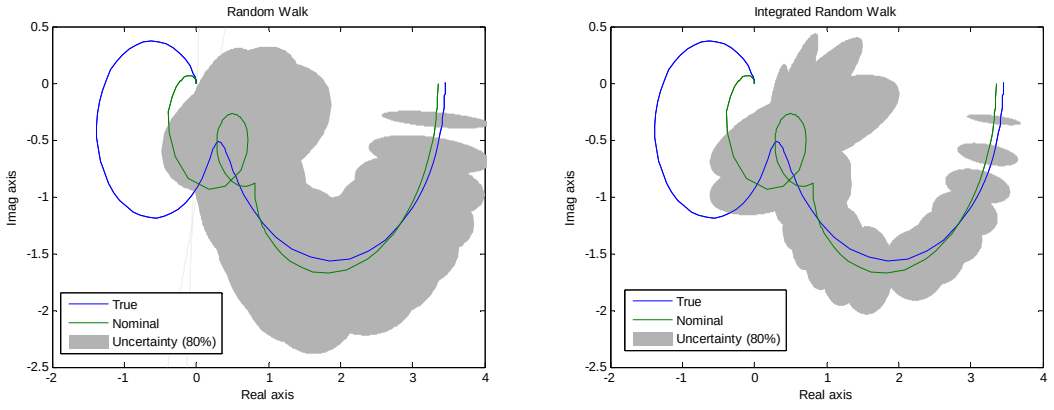


Fig. 2.7. NSSE. (a) Random walk vs. (b) integrated random walk for the case of resonant poles

■

2.3 Worst case Robust Identification methods

In the stochastic methods of Section 2.1, the obtained uncertainty regions are probabilistic since the measurement noise v_n is modeled as a stochastic process, characterized by means a probability distribution.

In the deterministic formulations, v_n is assumed to be *unknown but bounded* (UBB), that is, it satisfies hard constraints, e.g. $|v_n| \leq \delta_n$, $n = 0, \dots, N-1$, where the bounds δ_n are known. The deterministic noise description leads to hard bounded uncertainty regions. For this reason, they are also known as *bounding approaches* (Milanese et al., 1996).

2.3.1 Set-membership viewpoint on system identification

Deterministic methods rely on the *Set-membership Identification* (SMI) concept. This approach consists of characterizing the *model family* where both the nominal model and the hypothetical "true model" are assumed to belong. The *size* of such model set gives us a quantitative idea of the uncertainty around the nominal model. And the *bound* of the model set is a hard-type bound. See e.g. the book of (Milanese et al., 1996) and the references therein.

Originally, set-membership formulation takes many ideas and terminology from the *Information-Based Complexity* (IBC) (Traub et al., 1988). IBC is a branch of theoretical computing science which considers solving problems based on *partial* and *corrupted* information, hence the connection with the robust identification problem. A brief summary about IBC can be found in (Chen and Gu, 2000; App. A).

a. Feasible Parameter Set

In the parameter space, the model family is characterized by the so-called *Feasible Parameter Set* (FPS). To illustrate the construction of the FPS in the linear case consider again the output error model of previous sections

$$y_n = G(q, \boldsymbol{\theta})u_n + v_n \quad , \ |v_n| \leq \delta_n \quad , \ n = 0, \dots, N-1 \tag{27}$$

Under the premise that the system G can be truthfully represented by the output error model (27), the vector $\boldsymbol{\theta}$ has to be *consistent* with measurement data, that is, it must satisfy the inequalities

$$|y_n - G(q, \boldsymbol{\theta})u_n| \leq \delta_n \ , \ n = 0, \dots, N-1 \tag{28}$$

This is equivalent to say that any feasible parameter vector $\boldsymbol{\theta}$ belongs to the feasible parameter set, $\boldsymbol{\theta} \in$ FPS,

$$\text{FPS} \equiv \bigcap_{n=0}^{N-1} \{\boldsymbol{\theta}\colon\ |y_n - G(q,\boldsymbol{\theta})u_n| \le \delta_n\} \tag{29}$$

To estimate the parameter vector $\boldsymbol{\theta}$, and further quantify the estimation error, it suffices to characterize the FPS. To see that this FPS defines a region in the parameter space, consider that the model $G(q,\boldsymbol{\theta})$ is a map from $\mathbb{R}^d$ to $\mathbb{R}^N$,

$$\begin{aligned} G(q,\boldsymbol{\theta})\colon\ \mathbb{R}^d &\longmapsto \mathbb{R}^N \\ \boldsymbol{\theta} &\longmapsto [y_0 \pm \delta_0 \quad \ldots \quad y_{N-1} \pm \delta_{N-1}] \end{aligned} \tag{30}$$

then the FPS can be viewed as a pre-image of the map $G(q,\boldsymbol{\theta})$.

When the selected model structure satisfies the linear regression, $G(q,\boldsymbol{\theta})u_n = \boldsymbol{\varphi}_n^T\boldsymbol{\theta}$, the FPS is a *convex polytope*. The procedure for finding the exact FPS is quite simplified since $|y_n - \boldsymbol{\varphi}_n^T\boldsymbol{\theta}| \le \delta_n$ defines the region between the two parallel lines in the parameter space, orthogonal to $\boldsymbol{\varphi}_n$ and separated $2\delta_n$, $y_n - \boldsymbol{\varphi}_n^T\boldsymbol{\theta} = \pm\delta_n$. The intersection of the strips defined by all N pairs of parallel lines form a polytope in the parameter space and can be obtained exactly by recursive methods (Mo and Norton, 1990). For general nonlinear models one must use Monte Carlo techniques (Ninness and Goodwin, 1995).

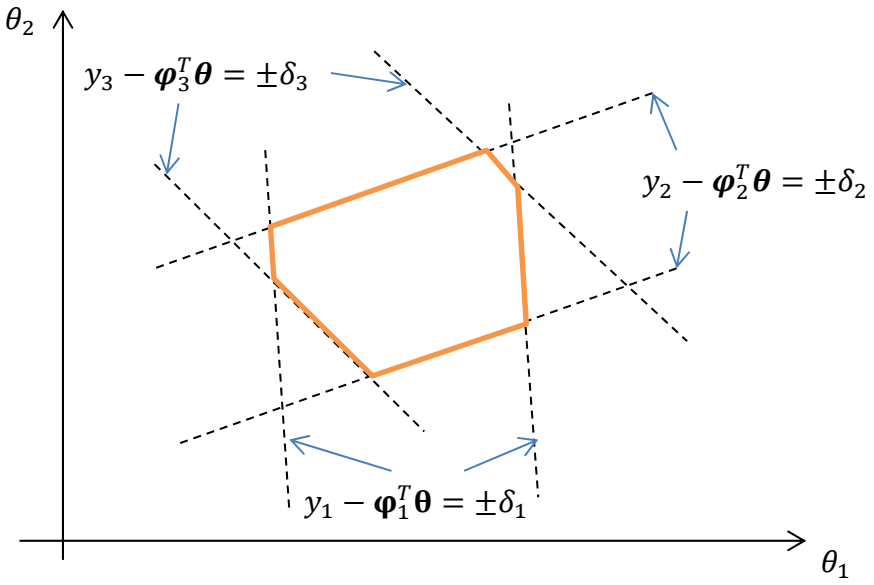


Fig. 2.8. Membership set as a convex polytope (linear regression model)

Example 2.7. Feasible Parameter Set

Fig. 2.9 shows the FPS regions obtained for several values of δ ranging from 0.2 to 0.8 for the same plant and experiment of (Ninness and Goodwin, 1995). The least squares nominal model is the one obtained in the Example 2.1, $\boldsymbol{\theta}_{LSE} = (0.1060 \quad 0.1673)^T$.

Since this is a simulation example, we know that $\|\mathbf{v}\|_\infty = 0.2477$ and that the total (bias plus variance) error bound is $\|\mathbf{y} - \mathbf{\Phi\theta}_{LSE}\|_\infty = 0.2783$. Another bound for this error can also be obtained by the triangle inequality

$$\|\mathbf{y} - \mathbf{\Phi\theta}_{LSE}\|_\infty < \|\mathbf{y}_{noiseless} - \mathbf{\Phi\theta}_{LSE}\|_\infty + \|\mathbf{v}\|_\infty$$

which leads to the result $\|\mathbf{y} - \mathbf{\Phi\theta}_{LSE}\|_\infty < 0.0998 + 0.2477 = 0.3475$. In summary, a tightened selection for δ is 0.3 so we know that $\delta > 0.4$ is too pessimistic and $\delta = 0.2$ is too optimistic.

Note in Fig. 2.9 that for values $\delta > 0.4$ the regions present the same form and their size depends on the δ value. Larger δ values lead to larger uncertainty regions. Moreover all the regions include the optimal value $\mathbf{\theta}_{LSE}$ and their centroids are near it.

On the other hand, the too much optimistic selection $\delta = 0.2$ leads to a small region but it does not include the optimal parameter vector $\mathbf{\theta}_{LSE}$. If we had to select a nominal model from the $\delta = 0.2$ region, a direct solution would be to take its centroid $\mathbf{\theta}_c = (0.1194 \quad 0.1742)^T$. But if we evaluate the total error for this latter model, we find that it is $\|\mathbf{y} - \mathbf{\Phi\theta}_c\|_\infty = 0.3117 > 0.2$, thus the region is not valid.

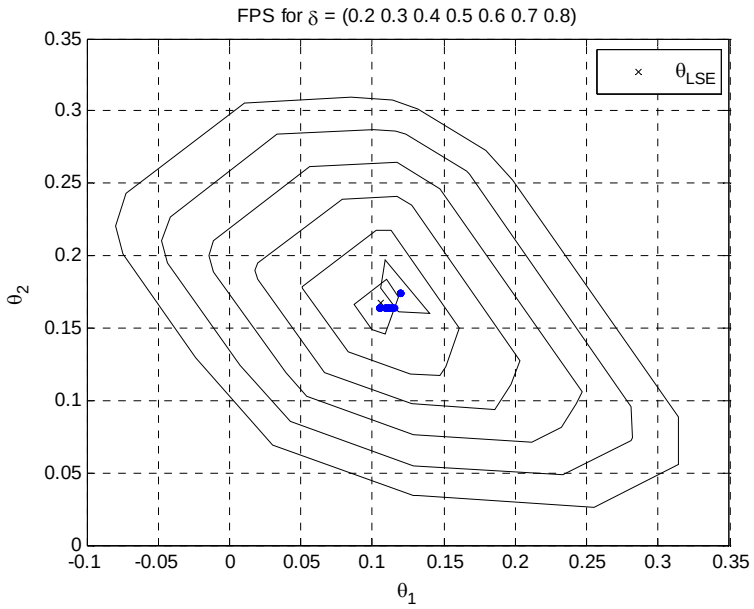


Fig. 2.9. Feasible parameter set for several values of $\boldsymbol{\delta}$ (blue points indicate the centroid of each region)

In summary, the selection of the δ value is critical. For too optimistic selections of δ (small values), the FPS regions may be constructed far from the "true" parameter value and thus it will be erroneous. For pessimistic selections of δ (large values) FPSs would be large and they will contain models that will never occur. This fact increases the conservativeness of the control design and hence it may penalize the control system performance. And in the fault detection framework, it may lead to the lack of detection of faults if they occur inside these large uncertainty regions.

■

b. Overbounding techniques

Most times the FPS shape is too complicated to work with. Also its complexity grows exponentially with the number of data *N*. This is a problem especially when we want to translate the uncertainty in the parameter space to uncertainty in the frequency domain since the straightforward solution consists of mapping the FPS onto the complex plane for each frequency of interest. This computation may be a prohibitive task depending on the FPS complexity.

Hence, it is usual to look for set approximations or *overbounding* regions of a simpler shape (Mbarek et al., 2003), such as ellipsoids (Fogel and Huang, 1982), orthotopes (Meassaoud and Favier, 1994), parallelotopes (Chisci et al., 1998), or limited complexity polytopes (Maraoui and Messaoud, 2001).

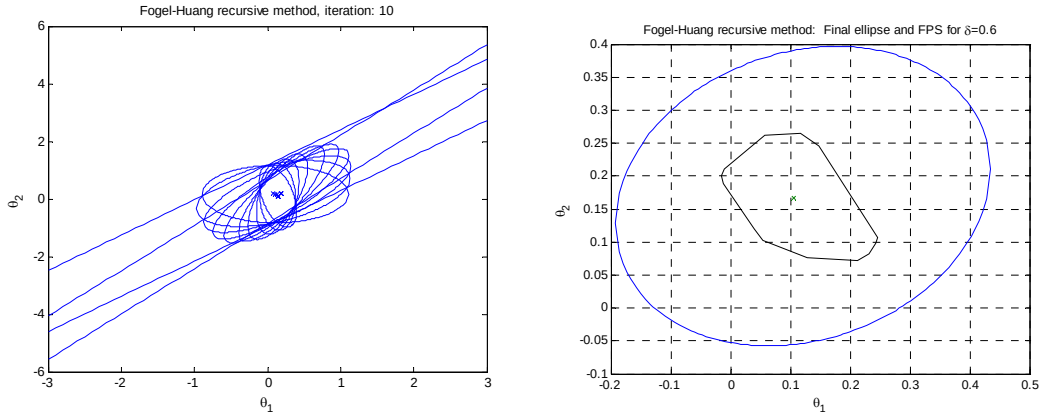
Example 2.8. Fogel Huang overbounding regions

The Fogel-Huang algorithm is an iterative method that finds the smallest overbounding ellipsoid around an arbitrary-shaped uncertainty region.

To derive the ellipsoid region, the algorithm uses the fact that any $\boldsymbol{\theta}$ consistent with the linear-in-the-parameters model and the disturbance assumption $|v_n| \leq \delta_n$ will satisfy the condition

$$\sum_{n=0}^{N-1} \frac{\rho_n}{\delta_n} (y_n - \boldsymbol{\varphi}_n^T \boldsymbol{\theta})^2 \leq \sum_{n=0}^{N-1} \rho_n \quad , \quad \forall\, \rho_n > 0$$

which corresponds to an ellipsoidal region.
It is clear that some of the $\boldsymbol{\theta}$ that satisfy the above condition will be out the FPS defined as in (29). To minimize this phenomenon, (Fogel and Huang, 1992) proposed an appropriate, recursive selection of the positive definite weightings ρ_n.



(a) **(b)**

Fig. 2.10. Fogel-Huang algorithm

Fig. 2.10(a) shows the iterative overbounding process for the same plant and experiment of Example 2.7 (Ninness and Goodwin, 1995). The FPS to be overbounded is the one

corresponding to the case $\delta = 0.6$. Fig. 2.10(b) shows the final ellipse obtained at the 35th iteration. The FPS is effectively inside this final ellipsoid.

■

c. Nominal model selection

In the worst case setting, feasible region bounds are *hard* bounds, that is, every parameter outside such a region is not consistent with actual data and should be discarded. Moreover, all models inside the hard-bounded feasible region are equally probable to occur. In this context a practical selection for the nominal model is the center of the FPS,

Projection estimate: Define the worst-case ℓ_p error as $e(\widehat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta} \in FPS} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_p$. Then, the central optimal estimate is given by $\widehat{\boldsymbol{\theta}}_c = \operatorname{argmin}_{\widehat{\boldsymbol{\theta}}} e[\widehat{\boldsymbol{\theta}}]$. However, for the case $\boldsymbol{\theta} \in \mathbb{R}^d$, $d > 1$, it cannot be guaranteed that the estimate is consistent with the FPS (Akçay, Hjalmarsson, and Ljung, 1996).

A suboptimal choice is obtained by minimizing the worst-case prediction error

$$\widehat{\boldsymbol{\theta}}_{PJ} = \arg\min_{\boldsymbol{\theta}} \sup_{n=1..N} |\varepsilon_n(\boldsymbol{\theta})| \tag{31}$$

This corresponds to a prediction error method in the ℓ_∞ norm, or equivalently, to finding the minimum δ for which the resulting FPS is nonempty. This estimate is usually addressed as *Projection Estimate* in the set-membership literature (Milanese and Vicino, 1991) or *Chebyshev Estimate* in a statistical context (Akçay, Hjalmarsson, and Ljung, 1996), and it does enjoy the useful property of being always feasible, that is, $\widehat{\boldsymbol{\theta}}_{PJ} \in$ FPS.

Restricted projection estimate: The optimization problem proposed in (31) is a too complicated min-max problem to deal with. Fortunately, if the noise is ℓ_∞-norm bounded, the FPS will form a polytope in $\mathbb{R}^d$ and it is possible to use an approximation known as the *Restricted Projection Estimate* (Garulli, Vicino, and Zappa, 2000), which involves estimating the suboptimal parameter vector, with limited computational effort, by linear programming

$$\widehat{\boldsymbol{\theta}}^r = \arg\inf_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\| \mathbf{y} - \sum_{i=1}^{d} B_i \theta_i \mathbf{u} \right\|_\infty \tag{32}$$

This approximation enjoys some nice properties. It does not depend on the actual value of the noise bound δ. And it also equals the maximum likelihood estimate when assuming that the innovations ε present symmetric, uniform distribution, with unknown bound.

Linear programming solution: The restricted projection estimate (32) can be solved by linear programming. The aim is to find $\boldsymbol{\theta}$ such that c in $|\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta}| \leq c$ is minimal. An alternative expression is $-c \leq \mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta} \leq c$. And this, in turn, is equivalent to the following two equations, $\boldsymbol{\Phi}\boldsymbol{\theta}-c \leq \mathbf{y}$, $-\boldsymbol{\Phi}\boldsymbol{\theta}-c \leq -\mathbf{y}$. In matrix notation,

$$\begin{bmatrix} \boldsymbol{\Phi} & -\mathbf{1} \\ -\boldsymbol{\Phi} & -\mathbf{1} \end{bmatrix} \begin{pmatrix} \boldsymbol{\theta} \\ c \end{pmatrix} \leq \begin{pmatrix} \mathbf{y} \\ -\mathbf{y} \end{pmatrix}$$

where $\mathbf{1}$ is a column vector of 1's.

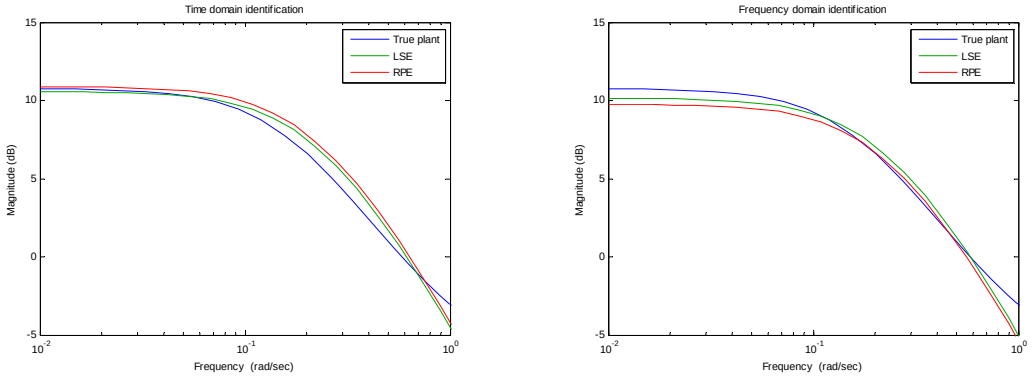
The problem can be solved either in time and frequency domain, provided the particular definition of $\boldsymbol{\Phi}$ and $\mathbf{y}$ (see previous sections). In both cases, we can use the Matlab function `x=linprog(f,A,b)` which solves the following optimization problem

$$\min_{\mathbf{x}} \mathbf{f}^T\mathbf{x} \quad \text{subject to} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}$$

Thus we can solve our problem by defining $\mathbf{f}^T = (\mathbf{0}_{\mathbf{1}\times \boldsymbol{d}} \quad 1)$ where $\mathbf{0}_{\mathbf{1}\times \boldsymbol{d}}$ is a d component row vector of 0's, $\mathbf{x} = \begin{pmatrix} \boldsymbol{\theta} \\ c \end{pmatrix}$, $\mathbf{A} = \begin{bmatrix} \boldsymbol{\Phi} & -\mathbf{1} \\ -\boldsymbol{\Phi} & -\mathbf{1} \end{bmatrix}$, and $\mathbf{b} = \begin{pmatrix} \mathbf{y} \\ -\mathbf{y} \end{pmatrix}$.

Example 2.9. Nominal model by restricted projection estimate

The restricted projection estimate has been computed for the plant of (Goodwin *et al.*, 2002) and the obtained nominal model is shown in Fig. 2.11. The results are compared to the true plant and to the model obtained by least squares estimation. Also, results for time domain data (Fig. 2.11(a)) and frequency domain data (Fig. 2.11(b)) are compared.



(a) **(b)**

Fig. 2.11. Restricted projection estimate, for (a) time domain data and (b) frequency domain data

■

2.3.2 Worst case system identification in $\mathcal{H}_\infty$

a. Feasible Model Set

The *Feasible Model Set* (FMS) can be viewed as an extension of the FPS and it is more general since it contains models, not only parameter values. The name FMS is not standard, in particular it is addressed as "feasible systems set" (Milanese and Taragna, 2005), as "unfalsified systems set" (Hjalmarsson, 2005), and even as "consistency set" (Mazzaro, Parrilo, and Sánchez Peña, 2004).

To construct the FMS one has to consider both *a priori* information and *a posteriori* information. The *a priori* information consists of the assumptions on the system dynamics (*model class* $\mathcal{G}$) and the assumptions on the measurement noise (*noise class* $\mathcal{N}$). Model class and noise class are combined to build the so-called Candidate Model Set (CMS). The *a posteriori* information consists of the measurement data $\{y_n, u_n\}_{n=0}^{N-1}$ obtained by means one experiment over the system. The combination of the CMS and the measurements leads to the FMS,

$$\text{FMS} = \{G \in \mathcal{G}:\ |y_n - G(q,\boldsymbol{\theta})u_n| \in \mathcal{N},\quad n = 0, \dots, N-1\} \tag{33}$$

Let us illustrate the construction of the FMS. For instance, a typical assumption on $\mathcal{G}$ is that the system is "exponentially stable", that is, the impulse response satisfies the following restriction: $|g_n| \le K\rho^{-n}$, $n = 0, 1, 2, ...$, with $K > 0$ and $\rho > 1$. The model class can be expressed as $\mathcal{G} = \{G \in \mathcal{H}_\infty(\mathcal{D}):\ |g_n| \le K\rho^{-n},\ K > 0,\ \rho > 1, \forall n \ge 0\}$, where $\mathcal{H}_\infty(\mathcal{D})$ is the space of all functions F *analytic* in the open unit disk $\mathcal{D} = \{z \in \mathbb{C}:\ |z| < 1\}$ and bounded in the $\mathcal{H}_\infty$ norm $\|F\|_\infty \equiv \sup_{z\in\mathcal{D}}|F(z)| < \infty$.

> **Remark**: In robust identification it is usual to define the z-transform in terms of z^k (instead of z^{-k}). Therefore causal stable systems $G(z)$ are analytic *inside* the unit circle. This is useful because there exist many identification and interpolation techniques on functions analytic in the unit disk. Some authors call it the λ-transform (Chen and Gu, 2000).

Regarding the disturbance class $\mathcal{N}$, additive measurement noise is usually assumed to be bounded in magnitude, i.e., $|v_n| \le \delta$, $n = 0, \dots, N-1$. Thus, $\mathcal{N} = \{\mathbf{v} \in \mathbb{R}^N:\ \|\mathbf{v}\|_\infty \le \delta\}$. Here we have considered ℓ_∞-bounded noise but other bounding criteria can be used. Of course, much more complex choices for $\mathcal{G}$ and $\mathcal{N}$ are possible. The combination of $\mathcal{G}$ and $\mathcal{N}$ with the actual measurements $\mathbf{y} = (y_0, \dots, y_{N-1})^T$, $\mathbf{u} = (u_0, \dots, u_{N-1})^T$ leads to $\text{FMS} = \{G \in \mathcal{G}:\ \|\mathbf{y} - G(q,\boldsymbol{\theta})\mathbf{u}\|_\infty \le \delta\}$.

b. Formulation of the Robust Identification problem in $\mathcal{H}_\infty$

Robust Identification first appeared to be used in Robust Control techniques, especially in $\mathcal{H}_\infty$ synthesis methods. See a survey in (Chen an Gu, 2000). In this context, the formulation of the robust identification problem in $\mathcal{H}_\infty$ is the following:

Given:

(i) The plant *a priori* information in the form of a set $\mathcal{G}$ such that $G_{true} \in \mathcal{G} \subset \mathcal{H}_\infty$,

(ii) the noise *a priori* information via a constant $\delta > 0$ such that $v \in \mathcal{N}(\delta) \subset \ell_\infty$, and

(iii) the experimental *a posteriori* information obtained via the experiment operator η_N: $\mathcal{H}_\infty \times \ell_\infty \mapsto \ell_\infty$ defined by $[\eta_N(G_{true}, v)]_n$, $n = 0..N-1$, where N is the number of samples.

Find:

An identification algorithm A_N such that it maps the *a priori* information and the *a posteriori* information data to an identified nominal model, $G_N \equiv A_N[\eta_N(G_{true}, v)]$, and that the worst case identification error

$$e(A_N, \delta, \mathcal{G}) \equiv \sup_{\substack{G \in \mathcal{G} \\ v \in \mathcal{N}(\delta)}} \|G_{true} - A_N[\eta_N(G_{true}, v)]\|_\infty \qquad \textbf{(34)}$$

converges in the sense that $\lim_{\substack{N\to\infty \\ \delta\to 0}} e(A_N, \delta, \mathcal{G}) = 0$.

In addition, derive explicit bounds on $e(A_N, \delta, \mathcal{G})$. □

The original formulation has been extended by many authors in order to include time domain data, frequency domain data and mixed time domain/frequency domain data, and in order to produce models with a parametric part and a nonparametric part. See (Mazzaro, Parrilo, and Sánchez Peña, 2004) for a benchmark example with these extensions.

c. Types of algorithms

The problem formulation of the previous section leads to different types of identification algorithms. An identification algorithm is just a rule that delivers a *nominal model* on the basis of the available information (FMS) and particular specifications for the nominal model (structure, order…). We speak of *conditional* algorithms when restrictions on the nominal model are posed. Also, an identification algorithm is said to be *linear* if it is a linear function of the *a posteriori* data, otherwise it is said to be *non-linear*. It is said to be *untuned* if it does not depend on *a priori* information about the plant and the measurement noise; otherwise it is said to be *tuned*.

Linear algorithms: Linear algorithms operate *linearly* on experimental data. They are simple and require low computational effort.

However their usefulness in robust identification is limited. Untuned linear algorithms are developed using *polynomial* approximation techniques, and are shown to be divergent in the worst-case. Tuned linear algorithms, on the other hand, are convergent but not robustly convergent, and are constructed based on *least squares* optimization. But this diverges on the worst-case.

Two stage nonlinear algorithms: Non-linear algorithms have been developed to overcome the robust convergence limitations of linear algorithms. In the case of frequency domain measurements, a basic technique consists of two steps:

1) Find a trigonometric polynomial T (i.e. a polynomial in z and $1/z$) that models the data closely.

2) Given T, find the rational function in the disc algebra F that minimizes $\|T - F\|_\infty$ over the unit circle. This second stage involves solving the Nehari's problem. See (Chen and Gu, 2000) for details.

Two stage non-linear algorithms present better properties than linear algorithms. However, they produce approximate models of excessive order and there is no guarantee for the identified model to belong to FMS.

Interpolatory algorithms: Interpolatory algorithms always yield nominal models belonging to FMS, see e.g. (Milanese and Taragna, 2002) and (Parrilo *et al.*, 1999). In general, a two-step procedure is carried out:

1) Validation of the FMS.

2) Identification of a model belonging to FMS by means of nonlinear interpolation techniques.

(Parrilo *et al.*, 1999) propose an interpolatory algorithm with direct application to $\mathcal{H}_\infty$ robust control design since the resulting model is presented in terms of a Linear Fractional Transformation (LFT) parameterized by a free function Q. Thus, this method is very close to the $\mathcal{H}_\infty$ robust control issues.

For another example of interpolatory algorithm, consider the nearly optimal algorithm of (Milanese and Taragna, 2002). The algorithm relies on the value set approximation (the value set is defined as the mapping of the FMS to the Nyquist plane). The procedure is as follows:

For any frequency, the first step is to compute the inner and outer approximations $\underline{VI}_m^d(\omega)$ and $\overline{VO}_m^d(\omega)$ of the value set $V(\omega)$ and to compute their centers. Then, on the basis of the value sets centers, a nearly optimal (usually FIR) model G_d of order d is obtained.

The identification error is minimized in two steps: (1) For given $n < d$, compute a reduced model $G_n^r(\widehat{\boldsymbol{\theta}})$ by Hankel norm approximation methods, and (2) using $G_n^r(\widehat{\boldsymbol{\theta}})$ as starting point, perform the nonlinear optimization $\boldsymbol{\theta}^{opt} = \arg\min_{\boldsymbol{\theta}} e(\mathcal{G}_n(\boldsymbol{\theta}))$ in order to obtain the model that minimizes the identification error.

Finally, the order is selected by evaluating the optimality level of $\mathcal{G}_n(\boldsymbol{\theta}^{opt})$ and choosing the order n by trading off between model set complexity and achievable optimality level. The final identified model set is given by $\mathcal{G}_n(\boldsymbol{\theta}^{opt}) = \{G(\boldsymbol{\theta}^{opt}) + \Delta: \ \|\Delta\|_\infty \leq e(\mathcal{G}(\boldsymbol{\theta}^{opt}))\}$.

Example 2.10. Value set computation

Fig. 2.12 shows the key step of the procedure described above which consists of the value set approximation. Results are given for the same example of (Milanese and Taragna, 2002). Fig. 2.12(a) gives the detail for the value set corresponding to the point $\omega_k = k\pi/1024,\ k = 5$, while Fig. 2.12(b) shows the true frequency response along with the value sets for $k = 208, 224, 240$. All polytopes have been computed with 16 vertices and assuming model order 130.

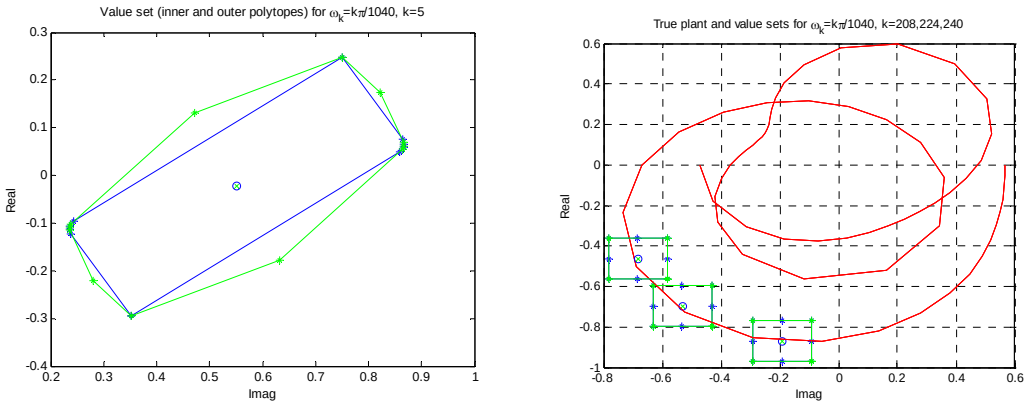


Fig. 2.12. Set value approximation in the nearly optimal algorithm: (a) Value set for k=5, (b) Value sets for k=208, 224, 240 and true frequency response

This method is intensive computationally since the computation of each polytope involves as much optimization steps (via the Matlab^R `linprog` function) as number of vertices.

■

2.4 Summary and conclusion

We have summarized the main features of classical system identification and robust identification. The major drawback of classical system identification is that it only characterizes properly the model uncertainty due to the variance error. Robust identification methods, both stochastic and deterministic, overcome this problem by explicitly assuming that the model uncertainty is due to the variance error (measurement noise and data length) and to the bias error (under-modeling). Stochastic methods use a probabilistic description of the errors and thus lead to probabilistic uncertainty regions. Deterministic methods rely on the concept of unknown but bounded errors and thus lead to hard bounded uncertainty regions.

CHAPTER 3

Bayesian Approach to Robust Identification

In this chapter, we define and characterize the *Bayesian Credible Model Set* (BCMS). The BCMS serves as a basis for the formulation of the Bayesian Robust Identification problem. The construction of the BCMS in the parametric case and in the frequency domain is illustrated. It is explained how to obtain the credible regions that constitute the uncertainty modeling in the Bayesian framework, and connections to the existing deterministic and stochastic robust identification methods are shown.

3.1 Bayesian Credible Model Set

One of the key ideas of the present thesis is to define a probabilistic model set containing all candidate models (*a priori* information) consistent with measurement data (*a posteriori* information). We call such a set the *Bayesian Credible Model Set* (BCMS) and we define it in terms of model posterior probability distributions.

3.1.1 Definition and main features

Definition 3.1. The Bayesian Credible Model Set (BCMS) is the set

$$\mathcal{B} \equiv \{G \in \mathcal{G}:\ p(G|\mathbf{y}) \geq c(\alpha)\} \tag{35}$$

that contains all the models G belonging to a model space $\mathcal{G}$ whose posterior probability distribution conditioned to measurement data, $p(G|\mathbf{y})$, is higher than a given critical value $c(\alpha)$ where $100(1-\alpha)\%$ is the desired credibility level. □

About the term "credible": The set $\mathcal{B}$ is inspired in the Feasible Model Set (FMS) of deterministic methods (see Chapter 2). "Feasible" is a term from the information based complexity theory, which is the origin of worst case set-membership identification methods. Since the underlying theory in the present approach is the Bayesian estimation theory, we rather use the term "credible".

Prior and posterior distributions: As in the FMS, the set $\mathcal{B}$ combines *a priori* information with *a posteriori* information. In the FMS the *a priori* information is contained in the candidate model set (CMS) which consists of a noise class and a model class. In the set $\mathcal{B}$, these two classes are defined by means the prior probability distributions of the noise $p_v(v)$ and of the model $p(G)$.

The measurement data $\mathbf{y}$, i.e. the *a posteriori* information, is introduced into the credible set by means the *likelihood function* of the observations $\mathbf{y}$ conditioned to the model G, $p(\mathbf{y}|G)$. Given the model G, this likelihood presents the same probability distribution as the noise, i.e., $p(\mathbf{y}|G) \equiv p_v(\mathbf{y}|G)$.

The posterior distribution $p(G|\mathbf{y})$ of the model G conditioned to the observations $\mathbf{y}$ is obtained by the application of the Bayes' rule,

$$p(G|\mathbf{y}) = \frac{p(\mathbf{y}, G)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|G)p(G)}{\int p(\mathbf{y}|G)p(G)dG} \tag{36}$$

where $p(G)$ is the model *prior distribution*, $p(\mathbf{y}, G)$ is the joint distribution of model and measurements, and the factor $p(\mathbf{y})$ is just a normalizing constant.

Finally, we have

$$p(G|\mathbf{y}) \propto p_v(\mathbf{y}|G) \cdot p(G) \tag{37}$$

where the prior distribution $p(G)$ contains the information about the plant before the data is obtained while the posterior distribution $p(G|\mathbf{y})$ contains the information about the plant updated by the measurements $\mathbf{y}$.

Time domain and frequency domain data: Equation (35) is expressed in terms of time domain data $\mathbf{y} = (y_0, \dots, y_{N-1})^T$, but it can accommodate frequency domain data $\mathbf{G} = (G(\omega_1), \dots, G(\omega_m))^T$ and mixed time domain/frequency domain data as well.

Stochastic nature: The set $\mathcal{B}$ is a *stochastic* characterization of the model set which is consistent with the measurements at hand. The set $\mathcal{B}$ can be viewed as an extension of the probabilistic likelihood-based regions of classical system identification (see Chapter 2) but it goes a step ahead in the sense that it allows the entry of prior knowledge. The appropriate choice of the prior $p(G)$ may reduce the bias error and thus yield smaller uncertainty regions.

Consistency tests: The Bayesian viewpoint allows updating/correcting the prior beliefs. A grossly erroneous $p(G)$ may be detected once the data is collected because, in such a situation, the resulting likelihood of the observations is far from $p(G)$. Also posterior distributions yielding disjoint credible regions may provide useful information about the consistency between the model and the measurements. This feature is useful in *model (in)validation* or *fault detection* procedures where one has to make a decision on the basis of the consistency between model distributions and likelihood of the observations.

Iteration: The possibility of iteration is another advantage of this approach. The posterior distribution obtained by an experiment $p(G|\mathbf{y})$ can be used as the prior distribution for a new experiment. In SMI, it is well known that making more experiments may reduce the size of the FMS. The Bayesian approach implements this process in a formal way.

Principle of stable estimation: The principle of *stable estimation* or *precise measurement* (Edwards, Lindman, and Savage, 1963), (Peterka, 1981), states that, in the case of large or medium length of observation data set (say for N of order of several tens or more), if the data do contain information about the unknown system, and if the likelihood function is well peaked, then even a rather drastic modification of the prior distribution does not significantly change the posterior distribution. Moreover (Berger, 1985) shows that in situations of stable estimation, the posterior can be approximated by a normal distribution.

However, it has to be said that the principle of stable estimation is not a generally valid principle. It applies only when data really carry the information about the parameters which are to be estimated. It does not apply in the cases of redundant, non-identifiable or weakly identifiable parameters.

The practical implication of the principle of stable estimation is that one does not need to worry too much about the choice of the prior distribution and that any prior distribution which is flat relatively to the likelihood function is good enough.

Next example (Eykhoff, 1974) illustrates how the Bayesian estimate effectively converges to the unknown true value when the principle of stable estimation applies.

Example 3.1. Bayesian point estimation of a single parameter

Consider that we want to obtain an estimate $\hat{\theta}$ for a single parameter θ which "unknown" true value is $\theta = 5$. Suppose that we perform $N = 10$ measurements of the parameter. We can express the generation of these measurement data by means the following linear regression model,

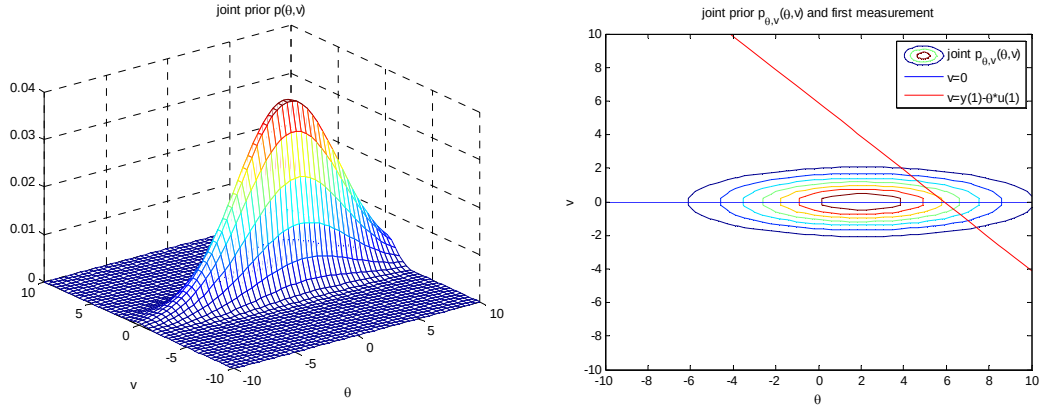
$$y_n = \theta u_n + v_n \quad , \quad n = 1, \dots, N$$

where the excitation u is taken as $u_n = 1, \forall n$. Regarding the measurement noise v, we assume that its probability distribution is standard normal, $v \sim \mathcal{N}(0,1)$ and statistically independent of θ. We also assume that it is stationary, that is, that the distribution $p_v(v)$ does not change with time.

Note that, for a fixed value of the parameter θ, the likelihood of the observation y, $p(y|\theta)$, presents the same probability distribution than the measurement noise, since $p(y|\theta) = p_v(y - \theta u) = p_v(y - \theta)$.

Before making any measurement, we can make a *guess* about the parameter value. For example, we think the value can be 2. But we are not too much convinced about this, so we recognise a standard deviation of, say, 4. The more uncertain we are the greater will be the assumed deviation. So, let us suppose that our prior knowledge about the parameter is normal distributed as $\theta \sim \mathcal{N}(2, 4^2)$.

Fig. 3.1(a) shows the joint prior distribution $p_{\theta,v}(\theta, v) = p_\theta(\theta)p_v(v)$, where $p_\theta(\theta)$ is the parameter subjective prior distribution.



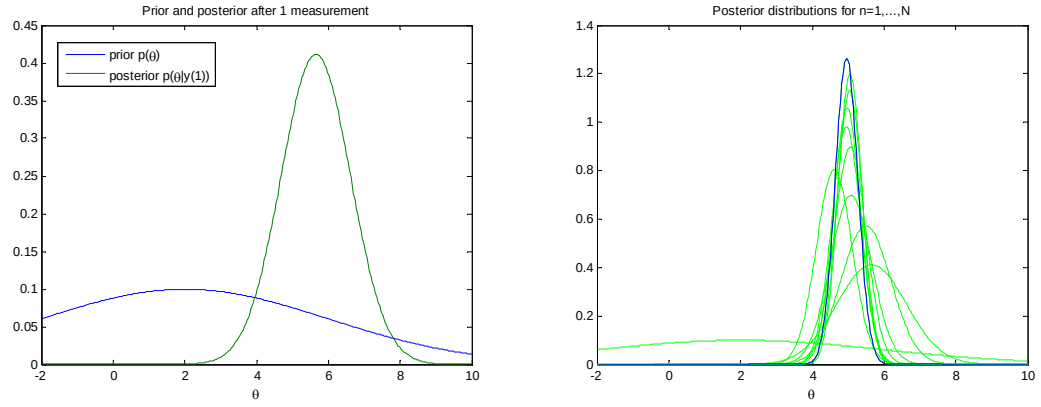
(a) **(b)**

Fig. 3.1. (a) Prior joint distribution of noise and parameter, (b) contour plot and first measurement

Suppose now that the first measurement gives the value $y_1 = 5.8810$. The posterior distribution obtained by means Bayes' rule (36) can be viewed as a "cut" through the prior joint distribution given by the line $v_1 = y_1 - \theta u = 5.8810 - \theta$. Note that the line corresponding to the parameter prior distribution $p_\theta(\theta)$ was $v = 0$ (see Fig. 3.1(b)).

Fig. 3.2(a) shows how this single measurement has improved a lot both our *knowledge* and *certainty feeling* about the unknown parameter.

Finally, suppose that we now use this posterior distribution as prior distribution before taking the second measurement and repeat the procedure for the N measurements. Fig. 3.2(b) shows the improvement as long as new measurements have entered to the model. At each iteration of Bayes' rule, the last posterior distribution obtained served as the new prior distribution.



(a) **(b)**

Fig. 3.2. (a) Prior and posterior distribution after the first measurement, (b) posterior distributions for all the measurements

■

3.1.2 Particular cases of the BCMS

Equation (35) shows a general case for the $\mathcal{B}$. Different classes of $\mathcal{B}$ can be defined if we select different supports for the probability distributions. Here we will consider that the support can be the parameter space $\boldsymbol{\Theta}$, the complex plane, or the model spaces ℓ_1 and $\mathcal{H}_\infty$.

a. BCMS defined in the parameter space

If the support for the model class is the parameter space $\boldsymbol{\Theta}$, we define the Bayesian Credible Parameter Set (BFPS) as

$$\mathcal{B}_{\boldsymbol{\theta}} \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}:\ p(\boldsymbol{\theta}|\mathbf{y}) \geq c(\alpha)\} \tag{38}$$

This set is useful when the structure of the model is fixed and the only uncertainty is in the parameters value. Note that this is the case considered in conventional system identification and set-membership methods which assume the existence of a $\boldsymbol{\theta}_{true}$ such that $G_{true} = G(\boldsymbol{\theta}_{true})$.

b. BCMS defined in the frequency domain

The set $\mathcal{B}$ can be defined in terms of the model frequency response. In this case the support is the complex plane. The Bayesian Credible Frequency Response Region (BCFR) is defined as

$$\mathcal{B}_\omega \equiv \{G(j\omega) \in \mathbb{C}:\ p(G(j\omega)|\mathbf{y}) \geq c(\alpha)\} \tag{39}$$

This case is interesting because we can define a credible region at each frequency (for all frequencies contained in the excitation signal) in the same way than MEM, NSSE and deterministic interpolatory algorithms do. We define such a Bayesian Credible Value Set (BCVS) as

$$\mathcal{B}_{\omega_i} \equiv \{G(j\omega_i) \in \mathbb{C}:\ p(G(j\omega_i)|\mathbf{y}) \geq c(\alpha),\ \ i = 1, \dots, m\} \tag{40}$$

In the robust control application, this latter set is interesting for various reasons. The frequency dependent uncertainty bands $W(\omega)$ needed by robust control techniques can be obtained by combining the credible regions defined in (40). Another advantage of this set is that, at each frequency, the support for the probability distributions is the Nyquist or Nichols plane so the probability distributions are two-dimensional. This fact facilitates the computations and also provides a visual, intuitive representation of the model uncertainty.

c. BCMS defined in the spaces ℓ_1 and $\mathcal{H}_\infty$

Strong connections to robust control can be obtained if the following two credible sets are defined. The Bayesian Credible ℓ_1 Set is defined as

$$\mathcal{B}_{\ell_1} \equiv \{h \in \ell_1:\ p(h|\mathbf{y}) \geq c(\alpha)\} \tag{41}$$

while the Bayesian Credible $\mathcal{H}_\infty$ Set is defined as

$$\mathcal{B}_{\mathcal{H}_\infty} \equiv \{H \in \mathcal{H}_\infty:\ p(H|\mathbf{y}) \geq c(\alpha)\} \tag{42}$$

In the case of $\mathcal{B}_{\ell_1}$ we can consider prior distributions on the impulse response h which satisfy the usual prior knowledge conditions, $|h_n| \leq K\rho^n,\ n = 0, \dots, N-1,\ K > 0,$ $\rho < 1$.

In the case of $\mathcal{B}_{\mathcal{H}_\infty}$ prior distributions on the frequency response G would satisfy $\|G\|_\infty \equiv \sup_{z\in\mathcal{D}}|G(z)| < K$, where $\mathcal{D}$ is the open unit disk $\mathcal{D} = \{z \in \mathbb{C}:\ |z| < 1\}$.

Both sets in spaces ℓ_1 and $\mathcal{H}_\infty$ could be dealt by using tools of the Bayesian nonparametric statistics, which also allows working with infinite dimensional spaces (Robert, 2001).

3.1.3 Bayesian robust identification problem. Methodology

The definition of the Bayesian Credible Model Set allows us dealing with the Bayesian Robust Identification problem in an analogous way than deterministic methods. The whole modelling procedure is described below. Next sections illustrate the development and results of the proposed methodology.

The experiment: In the next sections, we assume that we have collected N input/output measurement data obtained by applying an excitation sequence $\{u_n\}_{n=0}^{N-1}$ to an unknown system G_{true} and collecting the response samples $\{y_n\}_{n=0}^{N-1}$ corrupted by additive measurement noise $\{v_n\}_{n=0}^{N-1}$,

$$y_n = G_{true}(q)u_n + v_n \quad , \quad n = 0, \dots, N-1 \tag{43}$$

where q is the forward shift operator, $qu_n = u_{n+1}$. To simplify the notation, we define $\mathbf{y} = (y_0, \dots, y_{N-1})^T$, $\mathbf{u} = (u_0, \dots, u_{N-1})^T$, and $\mathbf{v} = (v_0, \dots, v_{N-1})^T$.

Prior information or assumptions: If we have any prior information about the plant G_{true} and measurement noise v, it will be used to select the prior distributions for the model $p(G)$ and measurement noise $p_v(v)$.

If we *do not have* any prior information, we must take *assumptions*, more or less educated, about the plant and noise.

A typical choice in stochastic methods is to assume that the noise $\{v_n\}_{n=0}^{N-1}$ is a sequence of stationary i.i.d. (independent identically distributed) white normal noise with zero mean and variance σ_v^2 and that it is independent to the excitation $\{u_n\}_{n=0}^{N-1}$.

A typical assumption about the plant is that it is BIBO (*Bounded Input Bounded Output*) stable with parameters belonging to a certain region of the parameter space.

Assign (subjective) prior probability distributions to the model and noise: Select $p(G)$ and $p_v(v)$ on the basis of the prior information and/or assumptions. In the case of $p(G)$, the support can be the parameter space, the complex plane, or a model space (ℓ_1, $\mathcal{H}_\infty$). Moreover, when there are several sources of uncertainty (structure, order…) mixture and hierarchical distributions can be used.

In a typical Bayesian framework, these distributions are *subjective*, thus indicating the degree of confidence of the engineer on her previous information about the system. Note that prior distributions can also be interpreted as an indicator of the *ignorance* degree.

Compute the likelihood function: That is, compute the sample distribution $p_v(\mathbf{y}|G)$ corresponding to the likelihood of observations $\mathbf{y}$ for the assumed model G, on the basis of the previously defined prior $p_v(v)$. This can be done numerically, for a grid of values for G.

Compute the posterior distribution: Apply the Bayes' rule to the likelihood function $p_v(\mathbf{y}|G)$ and to the prior distribution $p(G)$ in order to obtain the posterior distribution of the model $p(G|\mathbf{y})$. This can be done analytically, numerically o by means MCMC (Markov Chain Monte Carlo) methods.

Obtain the credible regions: Select a probability level $100(1-\alpha)\%$ and compute the corresponding threshold $c = c(\alpha)$ that establishes the size of the Bayesian credible model set $\mathcal{B} \equiv \{G \in \mathcal{G}:\ p(G|\mathbf{y}) \geq c(\alpha)\}$. Obtain the resulting highest posterior density (HPD) regions. These will constitute the uncertainty regions in this framework.

Identify the nominal model: Identify a nominal model on the basis of the posterior model distribution. Several criteria can be used in order to infer models from the posterior. One straightforward possibility is to select the maximum *a posteriori* (MAP) estimate. Another possibility is to define a penalty function and perform minimum risk (MR) estimation. This second choice is useful in order to penalise model complexity, improve the system robustness (by penalty functions derived from robustness theorems) or to take into account the effect of wrong modelling at critical frequencies.

3.2 Construction of the BCMS in the parametric case

In the parametric case, the model is characterized by means a parameter vector $\boldsymbol{\theta}$. To characterize the parametric uncertainty, we will obtain the Bayesian Credible Parameter Set (BFPS),

$$\mathcal{B}_{\boldsymbol{\theta}} \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}:\ p(\boldsymbol{\theta}|\mathbf{y}) \geq c(\alpha)\}$$

Note that the BFPS does not impose any restriction about the linearity of the model. It is also valid for nonlinear models, for instance, for the Wiener and Hammerstein models presented in Appendix B. However, in this section let us illustrate the procedure for the simplest case, namely the linear regression model case with Gaussian noise.

3.2.1 Likelihood of the observations

Measurement data, i.e. *a posteriori* information, is entered to the Bayesian credible model set (BCMS) by means of the sample distribution or likelihood function $p(\mathbf{y}|G, v)$ of the observations $\mathbf{y}$ conditioned to the measurement noise v and the plant model G.

Linear regression model with Gaussian noise: Let us assume that we have gathered N samples of time domain data $\mathbf{y}$, corrupted by additive Gaussian measurement noise v with variance $\lambda = \sigma_v^2$. The linear regression model is parameterized by means of the vector $\boldsymbol{\theta}$, $\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \mathbf{v}$, where the design matrix $\boldsymbol{\Phi}$ is the one defined in Chapter 2.

The likelihood of the observations jointly conditioned to noise and model (parameter vector) coincides in form with the probability distribution of the measurement noise.

$$p(\mathbf{y}|\boldsymbol{\theta},\lambda) \sim p_v(\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta}|\lambda) \tag{44}$$

For the case of Gaussian noise, we have $(\mathbf{y}|\boldsymbol{\theta},\lambda) \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\theta}, \lambda\mathbf{I})$, i.e.,

$$p(\mathbf{y}|\boldsymbol{\theta},\lambda) = \frac{1}{(2\pi\lambda)^{N/2}} \exp\left(-\frac{1}{2\lambda}(\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta})^T(\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta})\right) \tag{45}$$

Once obtained $p(\mathbf{y}|\boldsymbol{\theta},\lambda)$ subsequent unconditional likelihoods can be obtained by marginalization (law of total probability), that is,

$$p(\mathbf{y}|\lambda) = \int p(\mathbf{y}|\boldsymbol{\theta},\lambda)p(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta}$$

$$p(\mathbf{y}) = \int p(\mathbf{y}|\lambda)p(\lambda)d\lambda$$

and

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\boldsymbol{\theta},\lambda)p(\lambda|\boldsymbol{\theta})d\lambda$$

where $p(\lambda|\boldsymbol{\theta})$ can be obtained from $p(\lambda|\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}|\lambda)p(\lambda)}{p(\boldsymbol{\theta})}$.

Recursive computation of the likelihood: The likelihood function can be numerically obtained for a grid of candidate parameter vectors $\boldsymbol{\theta}_i$ by assuming that the error samples $e_n = y_n - \hat{y}_n$, where $\hat{y}_n = \boldsymbol{\varphi}_n^T\boldsymbol{\theta}_i$, are i.i.d. (independent identically distributed)

$$p_v(\boldsymbol{y}|\boldsymbol{\theta}_i,\lambda) = \prod_{n=0}^{N-1} p_v(y_n - \hat{y}_n|\boldsymbol{\theta}_i,\lambda) \tag{46}$$

The expression above is useful for the implementation *on-line*, where new measurements go entering to the model and modifying the likelihood function and posterior model distribution.

Model sets from the likelihood: Actually, the likelihood function is enough to define a model set (Hjalmarsson, 2005). A first model set can be defined by using the expression obtained in Chapter 2:

$$\mathcal{G}_1 = \left\{\boldsymbol{\theta}\colon\ (\widehat{\boldsymbol{\theta}}_{\boldsymbol{N}} - \boldsymbol{\theta})\mathbf{P}_N^{-1}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{N}} - \boldsymbol{\theta}) \le \chi_\alpha^2(d)\right\}$$

where d is the model order, $\mathbf{P}_N = \lambda\mathbf{R}_N^{-1}$ is the covariance matrix, and $\mathbf{R}_N = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is the precision matrix.

Another model set can be defined by using the negative log-likelihood function (see Appendix A),

$$L(\boldsymbol{\theta}|\mathbf{y},\lambda) = -\log p(\mathbf{y}|\boldsymbol{\theta},\lambda) = ct + \frac{1}{2\lambda}(\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta})^T(\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta})$$

The resulting set is:

$$\mathcal{G}_2 = \left\{\boldsymbol{\theta}:\ \frac{1}{2\lambda}(\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta})^T(\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta}) \le \chi^2_\alpha(N)\right\}$$

3.2.2 Computation of the posterior distribution

Model sets $\mathcal{G}_1$ and $\mathcal{G}_2$ are the ones used in classical system identification techniques. In the Bayesian approach these model sets are tuned by means of prior probability distributions containing the prior knowledge about the system. The result is the posterior distribution of the model, $p(\boldsymbol{\theta}|\mathbf{y})$, which allows the characterization of the model uncertainty by means the set $\mathcal{B}_{\boldsymbol{\theta}} \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}:\ p(\boldsymbol{\theta}|\mathbf{y}) \ge c(\alpha)\}$.

Let us illustrate the computation of the posterior distribution for the case of Gaussian prior distributions and Gaussian noise.

Gaussian prior on the parameters: Consider again time domain data and linear regression model. We assume that the prior distribution of the parameter vector is:

$$(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\theta}_0, \lambda\mathbf{R}_0^{-1}) \qquad \textbf{(47)}$$

where we can select arbitrary values for the prior parameter vector $\boldsymbol{\theta}_0$ and for the prior precision matrix $\mathbf{R}_0$. If the noise variance is not known we can substitute λ by $E[\lambda]$.

Likelihood function: Assuming zero mean Gaussian measurement noise with unknown variance λ, the likelihood function is

$$(\mathbf{y}|\boldsymbol{\theta},\lambda) \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\theta}, E[\lambda]\mathbf{I})$$

If the noise variance is known we can substitute $E[\lambda]$ by λ.

Posterior distribution: The resulting posterior distribution is obtained by applying the Bayes' rule. The result is:

$$(\boldsymbol{\theta}|\mathbf{y},\lambda) \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{P}(\mathbf{y})) \qquad \textbf{(48)}$$

with $\widehat{\boldsymbol{\theta}}(\mathbf{y}) = (\mathbf{R}_0 + \mathbf{R}_N)^{-1}(\boldsymbol{\Phi}^T\mathbf{y} + \mathbf{R}_0\boldsymbol{\theta}_0)$ and $\mathbf{P}(\mathbf{y}) = E[\lambda|\mathbf{y}](\mathbf{R}_0 + \mathbf{R}_N)^{-1}$, where $\mathbf{R}_N = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$. Again, if the noise variance were known we can substitute $E[\lambda|\mathbf{y}]$ by λ.

Proof:

The application of the Bayes' rule implies the product $p(\boldsymbol{\theta}|\mathbf{y},\lambda) \propto p(\mathbf{y}|\boldsymbol{\theta},\lambda)\cdot p(\boldsymbol{\theta})$, where

$$p(\mathbf{y}|\boldsymbol{\theta},\lambda) = \frac{1}{(2\pi\lambda)^{N/2}}\exp\left(-\frac{1}{2\lambda}(\mathbf{y}-\boldsymbol{\Phi\theta})^T(\mathbf{y}-\boldsymbol{\Phi\theta})\right)$$

$$p(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}|\lambda\mathbf{R}_0^{-1}|}\exp\left(-\frac{1}{2\lambda}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{R}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right)$$

One the one hand we have:

$$(\mathbf{y}-\boldsymbol{\Phi\theta})^T(\mathbf{y}-\boldsymbol{\Phi\theta}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\boldsymbol{\Phi\theta} - \boldsymbol{\theta}^T\boldsymbol{\Phi}^T\mathbf{y} + \boldsymbol{\theta}^T\mathbf{R}_N\boldsymbol{\theta}$$

$$(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{R}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = \boldsymbol{\theta}^T\mathbf{R}_0\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{R}_0\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^T\mathbf{R}_0\,\boldsymbol{\theta} + \boldsymbol{\theta}_0^T\mathbf{R}_0\boldsymbol{\theta}_0$$

The sum is:

$$\begin{aligned}(\mathbf{y}-\boldsymbol{\Phi\theta})^T(\mathbf{y}-\boldsymbol{\Phi\theta}) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{R}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\\ = \boldsymbol{\theta}^T(\mathbf{R}_0+\mathbf{R}_N)\boldsymbol{\theta} - \boldsymbol{\theta}^T(\mathbf{R}_0\boldsymbol{\theta}_0 + \boldsymbol{\Phi}^T\mathbf{y}) - \left(\boldsymbol{\theta}_0^T\mathbf{R}_0 + \mathbf{y}^T\boldsymbol{\Phi}\right)\boldsymbol{\theta} + \mathbf{y}^T\mathbf{y}\\ + \boldsymbol{\theta}_0^T\mathbf{R}_0\boldsymbol{\theta}_0\end{aligned}$$

Defining $\widehat{\boldsymbol{\theta}}_N = (\mathbf{R}_0+\mathbf{R}_N)^{-1}(\mathbf{R}_0\boldsymbol{\theta}_0 + \boldsymbol{\Phi}^T\mathbf{y})$, we have

$$\begin{aligned}(\mathbf{y}-\boldsymbol{\Phi\theta})^T(\mathbf{y}-\boldsymbol{\Phi\theta}) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{R}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\\ = \boldsymbol{\theta}^T(\mathbf{R}_0+\mathbf{R}_N)\boldsymbol{\theta} - \boldsymbol{\theta}^T(\mathbf{R}_0+\mathbf{R}_N)\widehat{\boldsymbol{\theta}}_N - \widehat{\boldsymbol{\theta}}_N^T(\mathbf{R}_0+\mathbf{R}_N)\boldsymbol{\theta} + \mathbf{y}^T\mathbf{y}\\ + \boldsymbol{\theta}_0^T\mathbf{R}_0\boldsymbol{\theta}_0\end{aligned}$$

Now, note that

$$\widehat{\boldsymbol{\theta}}_N^T(\mathbf{R}_0+\mathbf{R}_N)\widehat{\boldsymbol{\theta}}_N = \mathbf{y}^T\mathbf{y} + \boldsymbol{\theta}_0^T\mathbf{R}_0\boldsymbol{\theta}_0$$

since

$$\widehat{\boldsymbol{\theta}}_N^T(\mathbf{R}_0+\mathbf{R}_N)\widehat{\boldsymbol{\theta}}_N = \left(\boldsymbol{\theta}_0^T\mathbf{R}_0 + \mathbf{y}^T\boldsymbol{\Phi}\right)\widehat{\boldsymbol{\theta}}_N = \left(\boldsymbol{\theta}_0^T\mathbf{R}_0 + \mathbf{y}^T\boldsymbol{\Phi}\right)(\mathbf{R}_0+\mathbf{R}_N)^{-1}(\mathbf{R}_0\boldsymbol{\theta}_0 + \boldsymbol{\Phi}^T\mathbf{y})$$

Finally

$$(\mathbf{y}-\boldsymbol{\Phi\theta})^T(\mathbf{y}-\boldsymbol{\Phi\theta}) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{R}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = \left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_N\right)^T(\mathbf{R}_0+\mathbf{R}_N)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_N\right)$$

$$p(\boldsymbol{\theta}|\mathbf{y},\lambda) \propto \exp\left(-\frac{1}{2\lambda}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_N\right)^T(\mathbf{R}_0+\mathbf{R}_N)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_N\right)\right)$$

■

Note the influence of the prior distribution: if the prior precision matrix $\mathbf{R}_0$ were zero (that is, infinite prior covariance $\mathbf{P}_0 = \lambda\mathbf{R}_0^{-1}$, i.e., no prior knowledge at all), the results

coincide with the ones of the maximum likelihood estimate (least squares estimate) of the classical system identification presented in Chapter 2.

$$(\boldsymbol{\theta}|\mathbf{y},\lambda)\sim \mathcal{N}(\mathbf{R}_N^{-1}\boldsymbol{\Phi}^T\mathbf{y}, E[\lambda|\mathbf{y}]\mathbf{R}_N^{-1}) \tag{49}$$

Also, in a general case, if the number of samples N is small the prior precision $\mathbf{R}_0$ matrix will dominate in the posterior precision matrix, $(\mathbf{R}_0+\mathbf{R}_N)$. As the number of samples N increases, the experimental precision matrix $\mathbf{R}_N$ will dominate in the posterior precision matrix.

Finally, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y},\lambda)$ is the *joint* distribution of all the parameters $\theta_i,\ i=1..d$. If we want to compute the *marginal* distribution of a particular parameter θ_i, we need to solve the following integral,

$$p(\theta_i|\mathbf{y},\lambda)=\int p(\boldsymbol{\theta}|\mathbf{y},\lambda)\,d\theta_1,\ldots,d\theta_{i-1},d\theta_{i+1},\ldots,d\theta_d \tag{50}$$

3.2.3 Credible regions in the parameter space

a. Highest Posterior Density (HPD) regions

Once we have defined the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y},\lambda)$, we need to select a critical value $c(\alpha)$ to bound the Bayesian Credible Model Set and thus define the uncertainty (credible) region. This region will contain all the values of $\boldsymbol{\theta}$ such that $p(\boldsymbol{\theta}|\mathbf{y})\geq c(\alpha)$ where $100(1-\alpha)\%$ is the desired credibility level.

Selection of the credibility level: If the desired credibility level is 95% ($\alpha=0.05$), the "cut" value c corresponds to the region in the support of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ for which the probability content is the 95% of the total, i.e., the integral value of $p(\boldsymbol{\theta}|\mathbf{y})$ in such credible region is the 95% of the integral value of $p(\boldsymbol{\theta}|\mathbf{y})$ in the whole parameter space.

The threshold c must be selected such that there is a small probability that the "true" model may be falsified but not so small that $\mathcal{B}$ could contain models with extremely low probability to occur.

HPD region *vs.* classical confidence region: Fig. 3.3 illustrates the difference between the classical computation of confidence regions and HPD regions. In general, the HPD region is not symmetric about a Bayes point estimator and it is not invariant under transformations, unless the transformation is linear (Box and Tiao, 1973).

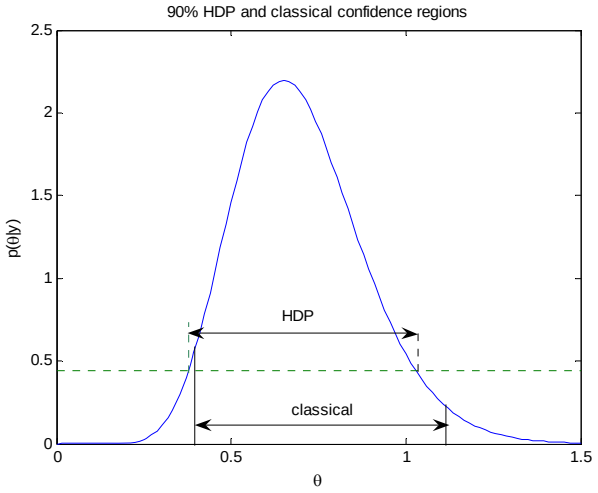


Fig. 3.3. HPD credible interval vs. classical confidence interval (for a posterior given by a Gamma distribution of shape parameter 14 and scale parameter 0.05)

The construction and interpretation of Bayesian credible sets is more straightforward than that of classical confidence sets. But as (Casella and Berger, 2002) point out, nothing comes free. The ease of construction comes because Bayesian models require more assumptions than classical models (definition of prior distributions, for instance).

HPD is optimal in the sense that it gives the smallest region for a given credible probability. In general, HPD regions are smaller than classical confidence regions. And it may happen that a HPD credible set consists of several disjoint intervals. This is a useful situation in model validation and fault detection procedures, since disjoint regions indicate *inconsistency*, e.g., situations where the prior model says one thing and the data another.

Computation of the HPD region in the normal case: One approximation to the computation of HPD credible sets is to consider that the posterior is approximately normal. This assumption is reasonable for large sample sizes and even for small number of samples if the likelihood is normal and the stable estimation principle applies.

For the scalar case, if the posterior $p(\theta|\mathbf{y})$ can be approximated by $\mathcal{N}(\hat{\theta}(\mathbf{y}), \sigma_\theta^2(\mathbf{y}))$, then the approximate $100(1-\alpha)\%$ HPD credible interval is

$$C = \left[\hat{\theta}(\mathbf{y}) - z_{\frac{\alpha}{2}} \cdot \sigma_\theta\ (\mathbf{y}) \quad , \quad \hat{\theta}(\mathbf{y}) + z_{\frac{\alpha}{2}} \cdot \sigma_\theta\ (\mathbf{y}) \quad \right] \tag{51}$$

where $z_{\frac{\alpha}{2}}$ is the $(1-\alpha/2)$-fractile of the standard normal distribution $\mathcal{N}(0,1)$.

For the multivariate case, the posterior density $\mathcal{N}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{P}(\mathbf{y}))$ is large when $\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{y})\right)^T \mathbf{P}(\mathbf{y})^{-1} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{y})\right)$ is small. Furthermore, this quadratic form has a chi-square distribution with d degrees of freedom, so the $100(1-\alpha)\%$ HPD credible interval for $\boldsymbol{\theta}$ is the ellipsoid:

$$C = \left\{ \boldsymbol{\theta}: \ \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\right)^T \mathbf{P}(\mathbf{y})^{-1} \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\right) \leq \chi_\alpha^2(d) \right\} \quad \textbf{(52)}$$

where $\chi_\alpha^2(d)$ is the $(1-\alpha)$-fractile of the chi-square distribution with d degrees of freedom.

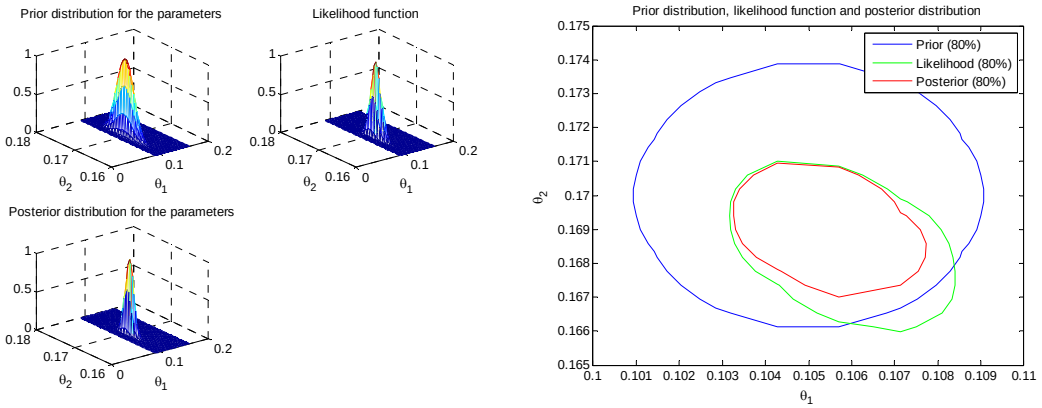
Example 3.2. Credible regions when the noise variance is known

Consider again the plant and experiment of the Example 2.1 (Ninness and Goodwin, 1995). Here, we have used the first $N = 500$ samples of the experiment and we assume that the noise variance is $\lambda = 0.005$.

Firstly, we arbitrarily select a prior distribution for the parameters given by $\mathcal{N}(\boldsymbol{\theta}_0, \lambda \mathbf{R}_0^{-1})$ with $\boldsymbol{\theta}_0 = (0.105 \quad 0.17)^T$ and $\mathbf{R}_0 = 1000 \times \mathbf{I}_{2\times 2}$. The 3D plot and 80% contour plot (blue line) are shown in Fig. 3.4.

Secondly, we compute the likelihood function for the first $N = 500$ samples assuming normal noise with zero mean and variance λ, i.e., $(\mathbf{y}|\boldsymbol{\theta}, \lambda) \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\theta}, \lambda \mathbf{I})$. The 3D plot and 80% contour plot (green line) are shown in Fig. 3.4.

Finally, we compute the posterior distribution for the parameters by combining the prior distribution with the likelihood distribution by means the Bayes' rule. The result is again a Gaussian distribution $\mathcal{N}(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{P}(\mathbf{y}))$ and its 3D plot and 80% contour plot (red line) are shown in Fig. 3.4.



(a) **(b)**

Fig. 3.4. Prior distribution, likelihood function and posterior distribution, (a) 3D plots (b) 80% contour plots

The maximum a posteriori (MAP) value for the parameter vector is $\widehat{\boldsymbol{\theta}}(\mathbf{y}) = (0.1055 \quad 0.1690)^T$ and the associated posterior covariance matrix is

$$\mathbf{P}(\mathbf{y}) = 10^{-5} \times \begin{bmatrix} 0.1319 & -0.0447 \\ -0.0447 & 0.1335 \end{bmatrix}$$

Next figure compares the Bayesian estimate and its credible region to the least squares estimate and its confidence region. In this example, the Bayesian uncertainty region is effectively smaller than the confidence region. Finally, to obtain the 80% credible region in Fig. 3.5 we have used the result (52), while in the Fig. 3.4 the same region was obtained numerically (by "brute force").

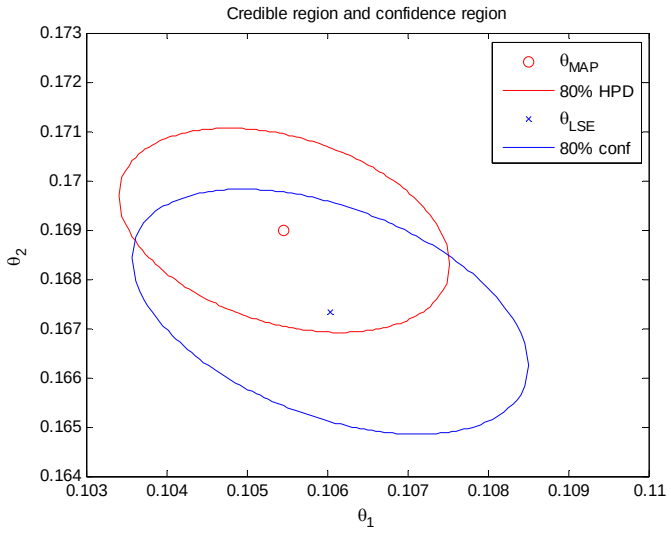


Fig. 3.5. Point estimates (maximum a posteriori and least squares) and 80% uncertainty regions (credible HPD region and confidence region)

■

Estimation of variance: If the noise variance λ is not a known value (it is not fixed), information about it coming from the measurement data is used. In Chapter 2, unbiased estimates for λ coming from time domain data and frequency domain data were presented.

An alternative is to take the variance λ as stochastic with a particular probability distribution. The usual case is to consider that λ follows an inverse Wishart distribution $\lambda \sim \mathcal{W}^{-1}(\sigma, m)$ (Box and Tiao, 1973), (Hjalmarsson and Gustafsson, 1995).

> **Remark**: The Wishart distribution is a generalization of the chi-square distribution to the multivariate case and it is often used as the distribution for the sample covariance matrix for multivariate normal random data. The inverse Wishart distribution, which is based in the Wishart distribution, is then used as the conjugate prior for the covariance matrix of a multivariate normal distribution.

The probability density function (PDF) of the inverse Wishart distribution is

$$p(\lambda) = \frac{\sigma^{\frac{m}{2}} \cdot e^{-\frac{\sigma}{2\lambda}}}{2^{\frac{m}{2}} \cdot \Gamma\left(\frac{m}{2}\right) \cdot \lambda^{\frac{m+2}{2}}}$$

and it has mean $E[\lambda] = \frac{\sigma}{m-2}$ and variance $Var[\lambda] = \frac{2\sigma^2}{(m-2)^2(m-4)}$.

With this selection, the *conditional* prior distribution for the parameter vector is the following *hierarchical* distribution:

$$(\boldsymbol{\theta}|\lambda) \sim \mathcal{N}(\boldsymbol{\theta}_0, E[\lambda]\mathbf{R}_0^{-1}) \ , \ \lambda \sim \mathcal{W}^{-1}(\sigma, m)$$

and the *unconditional* prior distribution for the parameter vector will be no longer Gaussian, it will be a multivariate generalisation of the Student's t distribution with m degrees of freedom

$$\boldsymbol{\theta} \sim t(\boldsymbol{\theta}_0, \sigma\mathbf{R}_0^{-1}, m) = \mathcal{T}(\boldsymbol{\theta}_0, \overline{\mathbf{P}}_0 \ , m)$$

which is given by the following PDF,

$$p(\boldsymbol{\theta}) = (m\pi)^{-d/2}|\overline{\mathbf{P}}_0|^{-1/2}\frac{\Gamma\left(\frac{d+m}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}\left(1 + \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\overline{\mathbf{P}}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}{m}\right)^{-\frac{d+m}{2}}$$

where d is the dimension of the parameter vector. This PDF has mean $E[\boldsymbol{\theta}] = \boldsymbol{\theta}_0$ and covariance $Cov[\boldsymbol{\theta}] = \frac{1}{m-2}\overline{\mathbf{P}}_0 = \frac{\sigma}{m-2}\mathbf{R}_0^{-1} = E[\lambda]\mathbf{R}_0^{-1}$. The value of m is taken as $m = N - d$ where N is the length of the data set.

Assuming Gaussian noise, the application of the Bayes' rule leads to the following posterior distribution for the noise variance (Hjalmarsson and Gustafsson, 1995),

$$\lambda|\mathbf{y} \sim W^{-1}(V_N(\widehat{\boldsymbol{\theta}}_N), N + m)$$

where $V_N(\widehat{\boldsymbol{\theta}}_N) = \sigma + (\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)^T\mathbf{R}_0(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) + \sum_{n=1}^{N}(y_n - \boldsymbol{\varphi}_n^T\widehat{\boldsymbol{\theta}}_N)^2$ and $\widehat{\boldsymbol{\theta}}_N = (\mathbf{R}_0 + \mathbf{R}_N)^{-1}(\boldsymbol{\Phi}^T\mathbf{y} + \mathbf{R}_0\boldsymbol{\theta}_0)$. The posterior expected value for the noise variance is $E[\lambda|\mathbf{y}] = \frac{V_N(\widehat{\boldsymbol{\theta}}_N)}{N+m-2}$.

Finally, the posterior distribution for the parameter vector is given by

$$\boldsymbol{\theta}|\mathbf{y} \sim \mathcal{T}(\widehat{\boldsymbol{\theta}}_N, V_N(\widehat{\boldsymbol{\theta}}_N)(\mathbf{R}_0 + \mathbf{R}_N)^{-1}, N + m)$$

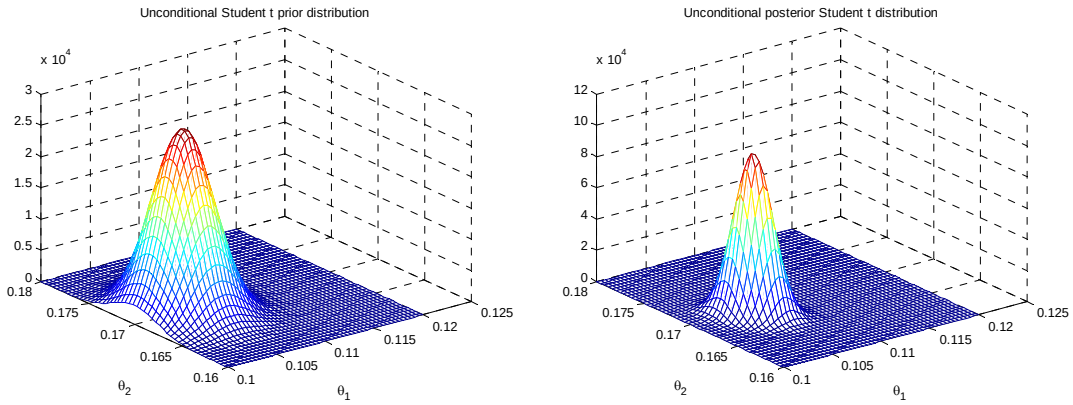
with posterior mean $E[\boldsymbol{\theta}|\mathbf{y}] = \widehat{\boldsymbol{\theta}}_N$ and posterior covariance $Cov[\boldsymbol{\theta}|\mathbf{y}] = E[\lambda|\mathbf{y}](\mathbf{R}_0 + \mathbf{R}_N)^{-1}$. Under mild conditions, this distribution tends asymptotically to be normal.

Example 3.3. Credible regions when the noise variance is unknown using a hierarchical prior.

Consider again the plant and the first $\mathrm{N} = 500$ samples of the experiment of the Example 3.2.

In the first place, we have extracted from data a noise realization by comparing the measured plant output with an estimated one (derived from a second order $d = 2$ Laguerre least squares model). This sequence fits an inverse Wishart distribution with parameter $\sigma = 2.749$ and $m = N - d$ degrees of freedom. The mean value for the prior noise variance is $E[\lambda] = 0.0055$ and the variance value for the prior noise variance is $Var[\lambda] = 1.2 \cdot 10^{-4}$. The mean value for the noise variance *a posteriori* obtained by application of the Bayes' rule to the $\mathrm{N} = 500$ samples and considering Gaussian noise has been $E[\lambda|\mathbf{y}] = 0.0061$.

Fig. 3.6 shows the unconditional Student t prior and posterior distributions obtained for the parameters. Since they are very close to the normal distribution, the results are very similar to the ones obtained in the previous example.



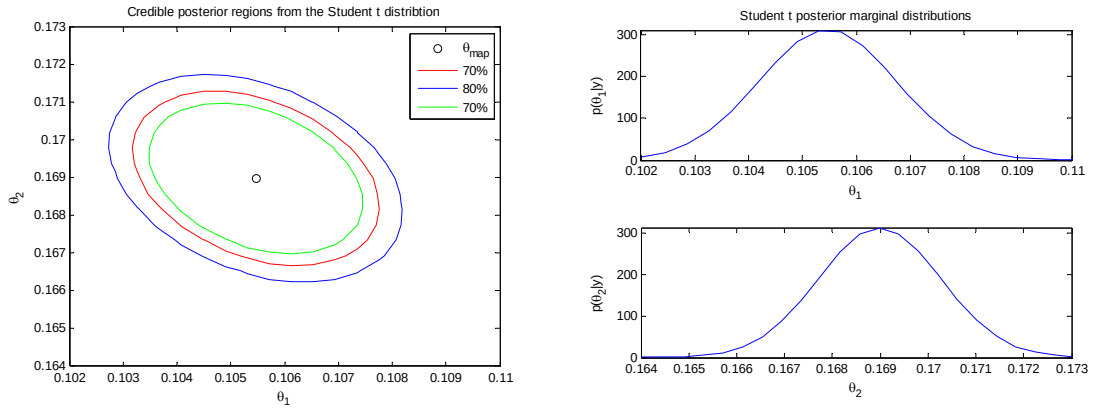
(a) **(b)**

Fig. 3.6. Unconditional Student t distributions for the parameters: (a) Prior distribution (b) Posterior distribution

The maximum a posteriori (MAP) value for the parameter vector is $\widehat{\boldsymbol{\theta}}(\mathbf{y}) = (0.1055 \quad 0.1690)^T$ and the associated posterior covariance matrix is

$$\mathbf{P}(\mathbf{y}) = 10^{-5} \times \begin{bmatrix} 0.1614 & -0.0548 \\ -0.0548 & 0.1633 \end{bmatrix}$$

Fig. 3.6(a) shows the 70%, 80% and 90% posterior credible regions obtained from the unconditional Student t posterior distribution of the parameters. And, finally, Fig. 3.6(a) shows the marginal posterior distributions $p(\theta_1|\mathbf{y})$ and $p(\theta_2|\mathbf{y})$ obtained by direct numerical integration of the unconditional posterior Student t distribution.



(a) **(b)**

Fig. 3.7. (a) Credible posterior regions (70%, 80%, 90%) (b) Marginal posterior distributions

■

b. MCMC implementation

In the previous section the computation of probability distributions and credible regions was easy. In the cases where the number of parameters is moderately high and/or distributions are not standard, the demand for computationally resources increases significantly. In such cases, one must use simulation strategies such as the Markov Chain Monte Carlo (MCMC) simulation. The idea of MCMC is to construct an ergodic Markov chain with invariant distribution equal to the desired posterior. This approach is also interesting because error bounds on estimates are derived from the sampled distribution and thus they do not rely on assumptions of large data records. See Appendix C for a throughout explanation.

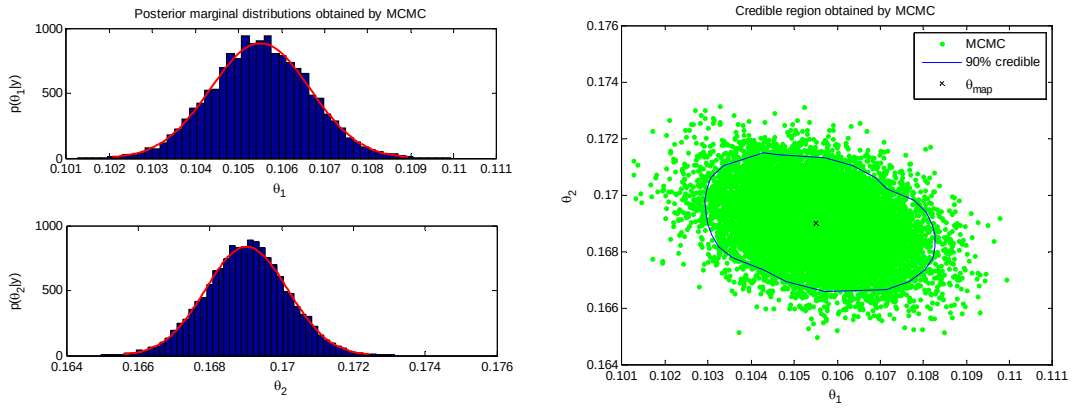
Example 3.4. Credible regions obtained by MCMC simulations

Consider again the posterior distribution for the parameters of the Example 3.2, $\mathcal{N}(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{P}(\mathbf{y}))$, where the mean value is $\widehat{\boldsymbol{\theta}}(\mathbf{y}) = (0.1055 \quad 0.1690)^T$ and the covariance matrix is

$$\mathbf{P}(\mathbf{y}) = 10^{-5} \times \begin{bmatrix} 0.1319 & -0.0447 \\ -0.0447 & 0.1335 \end{bmatrix}$$

Exact credible regions were obtained in the Example 3.2. Here we approximate the credible region by means the use of the slice sampler, which is a version of the Gibbs sampler.

After a burn-in stage, the computation of the histogram of the Markov chain associated at each parameter coincides with the target marginal distributions. Fig. 3.8(a) shows the results for a chain of 15000 samples.



(a) **(b)**

Fig. 3.8. (a) Histograms for the posterior marginals (b) Credible posterior region

The credible region in the parameter space can be obtained by combining the Markov chains (see Fig. 3.8(b)). Note that the region obtained after the burn-in period is effectively the same of the Example 3.2.

■

3.2.4 Relationship to robust identification deterministic methods

Many deterministic methods presented in the Chapter 2 can be viewed as particular cases of the Bayesian framework. In this section, we illustrate how the Feasible Parameter Set (FPS) regions of Chapter 2 can be obtained by means the Bayesian method by simply assuming that the noise is uniform-distributed.

As in the deterministic case the first step is to decide which the value of the bound δ is. In the FPS case this value is used to obtain the different strips corresponding to the measurements. The intersection of all the strips produces the FPS region.

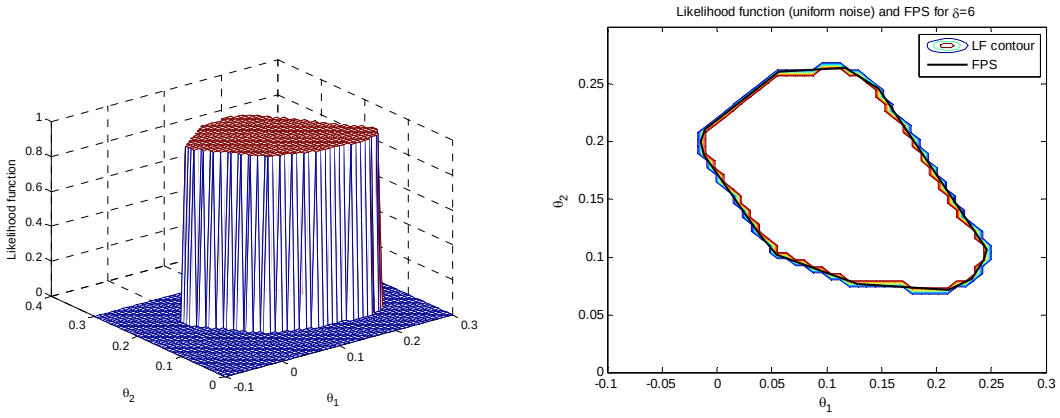
In the Bayesian methodology, the bound δ can be used to define the prior noise distribution. In this section we assume that the noise is uniform distributed $v \sim \mathcal{U}(-\delta, \delta)$. This distribution is used to compute the likelihood function $p_v(\mathbf{y}|\boldsymbol{\theta})$ by taking the parameter space as support. Since the distribution $p_v(v)$ is uniform, the resulting likelihood function will be nonzero and flat in the region where models (parameters) are consistent with measurements and it will be zero outside this region.

Example 3.5. Relationship between FPS and parametric BCMS

Consider again the plant and experiment of (Ninness and Goodwin, 1995). Even though we know that the measurement noise corrupting the data is Gaussian distributed, we chose to model it by means a uniform distribution $v \sim \mathcal{U}(-\delta, \delta)$ with $\delta = 0.6$.

The support values, i.e., the tentative values for the uncertain parameters θ_1 and θ_2, have been selected around the LSE Laguerre model obtained in (Ninness and Goodwin, 1995).

Fig. 3.9(a) shows the resulting normalized likelihood function (LF). In Fig. 3.9(b) the contour of this LF is compared to the FPS region obtained for $\delta = 0.6$ in Chapter 2 (Example 2.7). The two regions effectively coincide.



(a) **(b)**

Fig. 3.9. (a) (Normalized) likelihood function, **(b)** Likelihood function contour and FPS

■

3.2.5 Other features of the Bayesian approach

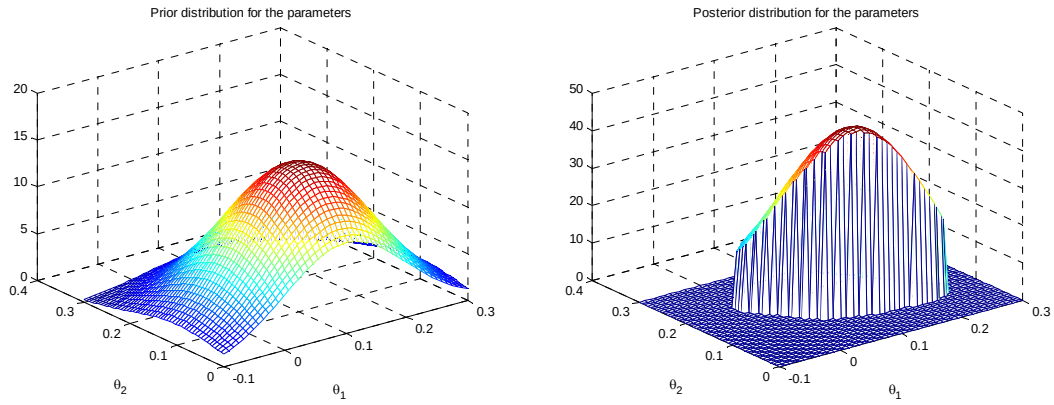
a. FPS with inner probability

In the previous example, all models at the top of the likelihood function, i.e. inside the FPS region, are equally probable to occur. This fact does not facilitate the selection a unique optimal parameter vector.

The Bayesian methodology can go one step beyond from the FPS since it allows assigning a probability to each model by the definition of a prior distribution on the parameters and subsequent computation of the posterior distribution.

Example 3.6. FPS with inner probability

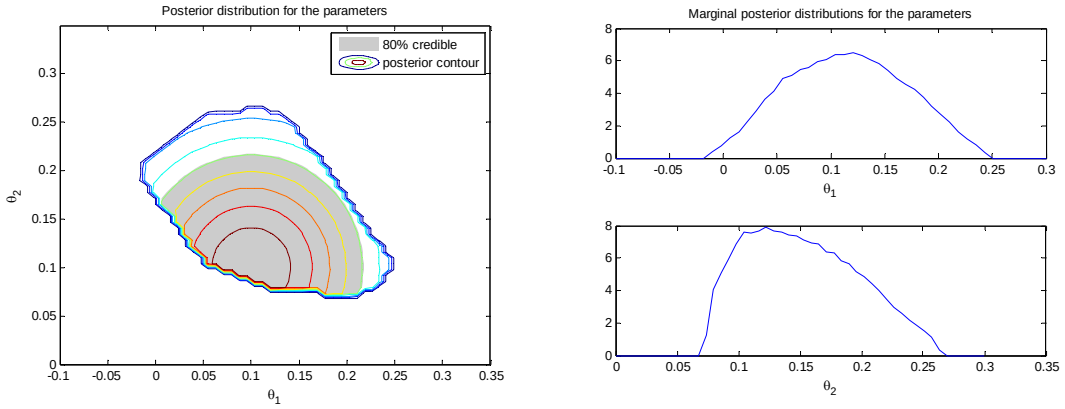
Consider again the plant and experiment of the Example 3.5. Now we assign a prior distribution to the parameters given by $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \mathbf{R}_0^{-1})$ with $\boldsymbol{\theta}_0 = (0.1 \quad 0.1)^T$ and $\mathbf{R}_0 = 100 \times \mathbf{I}_{2\times 2}$ (see Fig. 3.10(a)). The combination of this prior distribution to the uniform likelihood function of the Example 3.5 gives the posterior distribution of the parameters shown in Fig. 3.10(b).



(a) **(b)**

Fig. 3.10. (a) Prior distribution of the parameters, (b) Posterior distribution of the parameters

Fig. 3.11(a) shows the contour plot of the posterior distribution with the 80% credible region shadowed. And Fig. 3.11(b) shows the posterior marginal distributions obtained by numerical integration. The maximum a posteriori parameter vector is $\hat{\boldsymbol{\theta}}(\mathbf{y}) = (0.1062 \quad 0.1667)^T$. This value could be considered as the nominal model.



(a) **(b)**

Fig. 3.11. (a) Contour of the posterior distribution, (b) Marginal posterior distributions

■

b. Iterative computation

Another feature that is very interesting for on line fault detection purposes is that the computation of the likelihood function can be performed iteratively, sample to sample, by using the recursive expression of equation (46). See next example.

Example 3.7. Iterative computation of the uncertainty region.

Consider again the plant and experiment of the Example 3.5. Now we assume $\delta = 0.4$ and a parameter grid of 80×80. With the first 10 samples of the data record we have

obtained the region shown in Fig. 3.12(a). The other plots show the likelihood function updated as new measurements are used.

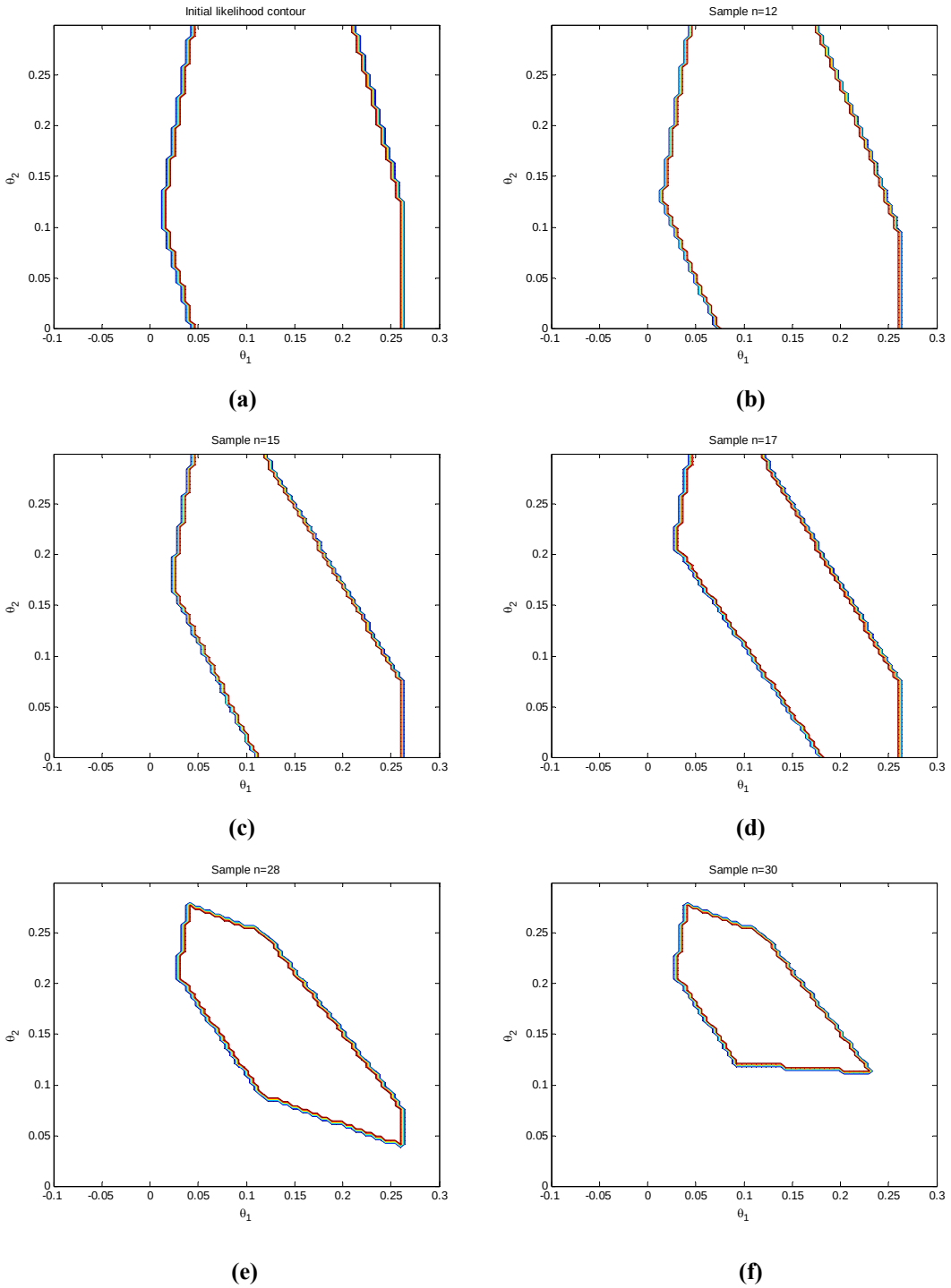


Fig. 3.12. (a) Initial uncertainty region (N=10 samples), and uncertainty regions obtained for the samples (b) 12, (c) 15, (d) 17, (e) 28, and (f) 30.

■

c. Disjoint credible regions

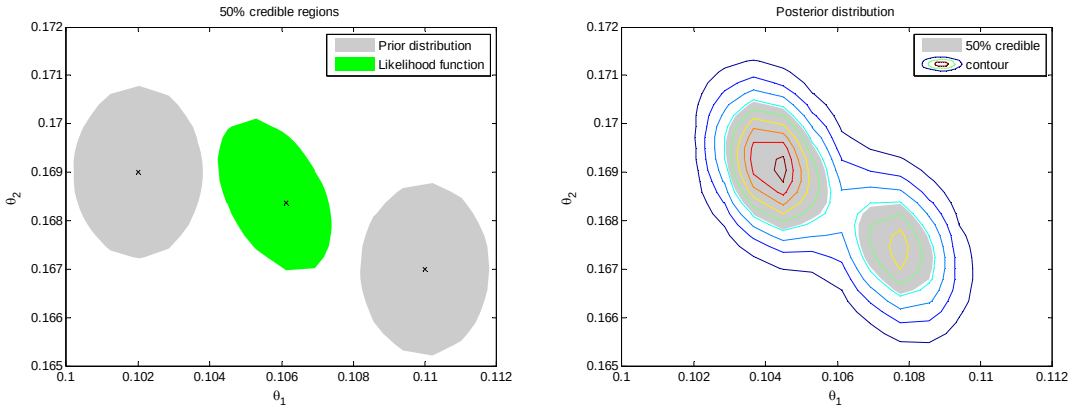
Unlike confidence regions, credible regions can be disjoint. This can be the case when we use mixture distributions to model the system. See next example.

Example 3.8. Disjoint credible regions

Consider again the plant and experiment of Example 3.2. The prior distribution of the system parameters is assumed to be a mixture of two Gaussian distributions. In the first one, the mean value is $\boldsymbol{\theta}_0^{(1)} = (0.102 \quad 0.169)^T$ and the precision matrix is $\mathbf{R}_0^{(1)} = 2000 \cdot \mathbf{I}_{2\times 2}$. And in the second one, the mean value is $\boldsymbol{\theta}_0^{(2)} = (0.11 \quad 0.167)^T$ and the precision matrix is $\mathbf{R}_0^{(2)} = 2000 \cdot \mathbf{I}_{2\times 2}$. The resulting 50% disjoint credible region is shown in Fig. 3.13(a).

The likelihood function is computed from N=500 samples assuming that the noise is Gaussian-distributed with zero mean and variance 0.005.

The resulting posterior distribution is shown in Fig. 3.13(b). The 50% posterior credible region is closer to the likelihood function but is still disjoint. However, one of the two peaks is taller. The maximum a posteriori value $\widehat{\boldsymbol{\theta}}_N^{(MAP)} = (0.0947 \quad 0.1717)^T$ indicates that $\boldsymbol{\theta}_0^{(1)}$ was closer to the measurement data (likelihood function) than $\boldsymbol{\theta}_0^{(2)}$.



(a) **(b)**

Fig. 3.13. (a) Prior distribution and likelihood function, (b) Posterior distribution

■

3.3 Construction of the BCMS in the frequency domain

In this section we will obtain uncertainty bands in the frequency domain. In particular we illustrate the computation of the Bayesian Credible Value Set (BCVS),

$$\mathcal{B}_{\omega_i} \equiv \{G(j\omega_i) \in \mathbb{C}: \ p(G(j\omega_i)|\mathbf{y}) \geq c(\alpha), \ \ i = 1, \ldots, m\}$$

Frequency domain data: The Bayesian Credible Value Set can also be specified in terms of frequency domain data

$$\mathcal{B}_{\omega_i} \equiv \left\{G(j\omega_i) \in \mathbb{C}:\ p\left(G(j\omega_i)\middle|\widehat{\mathbf{G}}\right) \geq c(\alpha),\ \ i = 1, \dots, m\right\} \tag{53}$$

where $\widehat{\mathbf{G}} = (\hat{g}^R(\omega_1) \quad \hat{g}^I(\omega_1) \quad \dots \quad \hat{g}^R(\omega_m) \quad \hat{g}^I(\omega_m))^T$ is the vector containing the estimates of the true frequency response at selected frequencies $\omega_1, \dots, \omega_m$. See Chapter 2.

3.3.1 Finite set of competing models

Let us assume that we are uncertain not only about the model parameters but also about the model structure. A simple approach to cope with the uncertainty in the model structure is to consider a finite set of competing models $\{M_1, \dots, M_K\}$, which constitute the candidate models. These models can be of different orders (if we are uncertain about the order), of different basis functions (if we are uncertain about the parameterization), and so on.

> **Remark**: Note that this approach is also valid for the case of parametric uncertainty. In such a case, the set of competing model contains only one model with uncertain parameters.

We can assume that, *a priori*, all models are equally probable, with pmf (probability mass function) $p(M_k) = 1/K,\ \ k = 1, \dots, K,$ or we may assign a prior belief or preference to each model.

Moreover, considering each of the models M_k, we can include the uncertainty in its parameters. In a general case, the joint uncertainty (model structure, model parameters, measurement noise v) can be expressed by means the use of hierarchical models of the type $(\boldsymbol{\theta}_k|M_k, v)$ where $\boldsymbol{\theta}_k$ is the vector of parameters of model M_k. The hierarchy is, for instance, $\boldsymbol{\theta}_k|M_k, \lambda \sim \mathcal{N}(\boldsymbol{\theta}, \lambda\mathbf{R}^{-1})$ and $(M_k) \sim \mathcal{U}$, where λ is the measurement noise variance, $\boldsymbol{\theta}$ is the mean value for the parameter vector and $\mathbf{R}$ is the precision matrix.

3.3.2 Credible regions in the frequency domain

a. Prior distributions in the Nyquist plane

For simplicity, let us assume Gaussian distributions. The prior distribution of the parameter vector conditioned to the model structure and noise variance is

$$\boldsymbol{\theta}_k|M_k, \lambda \sim \mathcal{N}(\boldsymbol{\theta}_0, \lambda\mathbf{R}_0^{-1}) \tag{54}$$

To translate this parametric uncertainty to uncertainty in the frequency response, we can proceed as in the Chapter 2 and define, for each frequency point ω_i,

$$\mathbf{\Gamma}\left(e^{j\omega_i}\right) = \begin{pmatrix} \mathrm{Re}\left(\mathbf{B}\left(e^{j\omega_i}\right)\right) \\ \mathrm{Im}\left(\mathbf{B}\left(e^{j\omega_i}\right)\right) \end{pmatrix}$$

where $\mathbf{B}\left(e^{j\omega_i}\right) = \left(B_0\left(e^{j\omega_i}\right) \quad \ldots \quad B_{d-1}\left(e^{j\omega_i}\right)\right)$ contains the frequency response of the basis functions $B_i(q)$ that parameterize the model

$$M_k \equiv G(q, \boldsymbol{\theta}_k) = \sum_{i=0}^{d-1} B_i(q)\theta_{k,i}$$

In the FIR case, these basis functions are simply $B_i(q) = q^{-i}$.

Now, the prior distribution for the frequency response of model M_k at frequency ω_i is:

$$G\left(e^{j\omega_i}, \boldsymbol{\theta}_k\right)|M_k, \lambda \sim \mathcal{N}\left(\mathbf{\Gamma}\left(e^{j\omega_i}\right)\boldsymbol{\theta}_0, \lambda\mathbf{\Gamma}\left(e^{j\omega_i}\right)\mathbf{R}_0^{-1}\mathbf{\Gamma}^*\left(e^{j\omega_i}\right)\right) \qquad \textbf{(55)}$$

where the superscript * means conjugate transpose.

For a fixed model M_k, the expression (55) defines a two dimension Gaussian bell at each frequency point ω_i in the Nyquist plane. The resulting credible regions at each frequency are ellipses that altogether define an uncertainty band for the frequency response of model M_k.

Since we wish to obtain a unique credible region characterizing the total uncertainty (for all the models M_k, i.e., structure plus parameters uncertainty) we can apply the law of total probability,

$$p\left(G\left(e^{j\omega_i}, \boldsymbol{\theta}_k\right)|\lambda\right) = \sum_{k=1}^{K} p\left(G\left(e^{j\omega_i}, \boldsymbol{\theta}_k\right)|M_k, \lambda\right) p(M_k) \qquad \textbf{(56)}$$

At each frequency point ω_i the result is a mixture distribution and the credible region is no longer an ellipse.

Example 3.9. Prior credible regions on the Nyquist plane for a set of competing models

Consider that the unknown plant is $ZOH\left\{\frac{1}{(s+1)(s+2)}\right\}$ with $T_s = 1s$. Before performing any experiment, we can obtain an uncertainty band in the Nyquist plane by simply translating to prior distributions our prior knowledge about the plant and measurement noise.

Regarding the model structure, the model set under consideration consists of four FIR-type competing models of orders 2 to 5. Each model M_k has associated a prior

probability of $p(M_k) = 1/4$, that is, we assume that the four structures are *a priori* equally probable.

Regarding the model parameters, we assume that vectors $\boldsymbol{\theta}_k$ are normally distributed, with mean $\boldsymbol{\theta}_0 = 0.2 \times \mathbf{1}_d$ and precision matrix $\mathbf{R}_0 = \mathbf{I}_{d\times d}$, being d the model order, $d = 2, \ldots, 5$.

And finally, regarding the measurement noise, we assume that the variance is $\lambda = 0.01$.

Fig. 3.14 shows the prior distributions for the four models at frequency $\omega = 1.1654 rad/s$.

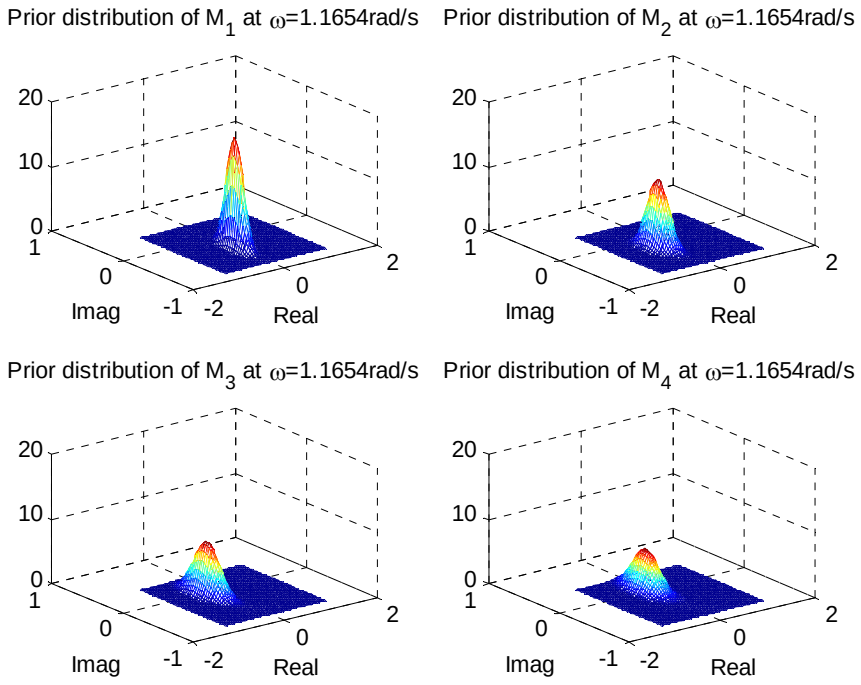
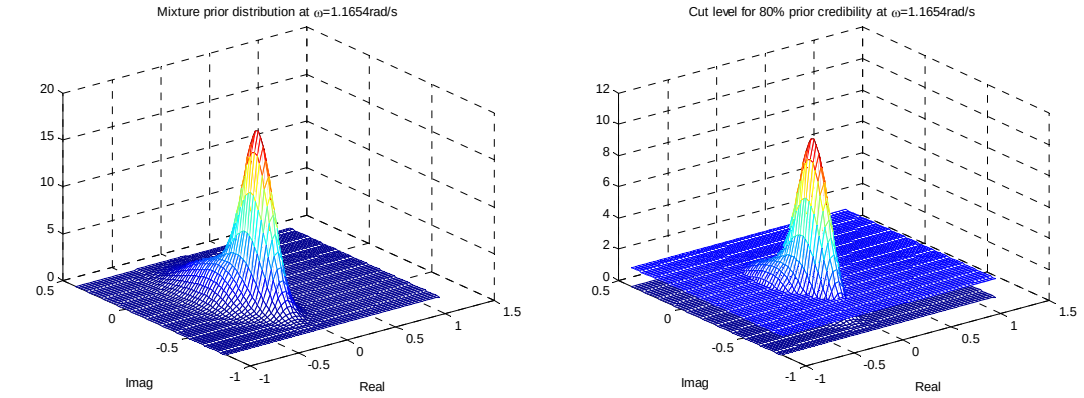


Fig. 3.14. Prior distributions of the four competing models at 1.1654rad/s

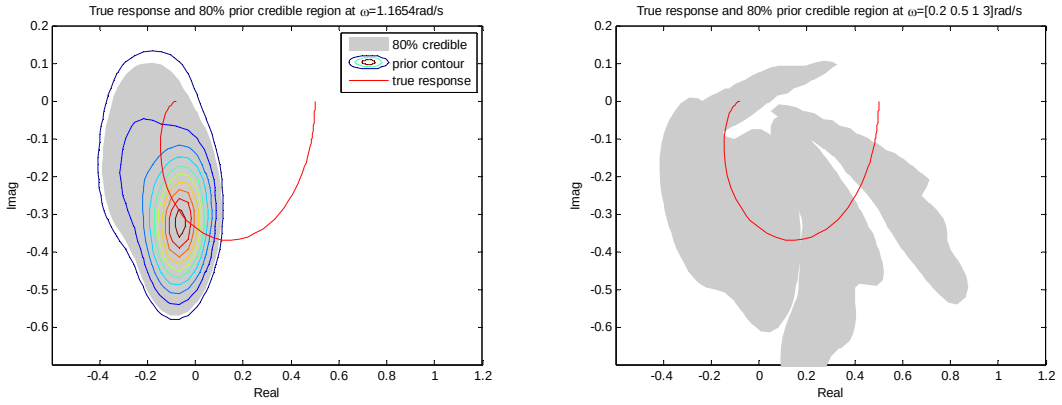
The combination of the four prior distributions by means the application of (56) gives the mixture distribution at frequency $\omega = 1.1654 rad/s$ shown in Fig. 3.15(a). The 80% credible region in the Nyquist plane is obtained by cutting this distribution at the level $c = 7.0607$, such that the integral above is the 80% of the total integral of the mixture distribution (see Fig. 3.15(b)).



(a) **(b)**

Fig. 3.15. (a) Mixture prior at 1.1654rad/s, (b) Cut to obtain the 80% credible region

Finally, Fig. 3.16(a) shows the prior 80% credible region at 1.1654rad/s and Fig. 3.16(b) shows the prior 80% credible region at several frequencies.



(a) **(b)**

Fig. 3.16. (a) Prior 80% credible region at 1.1654rad/s, (b) Prior 80% credible region at frequencies 0.2, 0.5, 1 and 3 rad/s

■

b. Posterior distributions in the Nyquist plane

The posterior probability of the model M_k can be obtained by means the application of the Bayes' rule and the law of total probability as follows:

The joint posterior distribution of the model M_k and parameters $\boldsymbol{\theta}_k$ is

$$p(M_k, \boldsymbol{\theta}_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k, \boldsymbol{\theta}_k) \cdot p(M_k, \boldsymbol{\theta}_k)}{\int_{\boldsymbol{\Theta}} p(\mathbf{y}|M_k, \boldsymbol{\theta}_k) \cdot p(M_k, \boldsymbol{\theta}_k) \cdot d\boldsymbol{\theta}_k} \tag{57}$$

Since the joint prior distribution can be factorized as $p(M_k, \boldsymbol{\theta}_k) = p(\boldsymbol{\theta}_k|M_k) \cdot p(M_k)$, we have

$$p(M_k, \boldsymbol{\theta}_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k, \boldsymbol{\theta}_k) \cdot p(M_k, \boldsymbol{\theta}_k)}{\sum_{i=1}^{K} \left(\int_{\boldsymbol{\Theta}} p(\mathbf{y}|M_i, \boldsymbol{\theta}_i) \cdot p(\boldsymbol{\theta}_i|M_i) \cdot d\boldsymbol{\theta}_i \right) \cdot p(M_i)} \tag{58}$$

where we can define $p(\mathbf{y}|M_i) \equiv \int_{\boldsymbol{\Theta}} p(\mathbf{y}|M_i, \boldsymbol{\theta}_i) \cdot p(\boldsymbol{\theta}_i|M_i) \cdot d\boldsymbol{\theta}_i$ as the integrated likelihood of model M_i. Equation (58) is then expressed as

$$p(M_k, \boldsymbol{\theta}_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k, \boldsymbol{\theta}_k) \cdot p(M_k, \boldsymbol{\theta}_k)}{\sum_{i=1}^{K} p(\mathbf{y}|M_i) \cdot p(M_i)}$$

The posterior probability of model M_k is given by

$$p(M_k|\mathbf{y}) = \frac{\left(\int_{\boldsymbol{\Theta}} p(\mathbf{y}|M_k, \boldsymbol{\theta}_k) \cdot p(\boldsymbol{\theta}_k|M_k) \cdot d\boldsymbol{\theta}_k \right) \cdot p(M_k)}{\sum_{i=1}^{K} p(\mathbf{y}|M_i) \cdot p(M_i)}$$

which can be expressed in a more compact form as:

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k) \cdot p(M_k)}{\sum_{i=1}^{K} p(\mathbf{y}|M_i) \cdot p(M_i)} \tag{59}$$

The expression above gives us the updated probability for each model M_k of the model set once we have gathered the experimental data. This is the solution to the problem known as "model classification" which is a classical problem in the field of Bayesian modelling (Peterka, 1981).

Finally, we use these model probabilities to obtain the mixture posterior distribution at each frequency,

$$p\left(G\left(e^{j\omega_i}, \boldsymbol{\theta}_k\right)|\lambda, \mathbf{y}\right) = \sum_{k=1}^{K} p\left(G\left(e^{j\omega_i}, \boldsymbol{\theta}_k\right)|M_k, \lambda, \mathbf{y}\right) p(M_k|\mathbf{y}) \tag{60}$$

where the posterior distribution for each model M_k and frequency ω_i is, in the Gaussian case,

$$G\left(e^{j\omega_i}, \hat{\boldsymbol{\theta}}_k(\mathbf{y})\right)|M_k, \lambda \sim \mathcal{N}\left(\boldsymbol{\Gamma}\left(e^{j\omega_i}\right)\hat{\boldsymbol{\theta}}_k(\mathbf{y}), \boldsymbol{\Gamma}\left(e^{j\omega_i}\right)\mathbf{P}_k(\mathbf{y})\boldsymbol{\Gamma}^*\left(e^{j\omega_i}\right)\right) \tag{61}$$

Example 3.10. Posterior credible regions on the Nyquist plane for a set of competing models

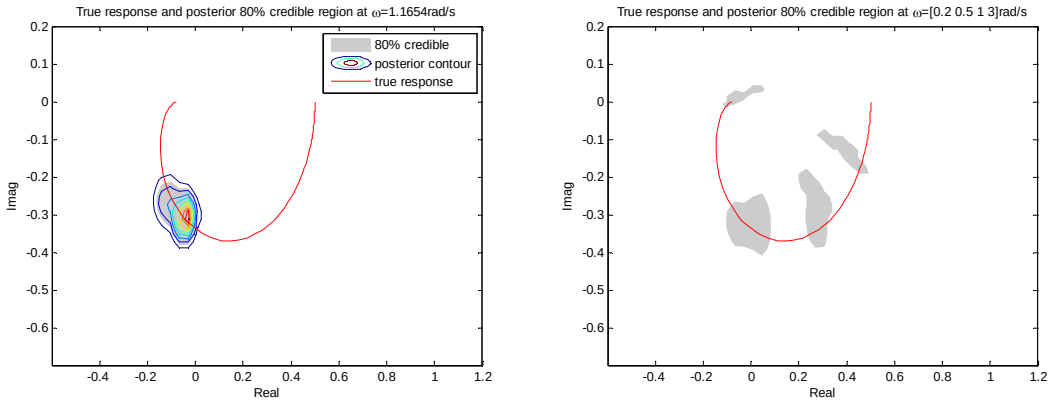
Let us continue with the Example 3.9. Now we excite the plant with $N = 1000$ samples of a PRBS (Pseudo Random Binary Signal) and collect the response samples. For each model we compute the integrated likelihood $p(\mathbf{y}|M_i)$ and apply (59) to obtain the new model probabilities.

After the experiment, the new model probabilities are:

$P(M_1|\mathbf{y}) = 0.3737$, $P(M_2|\mathbf{y}) = 0.2540$, $P(M_3|\mathbf{y}) = 0.2341$, $P(M_1|\mathbf{y}) = 0.1381$

which indicates that a low order model is more probable to have generated the data.

By means the law of total probability (60), these probabilities have been used to obtain the mixture distributions at each frequency, assuming that the individual (per model) distributions are Gaussian. Fig. 3.17 shows the resulting posterior 80% credible regions, (a) at a single frequency, and (b) for a set of frequencies.



(a) **(b)**

Fig. 3.17. (a) Posterior 80% credible region at 1.1654rad/s, (b) Posterior 80% credible region at frequencies 0.2, 0.5, 1 and 3 rad/s

■

3.3.3 Relationship to robust identification stochastic methods

The MEM-OE approach of Chapter 2 can be also viewed as a particular case of the Bayesian approach. In the MEM framework, it is assumed that the measurements explicitly depend on a nominal model $G\left(q,\widehat{\boldsymbol{\theta}}_N\right)$ and a model error $G_e(q)$ accounting for the undermodelling

$$y_n = G\left(q,\widehat{\boldsymbol{\theta}}_N\right)u_n + G_e(q)u_n + v_n \ , \ n = 1,\dots,N$$

This model error $G_e(q)$ can be interpreted as a black box model where the input are the excitation samples u_n and the output are the residuals $\varepsilon_n = y_n - G(q,\widehat{\boldsymbol{\theta}}_N)u_n$, and it can be obtained by means the same techniques used for nominal models, for instance, using Output Models and Least Squares Estimation. In this case, the uncertainty (confidence) regions are obtained assuming Gaussian noise.

Hence, similar results are obtained by the Bayesian methodology if we assume Gaussian noise and flat parameters prior (i.e., the only assumption about the model is the structure but no prior value is assigned to the parameters).

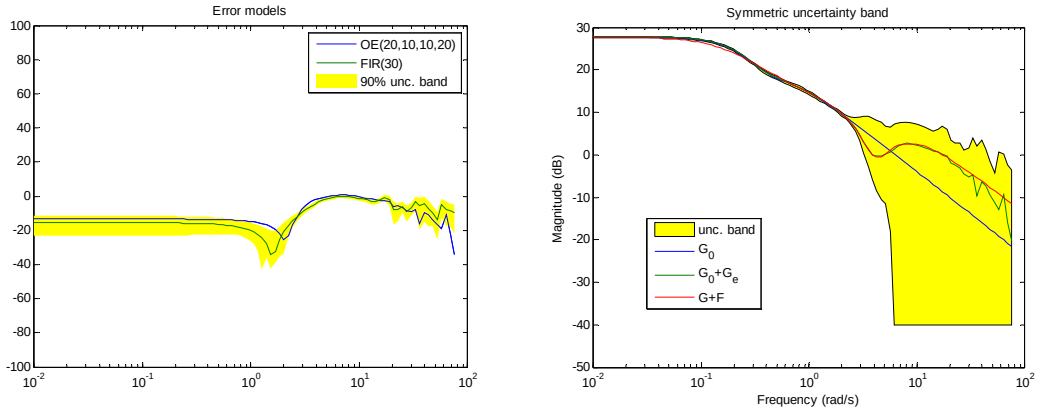
Now the likelihood of observations is substituted by the likelihood of residuals. If we assume that the noise is distributed as $p_v(v)$, then the likelihood function is

$$p(\boldsymbol{\varepsilon}|G_e) = p_v(\boldsymbol{\varepsilon} - G_e\mathbf{u}|G_e)$$

Therefore if we model the noise as Gaussian and compute the likelihood function of the observations, the different "cuts" of this function will lead to the stochastic uncertainty regions, either on the parameter space or in the Nyquist plane.

Example 3.11. Relationship with MEM-PEM and BCMS

Consider the first dataset of the (Reinelt *et al*., 2002). In the present example, we have modeled the model error by means a FIR model of order 30 and have assumed Gaussian noise of zero mean and variance estimated from the residuals by the expression (10) given in Chapter 2, $E[\lambda] = 0.0844$. Ellipses in the Nyquist plane have been obtained by cutting the resulting likelihood function at each frequency to obtain a confidence level of 95%. Fig. 3.18(a) shows the model error with its associated uncertainty band. For the sake of comparison it is show the OE-type model error obtained in (Reinelt *et al*., 2002).



(a) **(b)**

Fig. 3.18. (a) Model error and associated uncertainty band, (b) Final symmetric uncertainty band around the nominal model

Final uncertainty bands for the nominal model are then computed by combining the nominal model and the uncertainty regions of the error model, as explained in Chapter 2. Fig. 3.18(b) shows the final symmetric uncertainty band.

■

In the Chapter 2, we also presented the NSSE method which is the other main stochastic solution to the robust identification problem. The NSSE approach has little relation to the Bayesian one since in NSSE the uncertainty is quantified by means a non-stationary stochastic process. However, the Bayesian approach can obtain uncertainty bands similar to the NSSE ones. See next example.

Example 3.12. NSSE example solved with competing models.

In this example, we consider again the plant and experiment of Example 2.5. The model uncertainty is quantified by using a set of three competing models. The models considered are the ones suggested in (Goodwn, Braslavsky, and Seron, 2002) but, instead to take these models separately we use them together to quantify the uncertainty. The three competing models are parameterized by the following functions:

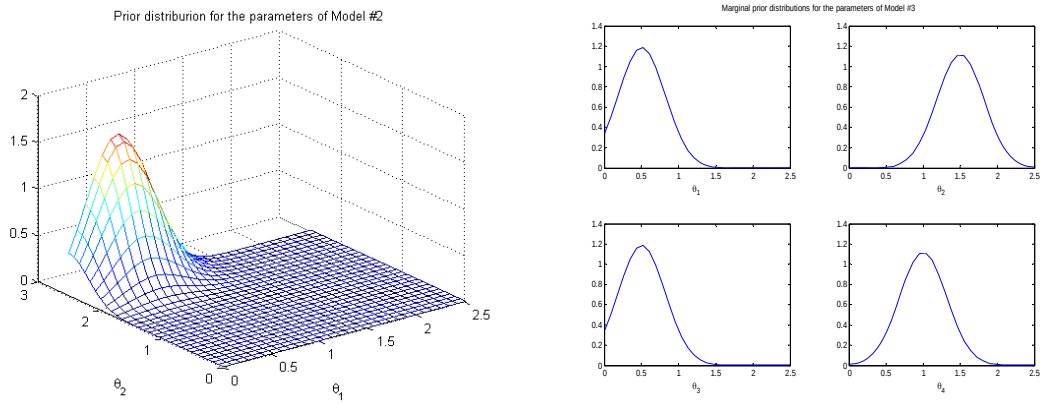
First model: $B_1(s) = \frac{3}{(0.5s+1)(5s+1)}$

Second model: $B_1(s) = \frac{1}{(0.5s+1)^2}$, $B_2(s) = \frac{1}{(3s+1)^2}$

Third model:

$$B_1(s) = \frac{1}{(0.5s+1)^2} \ , \ B_2(s) = \frac{1}{(3s+1)^2} \ , \ B_3(s) = \frac{1}{(0.5s+1)^3} \ , \ B_4(s) = \frac{1}{(3s+1)^3}$$

The prior distribution of the parameters of each model is assumed to be Gaussian $\mathcal{N}(\boldsymbol{\theta}_0, \lambda \mathbf{R}_0^{-1})$ with $\lambda = 1$. The mean value and precision matrix for each model is selected as: $\boldsymbol{\theta}_0^{(1)} = 1$, $\mathbf{R}_0^{(1)} = 10$, $\boldsymbol{\theta}_0^{(2)} = (0.5 \quad 2.5)^T$, $\mathbf{R}_0^{(2)} = 10 \cdot \mathbf{I}_{2\times 2}$, and $\boldsymbol{\theta}_0^{(3)} = (0.5 \quad 1.5 \quad 0.5 \quad 1)^T$, $\mathbf{R}_0^{(3)} = 10 \cdot \mathbf{I}_{4\times 4}$. The selected mean values are near the least squares estimates obtained from frequency data. We assign the same prior probability to each model, i.e., 1/3. Prior distributions for the second and third model are shown in Fig. 3.19.



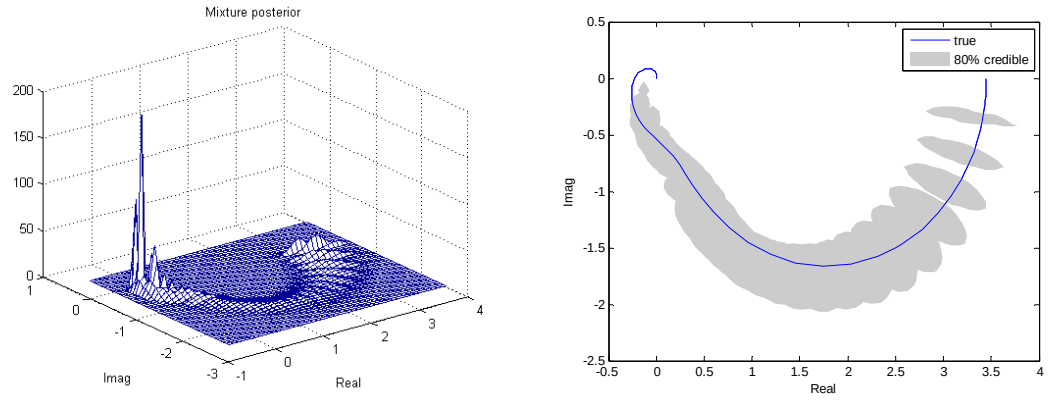
(a) **(b)**

Fig. 3.19. (a) Joint prior distribution for the parameters of Model 2, (b) Marginal prior distributions for the parameters of Model 3

Once collected the data and computed the integrated likelihoods, the posterior probabilities for each model are $p_{N1} = 0.2091$, $p_{N2} = 0.2098$, and $p_{N3} = 0.5812$. Posterior distributions in the Nyquist plane are combined with these probabilities to obtain the mixture posterior distributions of Fig. 3.20(a).

The final 80% posterior credible regions are shown in Fig. 3.20(b). The uncertainty band is tighter than the one obtained in Example 2.5 in which only Model 2 was considered and the uncertainty ellipses were obtained by assuming that the model error could be modeled by means a non-stationary stochastic process. Note also that the

credible regions at each frequency in general are not ellipses since they are obtained from mixture distributions.



(a) **(b)**

Fig. 3.20. (a) Mixture posterior distribution, (b) Final uncertainty band

■

3.3.4 Other features of the Bayesian approach

a. Effect of the prior distribution in the uncertainty size

One of the advantages of the Bayesian approach is that it is possible to reduce the uncertainty bands obtained from the likelihood function by means the adequate selection of the model prior distribution $p(G)$.

In the Gaussian case, this reduction can be attained by simply increasing the value of the prior precision matrix, $\mathbf{R}_0$. Next example illustrates this effect.

Example 3.13. Effect of the prior distribution in the variance error reduction

Consider the plant and experiment of the Example 2.5 (Goodwin, Braslavsky, and Seron, 2002). Regarding the prior information, we assume that the noise is zero mean Gaussian with variance $\lambda = 1$. And we assume that the nominal model is a second order model parameterized by the functions $B_1 = \frac{1}{(0.5s+1)^2}$ and $B_2 = \frac{1}{(3s+1)^2}$.

The prior distribution for the parameter vector $\boldsymbol{\theta}$ is assumed to be Gaussian with mean value $\boldsymbol{\theta_0} = (0.77 \quad 2.37)^T$, which is the least squares estimate obtained from frequency domain data. A selection of the prior precision matrix of $\mathbf{R}_0 = 100 \cdot \mathbf{I}_{2\times 2}$ leads to a spikier Gaussian bell than a selection of $\mathbf{R}_0 = \mathbf{I}_{2\times 2}$, therefore it leads to smaller prior and posterior uncertainty ellipses. See Fig. 3.21.

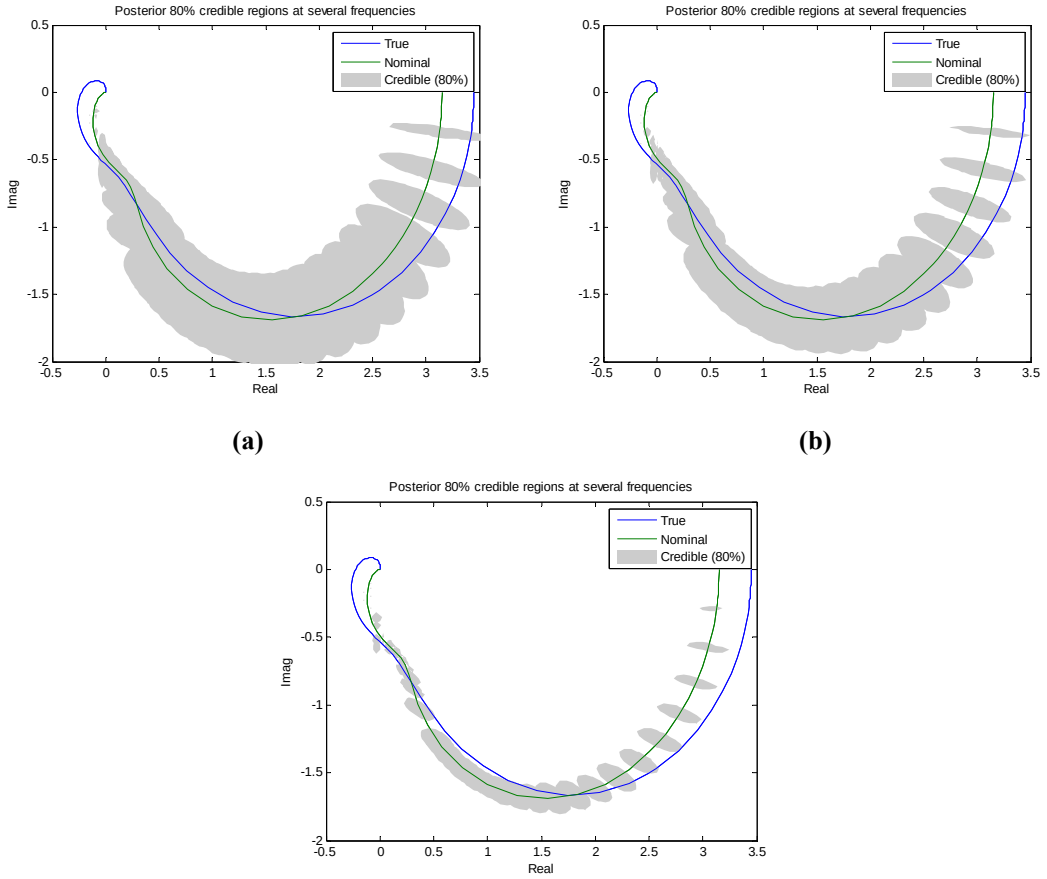


Fig. 3.21. Posterior 80% credible regions for (a) $\mathbf{R}_0 = \mathbf{I}_{2\times 2}$, (b) $\mathbf{R}_0 = 10 \cdot \mathbf{I}_{2\times 2}$, (c) $\mathbf{R}_0 = 100 \cdot \mathbf{I}_{2\times 2}$

Note also that the ellipses are around the nominal model. In fact these ellipses are quantifying only the variance error and not the bias error since we have considered only one model, i.e., we are considering only parametric uncertainty. However, this example shows how the bias/variance trade-off presented in Chapter 2 can be overcome if we use the Bayesian approach to tighten the variance error.

■

b. Resonant systems

Unlike the NSSE method, the Bayesian procedure does not distinguish plants with real poles from plants with resonant poles. It deals in the same way with all types of plants.

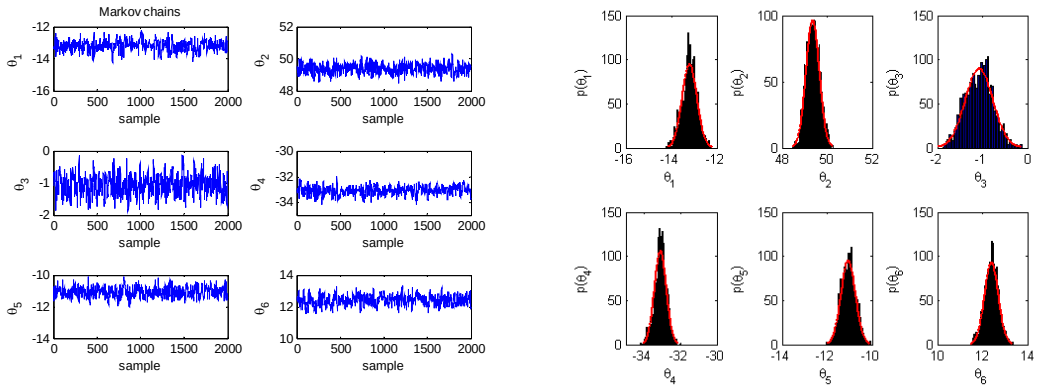
Example 3.14. Resonant plant and Markov Chain Monte Carlo (MCMC) implementation

Consider the resonant plant and experiment of the Example 2.6. The functions that parameterize the nominal model are $B_1(s) = \frac{1}{0.5s+1}$, $B_2(s) = \frac{1}{(0.5s+1)^2}$, $B_3(s) =$

$\frac{1}{(2s+1)^2}$, $B_4(s) = \frac{1}{(0.5s+1)^3}$, $B_5(s) = \frac{1}{(0.5s+1)^3}$ and $B_6(s) = \frac{1}{(2s+1)^3}$. Hence the model has 6 parameters to be identified.

The prior distribution for the parameter vector $\boldsymbol{\theta}$ is assumed to be Gaussian with mean value $\boldsymbol{\theta_0} = (-13.17 \quad 49.38 \quad -1.09 \quad -33.11 \quad -11.06 \quad 12.39)^T$, which is the least squares estimate obtained from frequency domain data, and prior precision matrix of $\mathbf{R}_0 = 10 \cdot \mathbf{I}_{6\times 6}$.

Even though the distribution is Gaussian, the high number of parameters (6) makes necessary the use of simulation techniques such as the Markov Chain Monte Carlo simulations presented in a previous section. Fig. 3.22(a) shows the Markov chains (2000 samples) obtained for each one of the prior parameters and Fig. 3.22(b) shows the resulting prior marginal distributions.



(a) **(b)**

Fig. 3.22. (a) Markov chains, (b) Simulated prior marginal distributions

Next figure shows the final 80% posterior credible regions. For this selection of the precision matrix the resulting uncertainty bands are tighter than the ones obtained by the NSSE method using both the random walk and the integrated random walk.

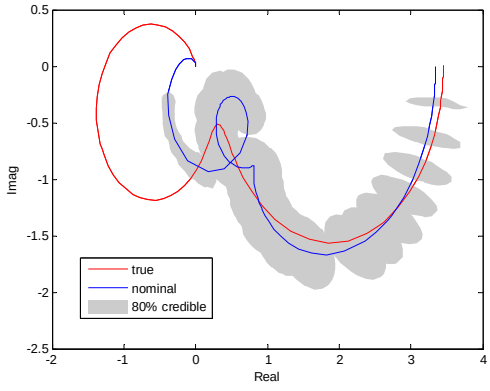


Fig. 3.23. Posterior 80% credible regions for $\mathbf{R}_0 = 10 \cdot \mathbf{I}_{7\times 7}$

■

3.4 Application of the Bayesian Decision Theory

It turns out that the Bayesian framework is highly unifying since it can consider many different aspects that constitute the robust identification problem. In the present section, we explore the connections between the robust identification problem and the Bayesian Decision Theory.

Three problems can be considered: the selection of the nominal model, the model (in)validation, and the optimal design of experiments.

3.4.1 Selection of a nominal model

When the application is the design of robust controllers, we need to select a nominal model G_0 from the credible model set. There exist several choices:

One possibility is to select the model corresponding to the maximum value of the posterior distribution $p(G|\mathbf{y})$.

$$\hat{G}(\mathbf{y}) = \arg\max_{G} p(G|\mathbf{y})$$

This is the *maximum a posteriori* (MAP) estimate.

Another possibility is to find the nominal model that minimises the Bayesian risk.

$$\hat{G}(\mathbf{y}) = \arg\min_{G} L(G(\mathbf{y}), G)$$

This is the minimum risk estimate (MR).

Still, a third possibility is to select the nominal model that minimises the maximal loss,

$$\hat{G}(\mathbf{y}) = \arg\min_{G(\mathbf{y})} \max_{G} L(G(\mathbf{y}), G)$$

This is a minimax (MML) approach.

a. Maximum a posteriori (MAP) nominal models

The optimal MAP nominal model is the one that maximises the posterior probability of the model conditioned on the observations $p(G|\mathbf{y})$. MAP estimation is sometimes called *unconditional* maximum likelihood (ML) estimation; and ML estimation is sometimes called *conditional* ML estimation.

Maximum a posteriori models do not require the definition of a loss function $L(G_{true}, G)$ quantifying the cost of selecting the value G for the nominal if the true model is G_{true}. Instead, it is supposed that the estimate G is in the neighbourhood of G_{true}, and hence a hypothetical loss function would be small.

The MAP nominal model can be estimated in the parameter space (case of parametric uncertainty) or in the Nyquist plane (case of dynamic uncertainty). For the case of credible regions in the Nyquist plane, it is possible to obtain the MAP estimate for each one of the set value distributions. Since the resulting nominal order will be equal to the number of value sets (number of exciting frequencies) one can apply Hankel norm model reduction techniques to produce a restricted complexity nominal model. This approach is widely used in deterministic methods, see for instance the works of (Milanese and Taragna, 2002) and (Malan *et al*., 2001).

Example 3.15. MAP nominal model

Consider the example of (Ninness and Henriksen, 2010). The data generating process (true plant) is $y_n = \frac{0.2}{q-0.8} u_n + v_n$, $n = 1..N$, so the true parameter vector is $\boldsymbol{\theta}^T = (\theta_1, \theta_2) = (0.2, -0.8)$.

The experiment consists of only $N = 20$ samples of the excitation signal $\{u_n\}_{n=0}^{N-1}$, where $u_n = \sin(n)$. The measurement noise sequence $\{v_n\}_{n=0}^{N-1}$ is i.i.d. uniform with zero mean and variance $E[v_n^2] = 0.01$.

Regarding the model prior information, since the plant is stable we know that the parameter in the denominator is such that $|\theta_2| \leq 1$, so we assume that the marginal distribution $p_{\theta_2}(\theta_2)$ is uniform between -1 and 1. And, since the gain is positive, we assume that the parameter in the numerator is $\theta_1 > 0$, and so we take the marginal distribution $p_{\theta_1}(\theta_1)$ uniform between 0 and 1. As θ_1 and θ_2 are independent, we can construct the joint distribution by simply taking $p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = p_{\theta_1}(\theta_1)\, p_{\theta_2}(\theta_2)$. See Fig. 3.24(a).

And regarding the prior noise information, even though we *know* that the noise is uniform, it is more convenient to assume it is Gaussian since, this way, the likelihood function (and the posterior model distribution) will present a unique maximum value. In this example we have assumed zero mean Gaussian noise with standard deviation 0.5.

The maximum value of the joint model posterior distribution corresponds to the model $(\hat{\theta}_1, \hat{\theta}_2) = (0.1975, -0.8051)$. This is the MAP estimate. See Fig. 3.24(b). The precision of the estimate depends on the grid used as a support for the probability distributions. In this example, we have used a linear grid of 80 values between -1.2 and 1.2 for both parameters.

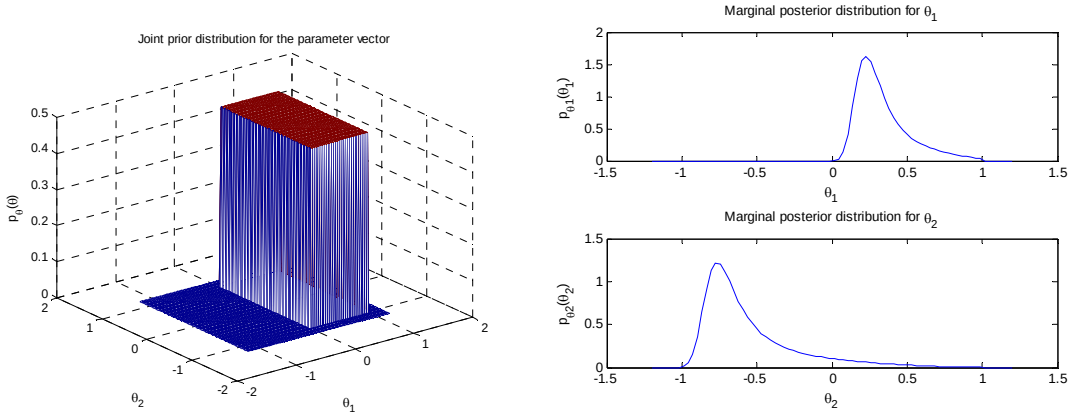


Fig. 3.24. (a) Joint model prior distribution, (b) Marginal model posterior distributions

■

b. Minimum risk (MR) nominal models

Minimum risk estimate is especially interesting since it allows introducing, in the modelling process, possible (quantitative) knowledge about the cost of a wrong estimate.

This cost may be identification-oriented or control-oriented. In the first case the aim is to minimise the identification error while in the second case the robust control relevancy can be evidenced by defining cost functions in terms of the robustness theorems, i.e., the stability robustness and the performance robustness specifications.

Selection of loss functions: The selection of suitable loss functions is important since each one produce a different estimate. For instance, if the quadratic loss

$$L\left(\hat{\boldsymbol{\theta}}(y), \boldsymbol{\theta}\right) = k\left(\hat{\boldsymbol{\theta}}(y) - \boldsymbol{\theta}\right)^2 \tag{62}$$

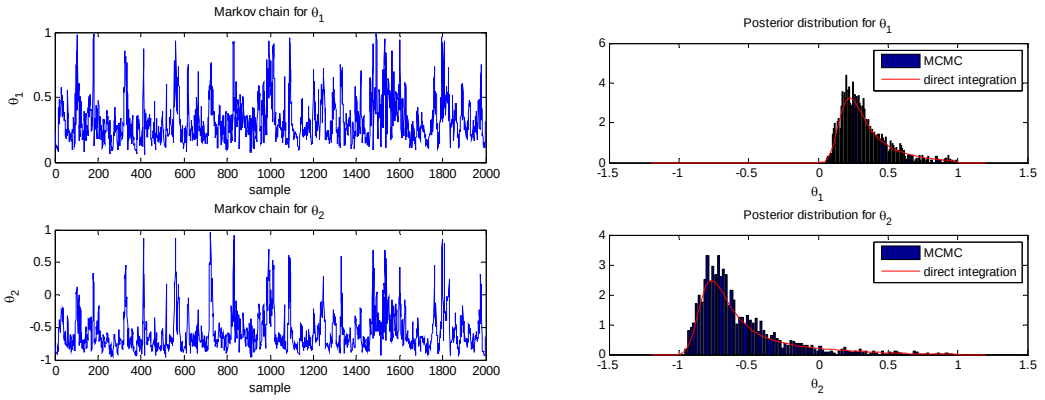
is selected then the minimum risk estimator is the posterior *mean*, and if the absolute value loss

$$L\left(\hat{\boldsymbol{\theta}}(y), \boldsymbol{\theta}\right) = k\left|\hat{\boldsymbol{\theta}}(y) - \boldsymbol{\theta}\right| \tag{63}$$

is selected, then the minimum risk estimator is the posterior *median*.

Example 3.16. MR nominal model

Consider again the plant and experiment of Example 3.15. Here we have obtained the model posterior probability distributions by means a 2000 points MCMC simulation. See Fig. 3.25.



(a) **(b)**

Fig. 3.25. (a) Markov chains, (b) Marginal model posterior distributions

The MR estimate for the quadratic loss defined in (62) with k=1 is the mean value $(\hat{\theta}_1, \hat{\theta}_2) = (0.3249, -0.5670)$ while the MR estimate for the absolute loss defined in (63) with k=1 is the median $(\hat{\theta}_1, \hat{\theta}_2) = (0.2853, -0.6657)$.

■

c. Case of the Bayesian Credible Model Set defined in a model space

In the case the probability distributions are defined on the model space, suitable measures for model distance must be used in order to define a loss function. For instance, weighted $\mathcal{L}_2$ norms on the transfer function space are commonly used to measure distance between operators, therefore they can be used as loss functions. These measures have the advantage that they are directly related to the identification error. Also, by the application of Fatou's Lemma (McVinnish, 2006) they can be related to the central estimate of the model set in set-membership techniques. Other norms may better reflect the ultimate objective of robust control design. It is the case of the infinity norm and ν-gap metric, see e.g. (Hildebrand and Gevers, 2003) and (Hjalmarsson, 2005).

Dealing with this kind of loss functions is similar than dealing with the standard parametric ones, since one can use MCMC simulations in order to obtain samples of the posterior loss. The reversible jump MCMC algorithm (see Appendix C) can generate a sample of operators $\{G_i\}_{i=1}^{M}$ that can be used in the calculation of the posterior expected loss by using sample averages

$$\frac{1}{M}\sum_{i=1}^{M} L(G_i, G_0)$$

The posterior expected loss can then be minimised by standard numerical optimisation techniques. For other forms of loss and nominal models with non-linear parameters, the posterior expected loss can be minimised numerically, for example with the Nelder-Mead simplex algorithm (McVinish *et al*., 2006).

3.4.2 Optimal experiment design

Experiment design is an important issue in robust identification since nominal models and uncertainty regions are obtained from measurement input/output data.

The objective of optimal experiment design is to determine the less costly identification experiment that delivers *sufficient* and *meaningful* information about the system dynamics for the design of a robust controller or for a fault detection procedure. The following is a motivating example.

Example 3.17. Selection of the excitation signal

Consider again the plant of Example 3.15. Fig. 3.26 shows the resulting Feasible Parameter Set (FPS) regions assuming uniform noise of variance 0.01. In both cases the measurement data length is N=20 but, in Fig. 3.26(a) the excitation is a period of a square signal and in Fig. 3.26(b) the excitation is a period of a sinusoid.

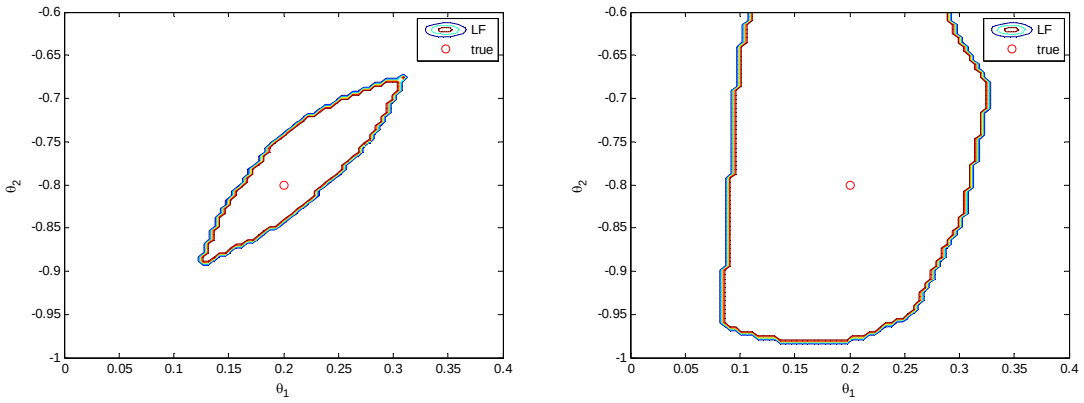


Fig. 3.26. (a) Square signal excitation, (b) sinusoid excitation

As expected, since the square signal is richer than the sinusoid, the size of the FPS is smaller in the first case, i.e., the uncertainty region is smaller.

■

In robust identification, the experiment must be designed to reduce the uncertainty region. This way, the resulting controllers will not be over-conservative and, in the fault detection procedures, we will reduce the risk of undetected faults.

The Bayesian framework allows considering the problem of the experiment design from a decision theoretic point of view. Let us point out some important concepts.

a. Utility function

A good way to design an experiment is to specify a utility function U reflecting the purpose of the experiment. Thus, one can regard the experiment design as a decision problem and can take the design that maximizes the expected utility.

The utility function depends on the decision d, the unknown parameters to be identified $\boldsymbol{\theta}$, the experiment η, and the measurement data $\mathbf{y}$, $U = U(d, \boldsymbol{\theta}, \eta, \mathbf{y})$.

In the Bayesian framework, the expected utility for the best decision d is given by:

$$U(\eta) = \int_{\mathbb{R}^N} \left(\max_{d \in \mathcal{D}} \int_{\mathbb{R}^d} U(d, \boldsymbol{\theta}, \eta, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}, \eta) p(\mathbf{y}|\eta) d\boldsymbol{\theta} \right) d\mathbf{y} \tag{64}$$

where $p(\cdot)$ denotes the probability density function. The Bayesian solution is provided by the η^* that maximises the expected utility

$$U(\eta^*) = \max_{\eta} U(\eta) \tag{65}$$

Informative experiment: In order to design informative experiments, it is reasonable to take as utility function the expected gain in Shannon information given by such an experiment.

Choosing a design that maximizes the expected gain in Shannon information is equivalent to choose a design that maximizes the expected Kullback-Leibler distance between the posterior and the prior distributions:

$$\int \ln \frac{p(\boldsymbol{\theta}|\mathbf{y}, \eta)}{p(\boldsymbol{\theta})} p(\mathbf{y}, \boldsymbol{\theta}|\eta) d\boldsymbol{\theta} d\mathbf{y} \tag{66}$$

since the prior distribution $p(\boldsymbol{\theta})$ is not a function of η, the η that maximizes the expected gain in Shannon information is the one that maximizes:

$$U(\eta) = \int \ln[p(\boldsymbol{\theta}|y, \eta)] \cdot p(y, \boldsymbol{\theta}|\eta) d\boldsymbol{\theta} d\mathbf{y} \tag{67}$$

b. Alphabetical optimality criteria

Different design criteria define the so-called Bayes A, C, D, E and G optimality. For a through explanation, see (Chaloner and Verdinelli, 1995).

Bayes D-optimality arises when we want to perform model discrimination and parameter estimation. For the case of linear regression models where the output is corrupted by additive i.i.d. Gaussian noise with known variance λ, the expected utility (design criterion function) is

$$U(\eta) = -\frac{d}{2}\ln(2\pi) - \frac{d}{2} + \frac{1}{2}\ln|\lambda(\mathbf{R}_N + \mathbf{R}_0)| \quad \textbf{(68)}$$

where $\mathbf{R}_N$ and $\mathbf{R}_0$ are the precision matrices defined in Chapter 3. Therefore, to maximize $U(\eta)$ is equivalent to maximize $|\lambda(\mathbf{R}_N + \mathbf{R}_0)|$.

A characteristic of optimal Bayesian designs is the dependence on the sample size *N,* since $\mathbf{R}_N = \mathbf{\Phi}^T\mathbf{\Phi}$. If N is large enough, there is no differences between a Bayesian design (where the prior knowledge on the parameters variance is introduced through $\mathbf{R}_0$) and its corresponding non Bayesian one, since, in this case $(\mathbf{R}_N + \mathbf{R}_0) \approx \mathbf{R}_N$. That is for large data records, the data dominates while for short data records the prior dominates, as we have seen previously in many examples.

3.4.3 Model validation

a. Bayesian hypothesis test

Model (in)validation can be viewed as a hypothesis testing problem. Suppose that we want to infer if a given model G belongs to the credible model set $\mathcal{B}$ or not. In a hypothesis testing problem, this is equivalent to define the two hypotheses "G belongs to the model set" and "G does not belong to the model set" and reject one and accept the other.

In the Bayesian framework the inference is based on the posterior model distribution, $p(G|\mathbf{y})$. This distribution is used to calculate which one of the corresponding null hypothesis H_0 and alternative hypothesis H_1 is true. Consider, for instance, the parametric case. These posterior probabilities are given as

$$\Pr(H_0 \text{ is true}|\mathbf{y}) = \Pr(\boldsymbol{\theta} \in \mathcal{B}|\mathbf{y})$$

and

$$\Pr(H_1 \text{ is true}|\mathbf{y}) = \Pr(\boldsymbol{\theta} \in \mathcal{B}^c|\mathbf{y})$$

These probabilities are not meaningful in a classical viewpoint, since it considers $\boldsymbol{\theta}$ to be a fixed number. Consequently, a hypothesis is either true of false, and the probabilities are 1 or 0. No intermediate values are possible. If $\boldsymbol{\theta} \in \mathcal{B}$, $\Pr(H_0 \text{ is true}|\mathbf{y}) = 1$ and $\Pr(H_1 \text{ is true}|\mathbf{y}) = 0$ for all values of $\mathbf{y}$. If $\boldsymbol{\theta} \in \mathcal{B}^c$, these values are reversed.

In a Bayesian formulation of a hypothesis testing problem, these probabilities depend on the sample $\mathbf{y}$ and can give useful information about the veracity of H_0 and H_1. The implementation of the Bayesian hypothesis test is performed by mean the use of the so-called Bayes factors.

b. Bayes factors

The Bayesian choice allows entering a prior guess about if the model at hand belongs or not to the model set. That is, we can assign to hypothesis H_0 and H_1 a prior probability. Selecting $\Pr(H_0) = \Pr(H_1) = 0.5$ indicates an unprejudiced starting point. To derive the Bayes factors, we apply Bayes' theorem to obtain

$$\Pr(H_k|\mathbf{y}) = \frac{\Pr(\mathbf{y}|H_k)\Pr(H_k)}{\Pr(\mathbf{y}|H_0)\Pr(H_0) + \Pr(\mathbf{y}|H_1)\Pr(H_1)} \quad , \quad k = 0,1$$

so that

$$\frac{\Pr(H_0|\mathbf{y})}{\Pr(H_1|\mathbf{y})} = \frac{\Pr(\mathbf{y}|H_0)}{\Pr(\mathbf{y}|H_1)} \cdot \frac{\Pr(H_0)}{\Pr(H_1)}$$

where the factor $B_{01} \equiv \frac{\Pr(\mathbf{y}|H_0)}{\Pr(\mathbf{y}|H_1)}$ is called the Bayes factor. Thus, in words, we have,

posterior odds = Bayes factor × prior odds

In the simplest case, when the two hypotheses are single distributions with no free parameters, B_{01} is the likelihood ratio (see Appendix A).

The application of the Bayes factor can be interpreted in terms of the so-called Occam's window (Hoeting *et al*., 1999). The Occam's window corresponds to the values of Bayes factor between O_L and O_H. Usual selections are O_L=1/20, O_H=1 and O_L=1/20, O_H=20.

Suppose that we want to validate the model G_0. The null hypothesis H_0 is "G_0 belongs to the model set" and the alternative hypothesis H_1 is "G_0 does not belong to model set". If there is evidence for H_0 then H_1 is rejected, but rejecting H_0 requires strong evidence for the H_1. If the evidence is inconclusive (falling in Occam's window) neither hypothesis is rejected.

3.5 Summary and conclusion

We have proposed a methodology to formulate and solve the robust identification problem in a probabilistic –Bayesian- framework. The methodology relies in the definition of a Bayesian credible model set to support both *a priori* information and *a posteriori* information. The BCMS is inspired in the FMS of SMI deterministic methods. Definitions for the BCMS in the parameter space and frequency domain have been derived.

The model uncertainty is described by means of credible regions. Credible regions are easier to compute than classical confidence regions and they enjoy some desirable properties compared to confidence regions. Credible regions may lead to smaller uncertainty regions (provided the adequate selection of the prior distributions), they can combine hard bounds with soft bounds, they can be computed iteratively (as new measurements are available), and they can be disjoint.

In the case of parametric uncertainty, the exact results for the case of linear regression models and Gaussian distributions have been presented. If the distributions are not Gaussian or the number of parameters increases, simulation techniques such as Markov Chain Monte Carlo techniques must be used to compute the posterior marginal distributions and the credible regions. Compared to the existing robust identification deterministic methods, it has been shown that the Feasible Parameter Set (FPS) can be obtained by means our methodology if the uniform distribution is used to model the measurement noise.

In the case of uncertainty regions in the frequency domain, we have illustrated the use of frequency domain data in the BCMS. In order to describe the bias error we have considered sets of competing models which lead to mixture (thus, non-ellipsoidal) credible regions in the Nyquist plane. The law of total probability is used to derive the credible regions and to compute the posterior probability for each model in the set of competing models. Exact posterior credible regions are presented for the case of linear regression models and Gaussian probability distributions. Compared to the existing methods, the same probabilistic regions of conventional system identification and Model Error Modeling can be obtained if no model prior distribution is used (i.e., by using only the likelihood function). In all the cases, the uncertainty regions can be tightened provided the adequate selection of the model prior distribution. In particular, it is illustrated how the variance error is reduced by selecting larger values for the prior precision matrix. Compared to the Non Stochastic Stationary Embedding, smaller uncertainty regions have been obtained and with no need to modify the methodology for the case of resonant systems.

Finally, three related problems have been presented and discussed under the viewpoint of the Bayesian Decision theory: the selection of the nominal model, the model (in)validation, and the optimal design of experiments.

CHAPTER 4

Application to Fault Detection

This chapter presents some results to illustrate the application of the Bayesian identification approach to fault detection. Two case studies are considered: a quadruple tank process and a three-bladed wind turbine.

4.1 Fault detection based on feasible parameter regions

Since in this chapter we are going to perform the fault detection on the basis of feasible parameter regions, in this section we present the background of this approach.

4.1.1 Background

Model parameterization: Let us assume that the system can be expressed by the following model

$$y(k) = F(k, \boldsymbol{\theta}) + e(k)\ , k = 1, \dots, N \qquad \textbf{(69)}$$

where the function $F(n, \boldsymbol{\theta}) = \hat{y}(k, \boldsymbol{\theta})$ can be linear or nonlinear and it can contain any function of the inputs $u(k)$ and outputs $y(k)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_o$ is the d dimension parameter vector which belongs to a set $\boldsymbol{\Theta}_o$, defined by the *a priori* bounds for the parameter values. And, finally, $e(k)$ is the additive error bounded by a constant $|e(k)| \leq \delta$.

Feasible parameter set (FPS): According to (Milanese *et al*., 1996) the parameter estimation problem consists of determining the parameter set that contains all the models consistent with the set of N input/output data. As explained in Chapter 2, the resulting feasible parameter set is defined as

$$\text{FPS} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_o \mid \ y(k) - \delta \leq F(k, \boldsymbol{\theta}) \leq y(k) + \delta \ , \ \ k = 1, \dots, N\} \qquad \textbf{(70)}$$

In the fault detection field, in order to avoid dealing with the exact description of the FPS, existing algorithms usually approximate the FPS by using inner/outer simpler shapes such as boxes, parallelotopes, ellipsoids or zonotopes (Vicino and Zappa, 1996), (Reppa and Tzes, 2011), and (Alamo, Bravo and Camacho, 2005). The approximated set is called Approximated Feasible Parameter Set (AFPS).

There exist inner and outer approximations. Inner approximations find the parameter set of maximum volume such that all the parameters of the AFPS are inside the FPS. Hence, for the k-th measurement we have $\text{AFPS}_k \subseteq \text{FPS}_k$. On the other hand, outer approximation algorithms find the parameter set of minimum volume that guarantees that the FPS is inside the AFPS, $\text{FPS}_k \subseteq \text{AFPS}_k$.

Recursive algorithms allow computing inner and outer approximations as follows

$$\begin{aligned} \text{A}_{in}\text{FPS}_{k+1} &\subseteq \text{A}_{in}\text{FPS}_k \cap \text{S}_k \\ \text{A}_{out}\text{FPS}_{k+1} &\supseteq \text{A}_{out}\text{FPS}_k \cap \text{S}_k \end{aligned}$$

where S_k is the region in the parameter space that contains all the parameters consistent with the measurement k and the function $F(k, \boldsymbol{\theta})$,

$$\text{S}_k = \left\{\boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{d}} \mid \ -\delta \leq y(k) - F(k, \boldsymbol{\theta}) \leq \delta\right\} \qquad \textbf{(71)}$$

Linear case: In the linear case, $F(k, \boldsymbol{\theta})$ can be expressed as a linear regression, $F(k, \boldsymbol{\theta}) = \boldsymbol{\varphi}(k)^T\boldsymbol{\theta}$, where the regression vector $\boldsymbol{\varphi}(k)^T$ can contain any function of inputs $u(k)$ and outputs $y(k)$. Here the set S_k is a strip and the FPS is a polytope that can be described in the H-polytope form (Blesa, Puig, and Saludes, 2013) as

$$\text{FPS} = \left\{\boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{d}} \mid \mathbf{A} \leq \boldsymbol{\theta}\mathbf{b}\right\} \qquad \textbf{(72)}$$

with $\mathbf{A} = (-\boldsymbol{\varphi}(1)^T, \boldsymbol{\varphi}(1)^T, \dots, -\boldsymbol{\varphi}(N)^T, \boldsymbol{\varphi}(N)^T)^T$ and $\mathbf{b} = (-y(1) + \delta, y(1) + \delta \ , \dots, -y(N) + \delta, y(N) + \delta)^T$.

Nonlinear case: Unfortunately, for the nonlinear case the optimization problem is nonconvex and obtaining a suitable solution is computationally hard. In (Milanese *et al.,* 1996) a minimum outer box is determined by means of a set of optimization problems. Alternatively, in (Jaulin *et al*., 2010) the FPS is approximated by using subpavings and the SIVIA (Set Inversion Via Interval Analysis) algorithm that is based on refining the initial *a priori* set $\boldsymbol{\Theta}_o$ by iteratively bisecting it.

Fault detection: Once the FPS has been estimated from non-faulty data, the fault detection test consists in checking the consistency of new measurements with the former FPS. The consistency is checked by means of the intersection of S_k (set of parameters consistent with data at instant k) with the FPS. A fault will be indicated if this intersection leads to an empty set

$$S_k \cap \text{FPS} = \emptyset \tag{73}$$

In the linear case, the fault detection test (73) can be solved easily, but in the nonlinear case, inner or outer approximations of this intersection must be used and missed alarms (in outer approximations) and false alarms (in inner approximations) may appear. For this reason, outer approximations are used rather than inner approximations for fault detection purposes (Blesa, Puig, and Saludes, 2011b).

4.1.2 Bayesian approach

Feasible parameter set estimation: As explained in Chapter 3, the same FPS region of the set-membership approach (70) can be obtained within our Bayesian methodology. Since the region defined in (70) describes parametric-type uncertainty, the Bayesian credible model set reduces to its parametric version,

$$\mathcal{B}_{\boldsymbol{\theta}} \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}:\ p(\boldsymbol{\theta}|\mathbf{y}) \geq c(\alpha)\} \tag{74}$$

where the process model is characterized by means of the parameter vector $\boldsymbol{\theta}$, and the model posterior probability is $p(\boldsymbol{\theta}|\mathbf{y}) \propto p_e(\mathbf{y}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})$. Now we have to decide which is the model prior probability distribution, $p(\boldsymbol{\theta})$. In the Bayesian framework this probability is a *subjective* probability. Here it is assumed that we have no information about which the value of the "true" parameter vector $\boldsymbol{\theta}$ will be and consequently we take a flat $p(\boldsymbol{\theta})$. This way the model posterior distribution is directly proportional to the likelihood function of the observations, $p(\boldsymbol{\theta}|\mathbf{y}) \propto p_e(\mathbf{y}|\boldsymbol{\theta})$.

The likelihood of the observations coincides in form with the noise probability distribution, i.e., $p_e(\mathbf{y}|\boldsymbol{\theta}, \sigma) \equiv p(\mathbf{y} - \hat{\mathbf{y}}|\sigma)$, where σ is a parameter that characterizes the noise and hence the error term.

Since we want to obtain a hard-bounded uncertainty/credible region, we select σ to be the additive error bound of the set-membership technique presented in the previous section and thus we assume that the additive error is uniform distributed, $e \sim \mathcal{U}(-\delta, \delta)$. Since $p_e(e)$ is uniform, the resulting likelihood function is constant and nonzero in the region where models (parameters) are consistent with measurements and it is zero outside this region.

The likelihood function can be numerically obtained for a grid of candidate parameter vectors $\boldsymbol{\theta}_i$ by assuming that the error samples $e(k) = y(k) - \hat{y}(k)$, where $\hat{y}(k) = F(k, \boldsymbol{\theta}_i)$, are i.i.d. (independent identically distributed)

$$p_e(\mathbf{y}|\boldsymbol{\theta}_i,\sigma)=\prod_{k=1}^{N} p_e(y(k)-\hat{y}(k)|\boldsymbol{\theta}_i,\sigma) \tag{75}$$

It is noteworthy that there is no difference in the computation of the likelihood function either in the nonlinear or linear case.

Fault detection: Once we have calibrated the model (i.e, obtained the likelihood function $p_e(\mathbf{y}|\boldsymbol{\theta}_i,\sigma)$ for all the points $\boldsymbol{\theta}_i$ in the parameter grid) the fault detection test can be carried out, for every new measurement $y(k)$, by computing the new likelihood function $p_e(y(k)-\hat{y}(k)|\boldsymbol{\theta}_i,\sigma)$ and verifying if there is at least one parameter vector $\boldsymbol{\theta}_j$ for which both $p_e(\mathbf{y}|\boldsymbol{\theta}_j,\sigma)$ and $p_e(y(k)-\hat{y}(k)|\boldsymbol{\theta}_j,\sigma)$ are nonzero. If this parameter (or set of parameters) exists we conclude that the new measurement is consistent with the feasible parameter set. The consistency can be checked by simply multiplying both likelihood functions for each parameter $\boldsymbol{\theta}_i$ in the grid. If the product is equal to zero for all the parameters in the grid,

$$p_e(\mathbf{y}|\boldsymbol{\theta}_i,\sigma)\cdot p_e(y(k)-\hat{y}(k)|\boldsymbol{\theta}_i,\sigma)=0 \quad \forall i \tag{76}$$

we decide that a fault has taken place since the new measurement is not consistent with the feasible parameter region.

Of course, the ability to detect "small" faults depends on the grid density. A denser grid will be able to detect smaller deviations of the parameter vector. This implies a more computationally intense calibration stage. However, the fault detection stage is not so intensive computationally since it can consider one sample at once. This feature also allows the on-line implementation of the method.

Let us illustrate the performance of this methodology by means of two case studies.

4.2 Case Study I: Quadruple tank process

4.2.1 Physical model

Fig. 4.1 shows the quadruple tank-process proposed as a benchmark problem by (Johansson, 2000). The process inputs are the voltages to the pumps v_1 and v_2. The process outputs are the tank levels h_1, h_2, h_3 and h_4.

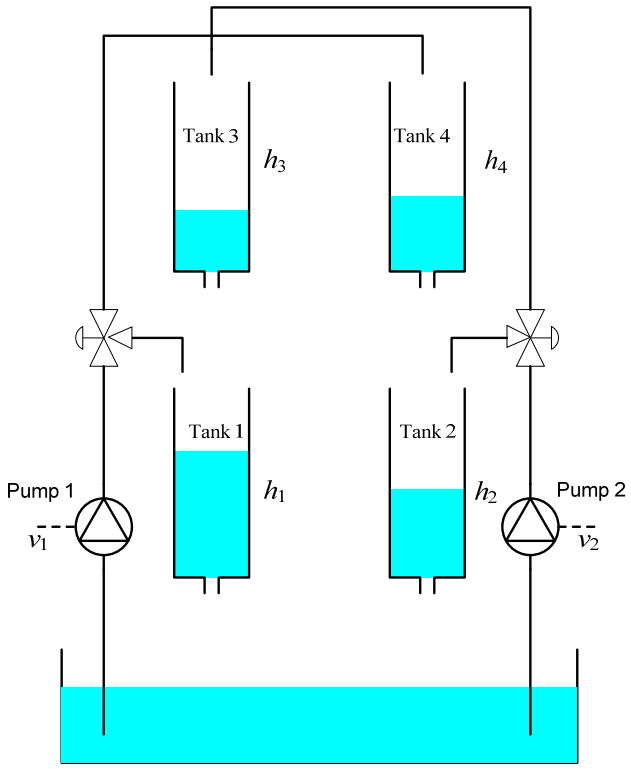


Fig. 4.1. The quadruple-tank process

The equations that describe the dynamical behavior of the system are obtained by means of the mass balances and the Bernouilli's law:

$$\begin{aligned}
\frac{dh_1}{dt} &= -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}v_1 \\
\frac{dh_2}{dt} &= -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2}v_2 \\
\frac{dh_3}{dt} &= -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}v_2 \\
\frac{dh_4}{dt} &= -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4}v_1
\end{aligned} \tag{77}$$

where A_i is the cross-section of tank i, a_i is the cross-section of the outlet hole of tank i, and $k_1 v_1$ and $k_2 v_2$ are the corresponding flows of pumps 1 and 2. The parameters $\gamma_1, \gamma_2 \in (0,1)$ are determined from how the valves are set prior to the experiment. The gravity acceleration is denoted as g.

The initial conditions are $h_1(0) = 12.4cm$, $h_2(0) = 12.7cm$, $h_3(0) = 1.8cm$, $h_4(0) = 1.4cm$, $v_1(0) = 3V$ and $v_2(0) = 3V$.

The operation range is assumed to be $h_1 \in [2,\ 11]$ cm and $h_3 \in [1,\ 15]$ cm.

Table 4.1 shows the values of the plant parameters.

A_1, A_3	28cm^2
A_2, A_4	32cm^2
a_1, a_3	0.071cm^2
a_2, a_4	0.057cm^2
γ_1	0.7
γ_2	0.6
k_1	3.33cm^3/Vs
k_2	3.35cm^3/Vs
g	981cm/s^2

Table 4.1. Parameter values

4.2.2 MISO case

Firstly we consider the MISO (Multi Input Single Output) case. In this case, the plant output is the level of tank 1, h_1, while the inputs are the level of tank 3, h_3, and the first pump voltage, v_1. The uncertain parameters will be a_1 and a_3.

The fault detection procedure consists of two steps. In the first step, *calibration*, a fault-free scenario is used to generate the data needed to determine the uncertainty region for the parameters a_1 and a_3. In the second step, *fault detection*, data containing faults are generated and the former uncertainty region is used to detect them.

a. Calibration in a fault-free scenario

A set of N=140 measurements has been obtained in a fault-free scenario (see Fig. 4.2). These data will be used to calibrate the model, i.e., to obtain the uncertainty region for a_1 and a_3 in the parameter space.

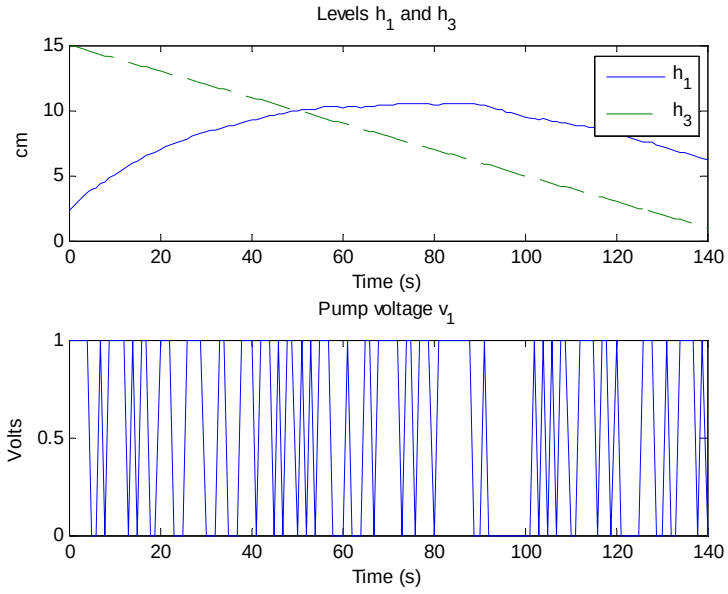


Fig. 4.2. Fault-free scenario data

Discrete-time linear regression model: The nominal model will be a discrete-time linearized version of the first equation in (77). To discretize we use the Euler method with sampling time $T_s = 1s$.

$$\dot{h}_1 \approx \frac{h_1(k) - h_1(k-1)}{T_s} \tag{78}$$

Thus,

$$h_1(k) = h_1(k-1) - \frac{a_1}{A_1}\sqrt{2gh_1(k-1)} + \frac{a_3}{A_1}\sqrt{2gh_3(k-1)} + \frac{\gamma_1 k_1}{A_1}v_1(k-1) + e(k)$$

where $e(k)$ is the additive error (it includes sensor and discretization error) and it is assumed to be bounded by a constant $|e(k)| \leq \delta$, $\delta = 0.05cm$.

The process output, expressed in the linear regression form, is

$$y(k) = h_1(k) = \boldsymbol{\varphi}^T(k) \cdot \boldsymbol{\theta} + e(k) \tag{79}$$

where $\boldsymbol{\varphi}^T(k) = \left(h_1(k-1) \quad -\frac{1}{A_1}\sqrt{2gh_1(k-1)} \quad \frac{1}{A_1}\sqrt{2gh_3(k-1)} \quad \frac{k_1}{A_1}v_1(k-1)\right)$ is the regression vector and $\boldsymbol{\theta} = (1 \quad a_1 \quad a_2 \quad \gamma_1)^T$ is the parameter vector.

Uncertainty region obtained by strips intersection: In this case, the Feasible Parameter Set (FPS) is obtained by intersecting all N strips defined by the pairs of parallel lines separated 2δ, $y(k) - \boldsymbol{\varphi}^T(k) \cdot \boldsymbol{\theta} = \pm\delta$. See Fig. 4.3.

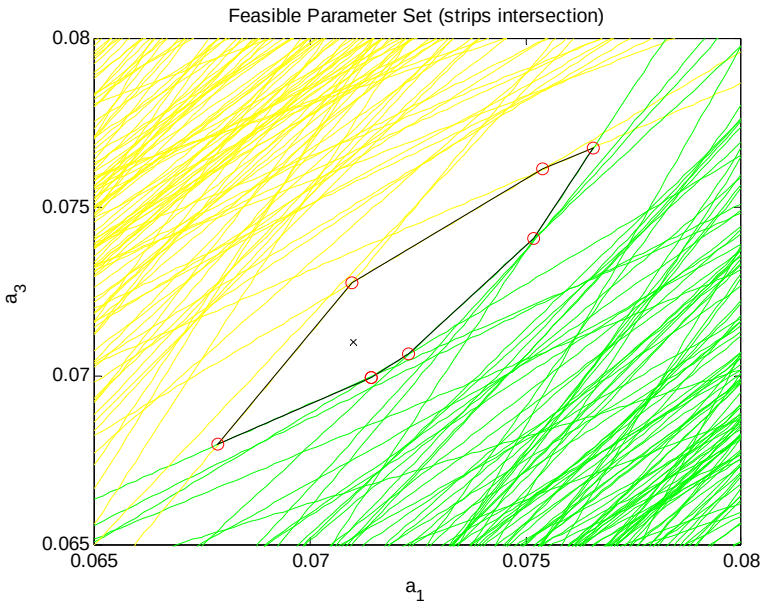


Fig. 4.3. FPS obtained by strips intersection (the red little circles indicate the final polytope vertices)

Uncertainty region obtained as the likelihood function contour: In this case the FPS region is obtained as the contour of the likelihood function assuming that the noise is uniform distributed as $\mathcal{U}(-\delta, \delta)$. Fig. 4.4 shows the result for a 60×60 parameters grid.

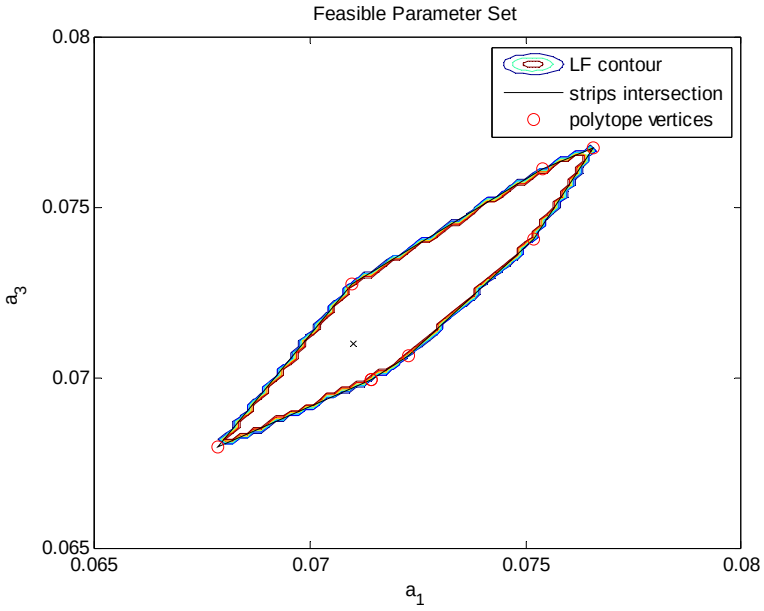


Fig. 4.4. FPS obtained as the likelihood contour

This FPS region coincides with the one obtained by intersecting the strips. Thus, in the linear case, the computation of the likelihood function does constitute an alternative to the strips intersection technique of the set-membership approach.

b. Fault detection stage

Generation of the faulty behavior: In order to show the fault detection behavior, different fault scenarios have been created by introducing faults when the system is under the operation point shown in Fig. 4.5.

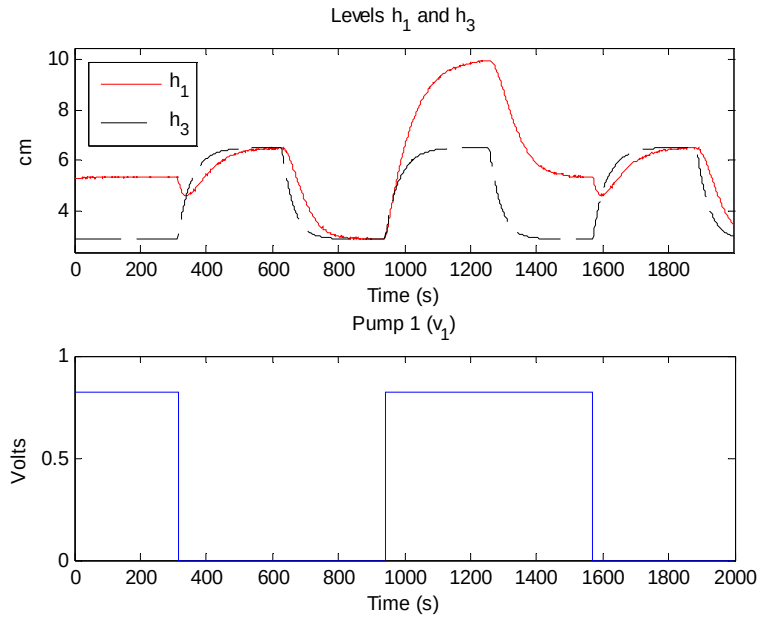


Fig. 4.5: Nonfaulty scenario

In particular, a fault has been introduced at sample 1201 consisting of an additive constant of value 0.035 acting over the parameter a_1.

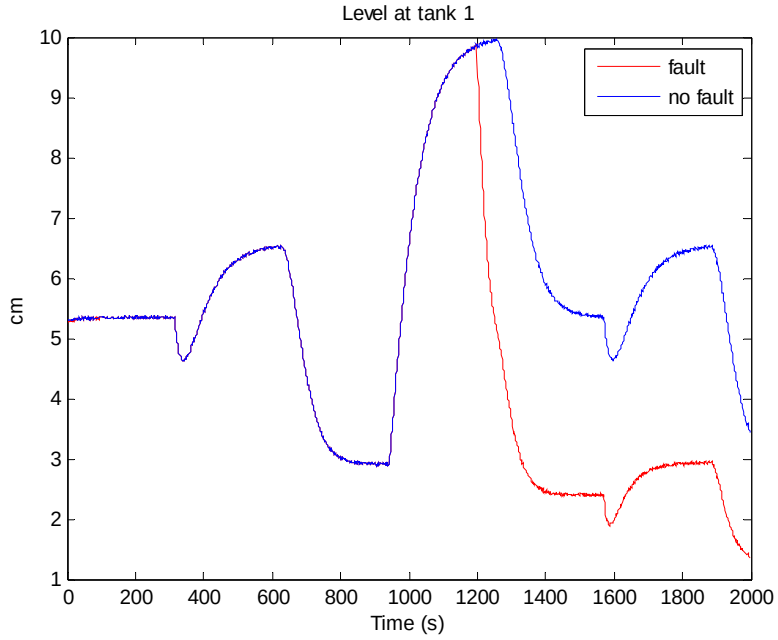


Fig. 4.6. Faulty scenario

Fault detection by means of the set-membership technique: In this technique, each new measurement is used to obtain a new strip in the parameter space and analyze its

consistence to the FPS. No fault is decided when the strip intersects or contains the FPS (see Fig. 4.7(a)), otherwise we decide a fault has taken place (see Fig. 4.7(b)).

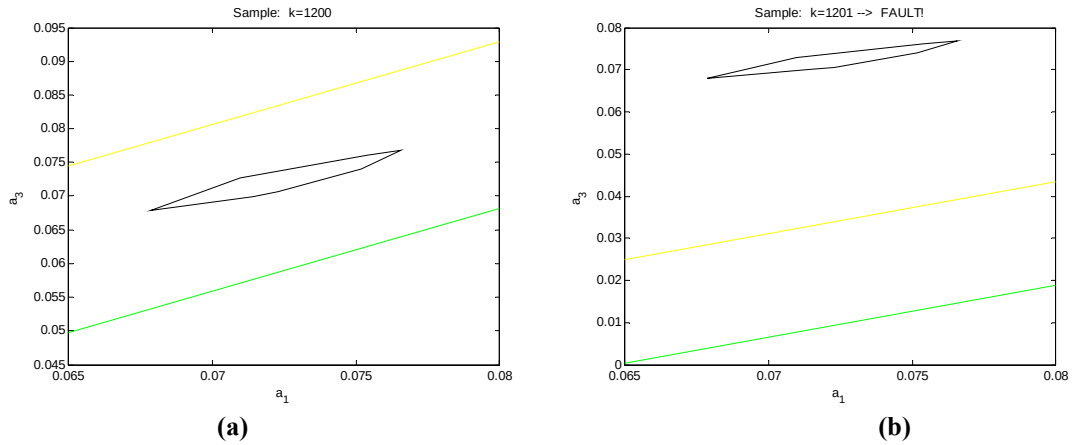


Fig. 4.7. (a) No fault detected, (b) Fault detected

Fault detection by means of the likelihood function: In the uniform case, the uncertainty region obtained in the fault-free scenario corresponds to the values in the parameter space grid where the likelihood function is nonzero. In the fault detection stage, we can use this region to test if new samples of the system are consistent with it or not.

When a new measurement enters, we compute the likelihood that every pair $(a_1,\ a_3)$ in the grid has generated it. If the new likelihood covers (totally or partially) the fault-free likelihood function, we conclude that data are consistent with the model and thus we decide that no fault has taken place (see Fig. 4.8).

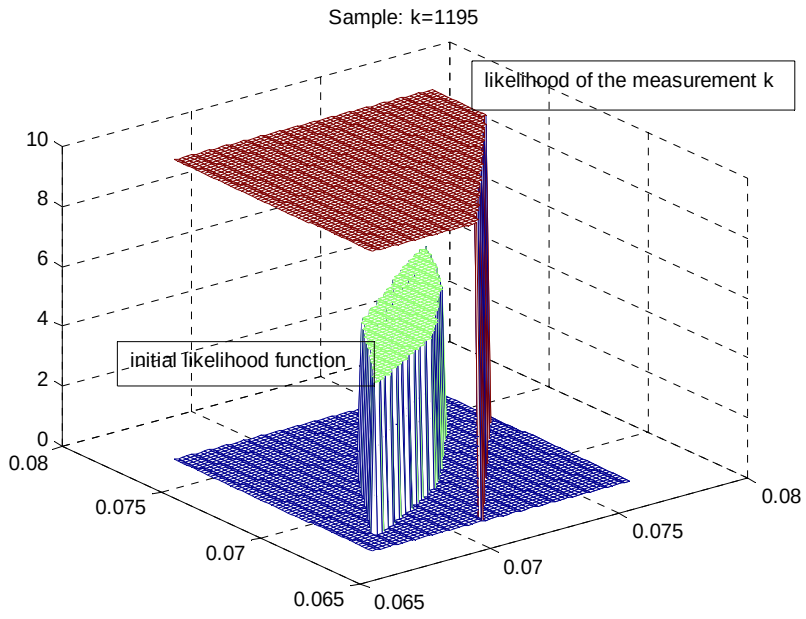


Fig. 4.8. No fault detected (measurement *k* is consistent with the uncertainty model). Remark: the likelihood functions z-values have been scaled for comparison purposes at 5 and 10 respectively.

In the case that a fault has taken place, the two likelihoods will be disjoint and thus their product will be zero for all the grid values. In such a case, we will decide that we have a fault (see Fig. 4.9).

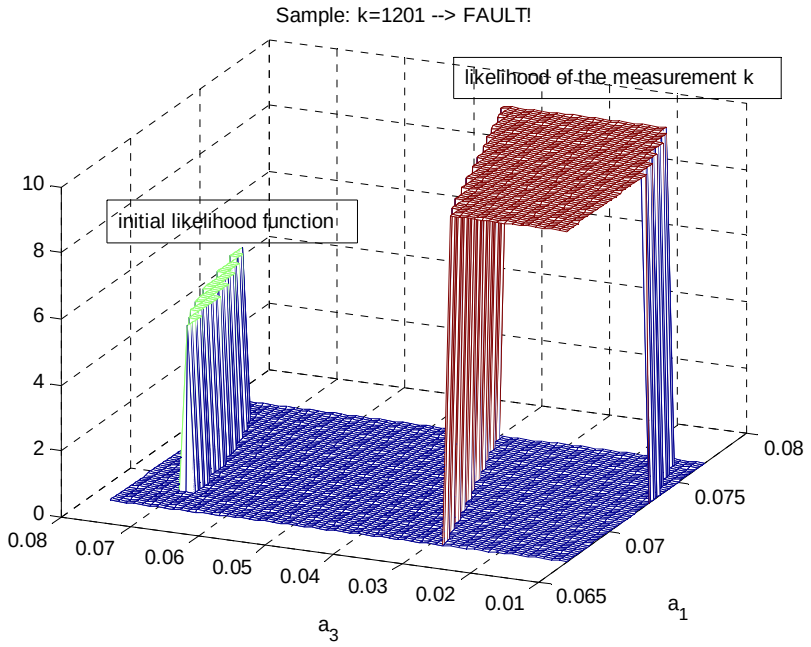


Fig. 4.9. Fault detected (measurement *k* is not consistent with the uncertainty model). Remark: the likelihood functions z-values have been scaled at 5 and 10 respectively.

This procedure has been tested, for a 60×60 parameters grid, with the same data than the set-membership case and it successfully has detected the fault at sample 1201. The elapsed time per sample is similar to the set-membership case. Note that, again, the results coincide with the set-membership case. Both methods are equivalent in the linear case (since the contour of the measurements likelihood coincides with the set-membership strips).

Minimum fault detected: Both procedures detect additive faults in parameter a_1 equal or greater than 0.0053cm. If the fault magnitude is 0.0052cm or smaller, since many few values of the uncertainty set are consistent with the data, the fault is not detected (the deviation of the behavior is considered due to the model uncertainty and not a fault).

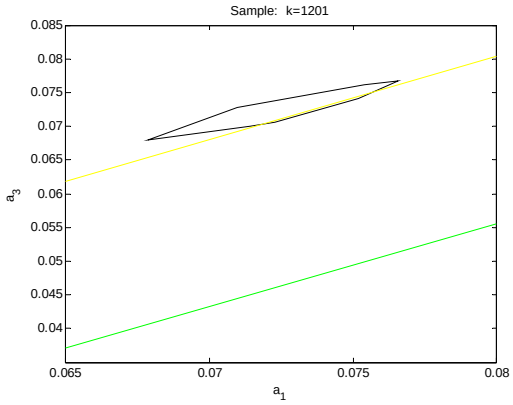
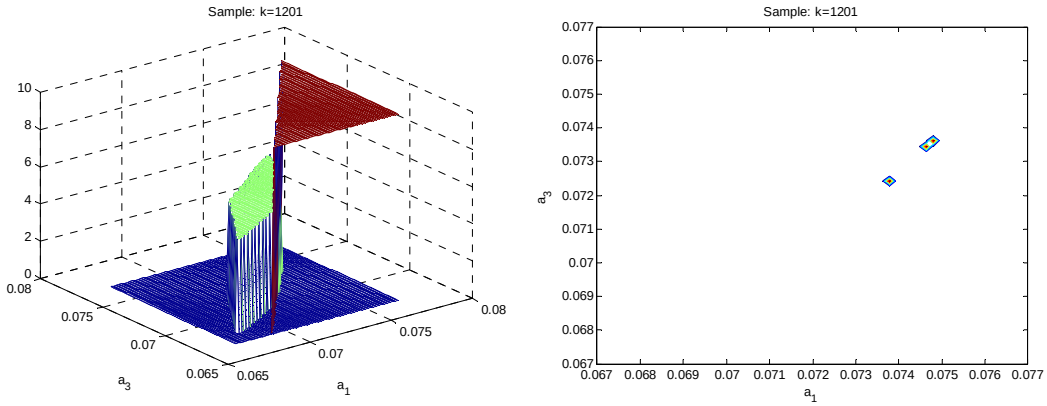


Fig. 4.10. Fault not detected in the set-membership technique



(a) **(b)**

Fig. 4.11. Fault not detected in the likelihood function technique. (a) free-fault likelihood and likelihood at sample k, (b) contour plot of the likelihood updated by sample k.

4.2.3 MISO case with observer

The use of diagnostic interval observers is reported in (Puig *et al*., 2008), (Raïssi *et al*., 2010). Observers improve the ability of detecting output faults but lead to structures nonlinear in the parameters. Set-membership techniques cannot deal in a simple manner with this type of systems, but the likelihood approach presented in this dissertation does.

Next figure shows the observer configuration for the MISO plant considered. It consists of a model of the plant with an additive correction term depending on the error between the measured output $h_1(n)$ and the predicted output $\hat{h}_1(n)$. L is the observer gain.

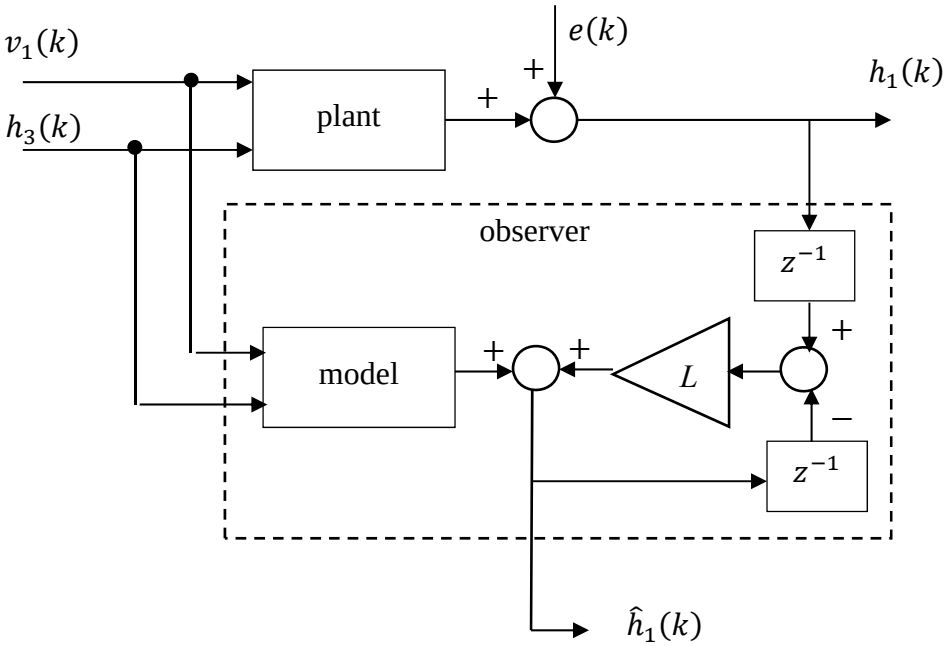


Fig. 4.12. MISO plant with output observer

The general expression is

$$\hat{h}_1(k) = \hat{h}_1(k-1) - \frac{a_1}{A_1}\sqrt{2g\hat{h}_1(k-1)} + \frac{a_3}{A_1}\sqrt{2gh_3(k-1)} + \frac{\gamma_1 k_1}{A_1}v_1(k-1) + e(k) + L\left(h_1(k-1) - \hat{h}_1(k-1)\right) \quad \textbf{(80)}$$

a. Calibration in a fault-free scenario

Uncertainty regions for several values of L have been obtained. **Fig. 4.13** shows the case for $L = 0.5$.

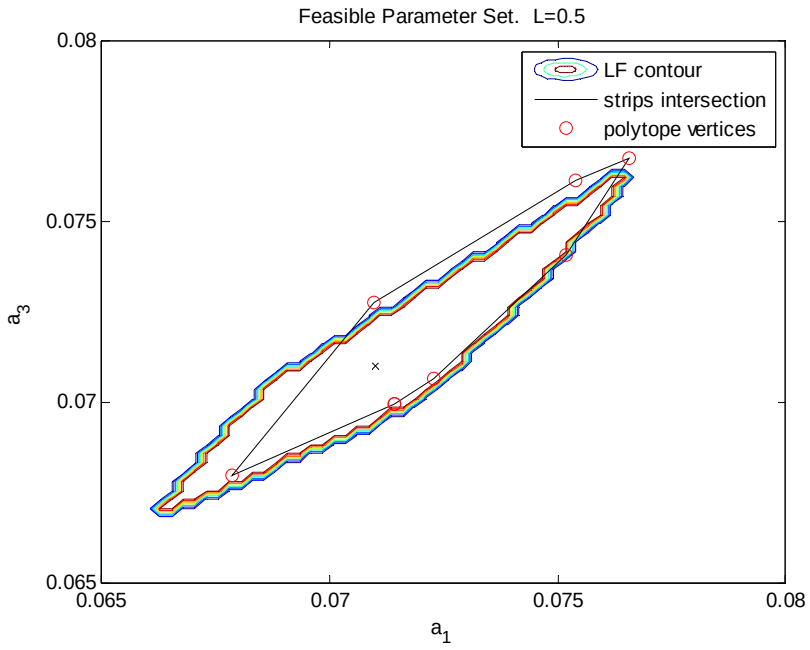


Fig. 4.13. Uncertainty region for the MISO case with output observer. $L = 0.5$

For small values of the observer gain, the uncertainty region is small (see **Fig. 4.14**). For the case $L = 0$, the system is equivalent to an Output Error (OE) model,

$$\hat{h}_1(k) = \hat{h}_1(k-1) - \frac{a_1}{A_1}\sqrt{2g\hat{h}_1(k-1)} + \frac{a_3}{A_1}\sqrt{2gh_3(k-1)} + \frac{\gamma_1 k_1}{A_1}v_1(k-1) + e(k) \quad \textbf{(81)}$$

and the observer output tracks the nonfaulty behavior.

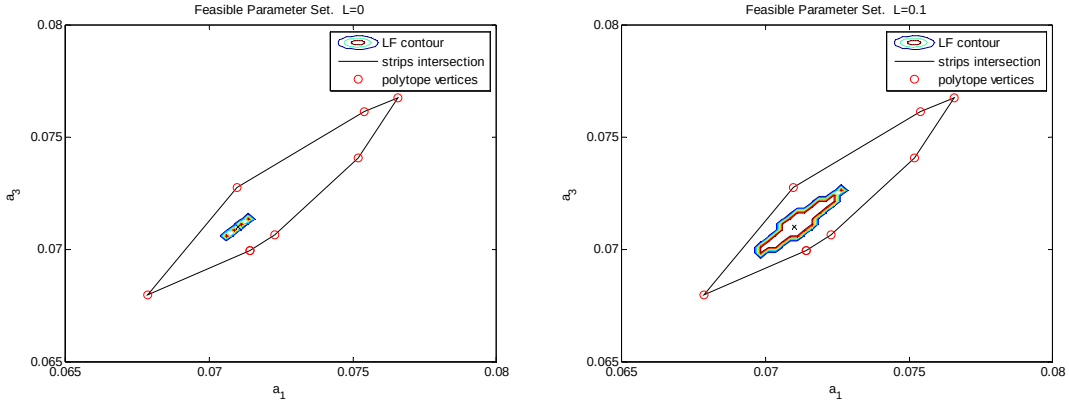
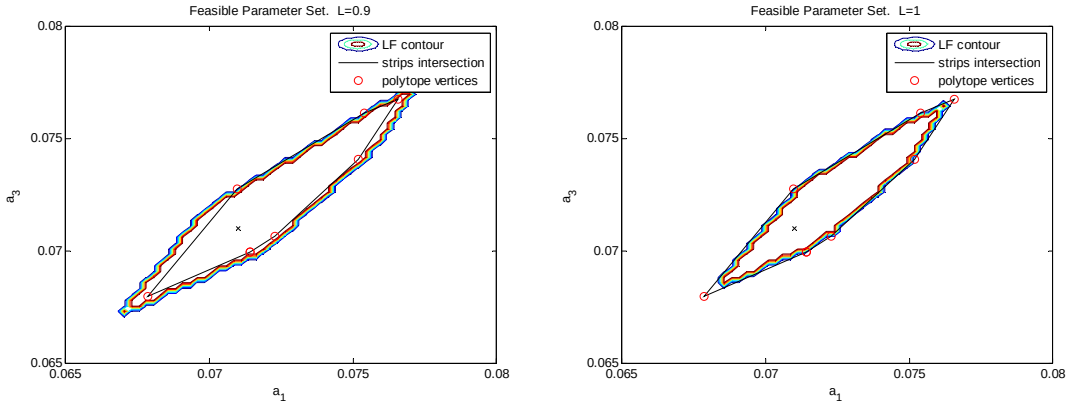


Fig. 4.14. Uncertainty region for the MISO case with output observer. $L = 0, 0.1$

For large values of the observer gain, the uncertainty region tends to the FPS region of the previous section (see Fig. 4.15). For the case $L = 1$, the system is equivalent to an Auto Regresive with eXogenous input (ARX) model,

$$\hat{h}_1(k) = -\frac{a_1}{A_1}\sqrt{2g\hat{h}_1(k-1)} + \frac{a_3}{A_1}\sqrt{2gh_3(k-1)} + \frac{\gamma_1 k_1}{A_1}v_1(k-1) + e(k) + h_1(k-1) \quad \textbf{(82)}$$

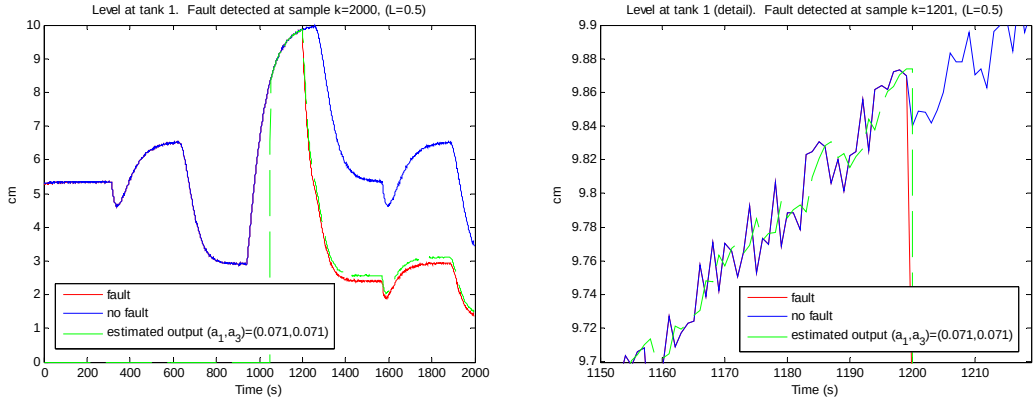


(a) **(b)**

Fig. 4.15. Uncertainty region for the MISO case with output observer. $L = 0.9, 1$

b. Fault detection stage

As in the previous section, the likelihood approach and the obtained uncertainty regions for a 60×60 parameters grid have been used to detect faults induced by changes in the parameter a_1. At the sample 1201, a value of 0.035 is added to this parameter. The observer is applied to the plant at sample 1050. The fault detection procedure has been implemented on-line and the fault has been detected at the correct sample in all the cases ($0 < L \leq 1$).



(a) **(b)**

Fig. 4. 16. Fault detection for the case $L = 0.5$. (a) The observer output for the case $(a_1, a_3) = (0.071, 0.071)$. (b) Detail of the faulty, nonfaulty and observer behavior.

Regarding the minimum detectable fault, the observer $L = 0.1$ can detect faults as small as 0.0011 (but with a delay of 63 samples), and the observers $L = 0.5$ and $L = 0.9$ can detect faults as small as 0.0035 and 0.0030 respectively (but with a delay of 528

samples, this latter is due to the fact that they detect the fault thanks to the change of the pump voltage)

Finally, Fig. 4.17 shows the smallest faults that the system can detect and the number of samples elapsed until the fault is detected. The good behavior of small values of L is explained here because we are using an ideal model (i.e., the data have been generated by this model). If the model was not exact, the behavior would not be so good for small observer gains.

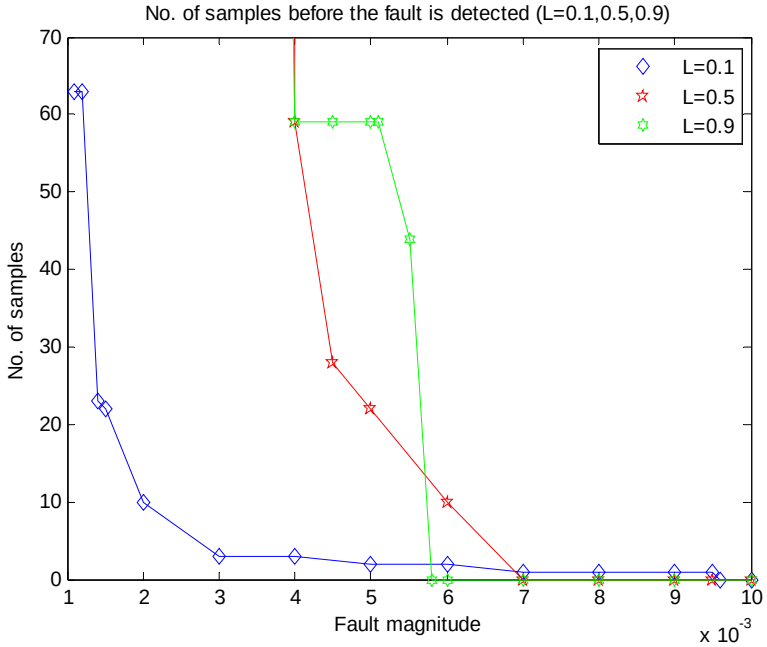


Fig. 4.17. Number of samples before the fault is detected

4.2.4 MIMO case

.

Now we consider the MIMO (Multi Input Multi Output) case. A set of 21000 measurement data have been obtained for the whole system. Fig. 4.18 shows the steady state final 1400 samples for each tank level. The first $N = 500$ samples of this record will be used for calibration purposes.

The system in (77) can be viewed as two independent MIMO systems. In the first one, the inputs are v_1 and v_2, and the outputs are h_1 and h_3. The uncertain parameters are, again, a_1 and a_3.

$$\frac{dh_1}{dt} = -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}v_1$$
$$\frac{dh_3}{dt} = -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}v_2 \qquad \textbf{(83)}$$

In the second one, the inputs are v_1 and v_2, and the outputs are h_2 and h_4. The uncertain parameters are a_2 and a_4.

$$\begin{aligned}\frac{dh_2}{dt} &= -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2}v_2 \\ \frac{dh_4}{dt} &= -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4}v_1\end{aligned} \tag{84}$$

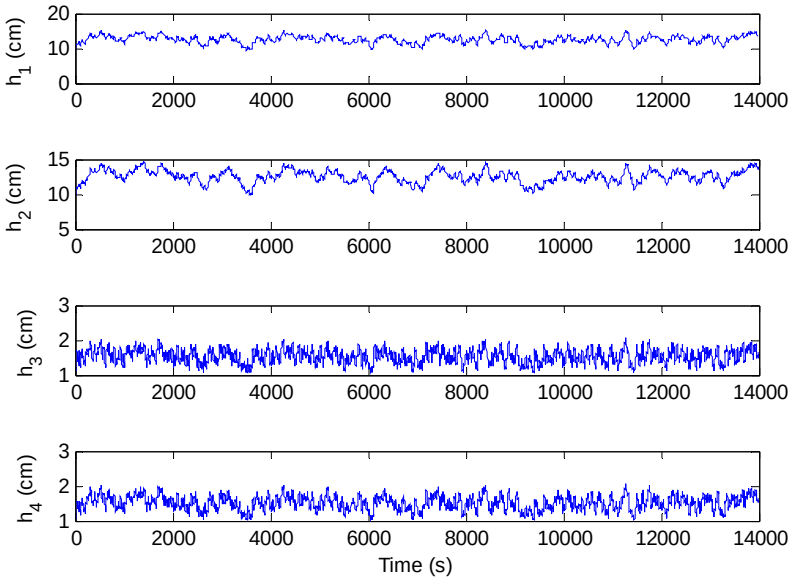


Fig. 4.18. Measurement data for the MIMO case

The identified error bounds are $\delta_1 = 0.1134$, $\delta_2 = 0.1098$, $\delta_3 = 0.1036$, and $\delta_4 = 0.1024$.

c. Set-membership approach

Firstly, we obtain the uncertainty region for the parameters a_1 and a_3 by considering the constraint h_1 (see Fig. 4.19(a)) and then the uncertainty region for the parameters a_1 and a_3 by considering the constraint h_3 (see Fig. 4.19(b)).

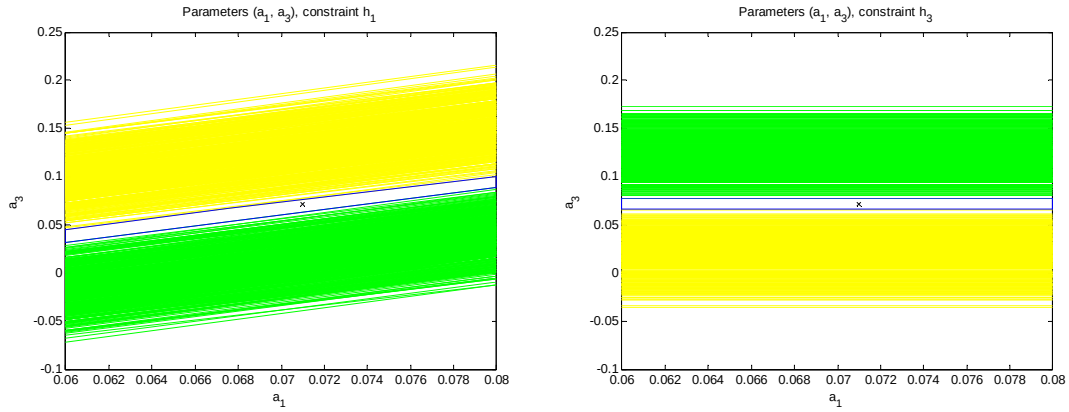


Fig. 4.19. MIMO case. Uncertainty region for a_1 and a_3 considering constraints (a) h_1 and (b) h_3

The combination of the previous regions leads to the uncertainty region shown in Fig. 4.20(a). Fig. 4.20(b) shows the resulting region for the parameters a_2 and a_4 and constraints h_2 and h_4.

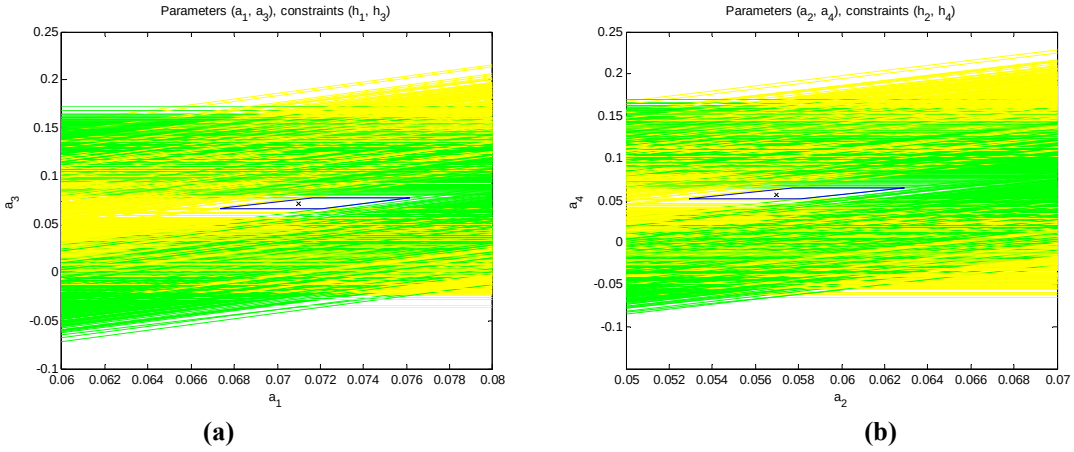


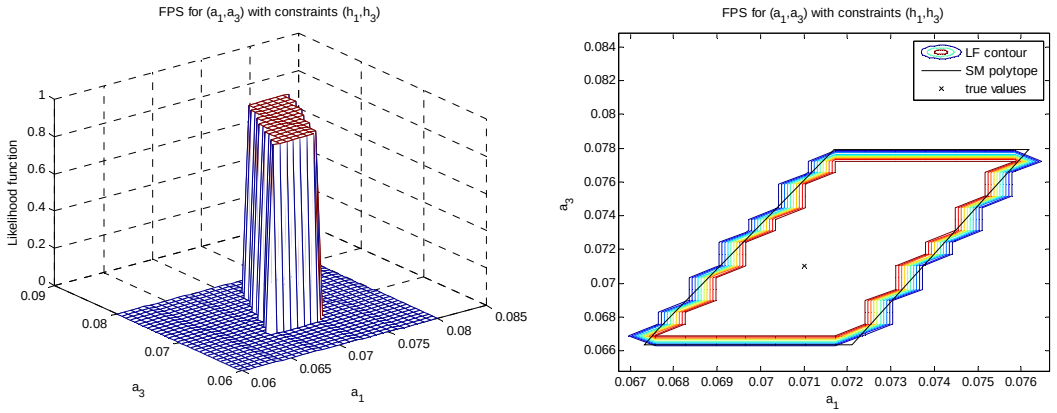
Fig. 4.20. MIMO case: (a) Final uncertainty region for a_1 and a_3 (b) Final uncertainty region for a_2 and a_4

d. Likelihood approach

The same region shown in Fig. 4.20(a) can be obtained by computing the likelihood to obtain the measurements $h_1,\ h_3$ for each pair of parameters $a_1,\ a_3$,

$$p(h_1, h_3|a_1, a_3) = \prod_{n=0}^{N-1} p_1(h_1 - \hat{h}_1|a_1, a_3)p_3(h_3 - \hat{h}_3|a_1, a_3) \quad \textbf{(85)}$$

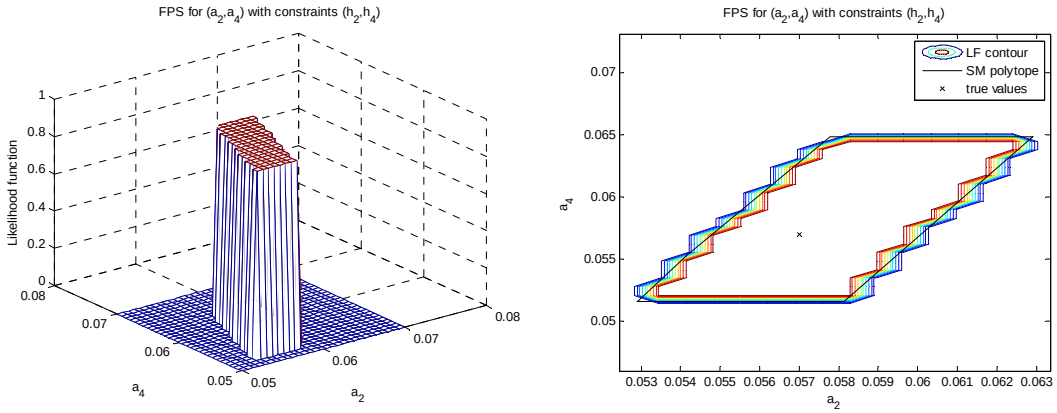
taking a 30×30 parameters grid, and considering uniform probability distributions for the residuals, $(h_1 - \hat{h}_1|a_1, a_3) \sim \mathcal{U}(-\delta_1, \delta_1)$ and $(h_3 - \hat{h}_3|a_1, a_3) \sim \mathcal{U}(-\delta_3, \delta_3)$. Fig. 4.21 shows the results for $N = 500$ and a grid of 30×30 values for $a_1,\ a_3$ between 0.06 and 0.08.



(a) **(b)**

Fig. 4.21. MIMO case, parameters a_1, a_3. (a) normalized likelihood function, (b) likelihood function contour plot.

Similar results are obtained for each pair of parameters a_2, a_4, by computing the likelihood to obtain the measurements h_2, h_4 (see Fig. 4.22).



(a) **(b)**

Fig. 4.22. MIMO case, parameters a_2, a_4. (a) normalized likelihood function, (b) likelihood function contour plot.

4.3 Case Study II: Wind turbine

In this section we consider the generic three-bladed horizontal variable speed wind turbine with a full converter coupling that was proposed by (Odgaard, Stoustrup, and Kinnaert, 2009) as a fault detection benchmark. Recently, a second benchmark based on the same plant has been proposed by (Odgaard and Johnson, 2012). We will focus on three faults of this second challenge, one for each blade, in order to illustrate the application of the methodology developed in this dissertation.

4.3.1 Physical model

A Simulink-based model of the wind turbine is available at (kk-electronic, 2012). It corresponds to a 5MW turbine with a hub height of 89.6m, rotor radius of 63m, rated rotor speed of 12.1rpm, and maximum pitch rate limited to 8deg/s. See (Odgaard and Johnson, 2012) for more details.

The wind turbine dynamics are implemented by means of a FAST (Fatigue, Aerodynamics, Structures, and Turbulence) code. FAST is an aeroelastic wind turbine simulator designed by the U.S. National Renewable Energy Laboratory's (NREL) National Wind Technology Center.

Sensor models, actuator models and faults are implemented in Simulink, making no changes in the underlying FAST code.

Actuator model: The benchmark presents several actuators: for the pitch, for the torque, and for the yaw systems. Here we will focus on the hydraulic pitch actuator model. This is a piston servo system which can be modeled as a closed loop transfer function between the pitch angle β and its reference β_r. A good approximation is a second order transfer function

$$H(s) = \frac{\beta(s)}{\beta_r(s)} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{86}$$

where ζ is the damping factor and ω_n is the natural frequency. In a fault-free scenario we consider that the all three systems are equal and their nominal values are $\zeta = 0.6$ and $\omega_n = 11.11\, rad/s$. In addition, the pitch angle is restricted to be within $\beta \in [-2°, 90°]$ and the pitch rate is restricted to $\dot{\beta} \in [-8°/s, 8°/s]$.

Sensor model: The pitch angles β_i, $i = 1,2,3$ are provided by FAST. In order to simulate the measurement noise and the effect of the electrical noise, signals from Band Limited White Noise blocks with a noise power of $1.5 \cdot 10^{-3}$ are added to the pitch angles generated by FAST.

Discrete model: The nominal model in (86) can be discretized by means of several methods. If we choose the forward approximation of the derivative $\dot{\beta} \approx \frac{\beta(k+1)-\beta(k)}{T_s}$ (which is equivalent to substitute $s = \frac{z-1}{T_s}$ in (86)), the resulting transfer function is

$$H(z) = \frac{\omega_n^2 T_s^2}{z^2 + [-2 + 2\zeta\omega_n T_s]z + [1 - 2\zeta\omega_n T_s + \omega_n^2 T_s^2]} \tag{87}$$

where T_s is the sampling time. The backward approximation ($\dot{\beta} \approx \frac{\beta(k)-\beta(k-1)}{T_s}$, $s = \frac{z-1}{zT_s} = \frac{1-z^{-1}}{T_s}$),

$$H(z) = \frac{\omega_n^2 T_s^2}{[1 + 2\zeta\omega_n T_s + \omega_n^2 T_s^2] + [-2 - 2\zeta\omega_n T_s]z^{-1} + z^{-2}} \quad \textbf{(88)}$$

and bilinear transform (Tustin transform, $s = \frac{2}{T_s}\frac{z-1}{z+1}$),

$$H(z) = \frac{\omega_n^2 T_s^2 z^2 + 2\omega_n^2 T_s^2 z + \omega_n^2 T_s^2}{z^2[4 + 4\zeta\omega_n T_s + \omega_n^2 T_s^2] + z[-8 + 2\omega_n^2 T_s^2] + [4 - 4\zeta\omega_n T_s + \omega_n^2 T_s^2]} \quad \textbf{(89)}$$

may be used instead.

The sampling time is chosen as 80 samples per second, i.e., $T_s = \frac{1}{80} = 0.0125s$.

It is important to note that, in all the three discrete models, the relationship between the two model parameters ζ and ω_n is nonlinear. Therefore, linear system identification techniques such as the strips set-membership technique considered in the previous case study cannot be used.

Next sections show how the uncertainty region identification and fault detection can be successfully performed by means of the Bayesian methodology developed in this dissertation.

4.3.2 First blade. Sensor fault

Fault description: The fault considered here is the Fault #4 of the benchmark. It consists of Blade 1 having a stuck pitch angle sensor, which holds a constant value of 1 deg. Fault #4 is active from 185s to 210s (i.e., from samples 14800 to 16800).

The requirement is that this fault must be detected in less than ten samples, that is, the detection time must be $T_D < 10T_s$.

a. Calibration in a fault-free scenario

The Simulink model provided by the benchmark has been used to generate a record of $N = 50.000$ samples in a fault-free scenario.

Error bound: The fault-free samples have been compared to the nominal model response to the same reference signal in order to obtain an estimate of the error bound. Since the model implemented is Simulink is the nominal model (86), the resulting error is only due to the measurement noise and to the discretization method. The maximum values obtained for $|e|$ have been 1.4048 for the forward approximation, 1.4184 for the Tustin approximation, and 1.4551 for the backward approximation. Note that, although

the Tustin transform is the method that best approximates the analogic behavior, the error bound is greater than the one obtained by the forward approximation. We conclude that the effect of the measurement noise is much more important than the discretization error. Hence, the discretization method is not a relevant issue here and we will use the forward approximation from now on.

Uncertainty region: The feasible parameter set has been obtained for a parameters grid of 40×40 as the contour of the likelihood function of the nonfaulty measurements assuming uniform measurement noise, $\mathcal{U}(-\delta, \delta)$. We have selected $\delta = 1.1 \cdot \max|e|$, which in the case of the forward approximation is $\delta = 1.5452$. Fig. 4.23 shows the resulting uncertainty region for (a) N=200 samples and (b) N=50.000 samples. The computation time (in a general purpose laptop) has been 9.14s and 29.21s, respectively.

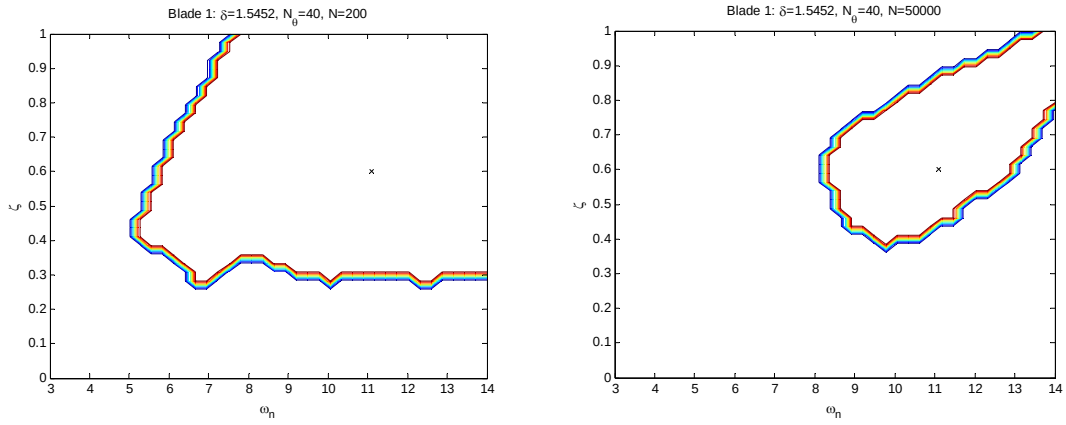
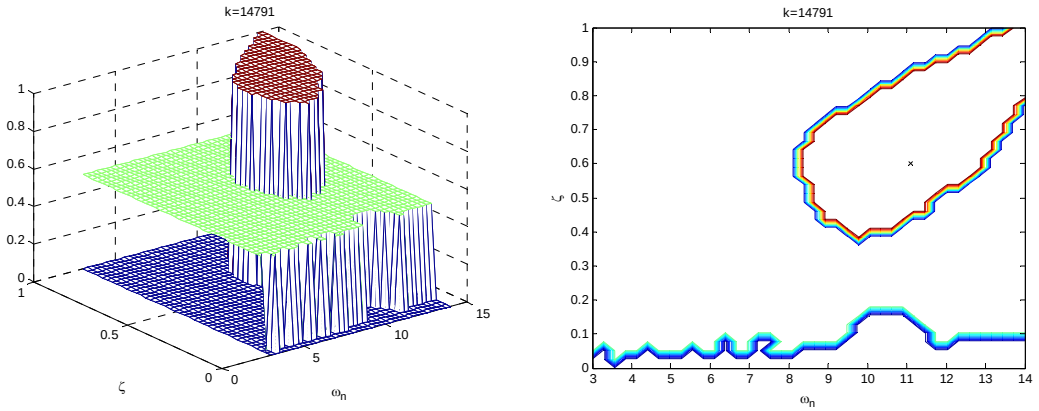


Fig. 4.23. Blade 1. Likelihood function contour plot for (a) N=200 samples and (b) N=50.000 samples. The black cross is the nominal model.

b. Fault detection

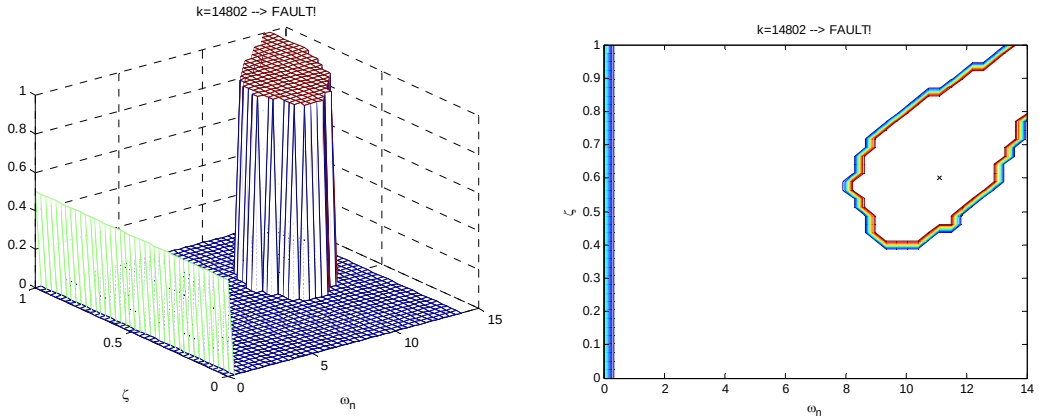
The uncertainty region obtained with N=50.000 samples has been used to check the existence of faults. A new record of measurements has been generated but now the data contain the Fault #4. For each new measurement, we compute the likelihood function assuming uniform distributed noise in a 40×40 grid and compare it to the likelihood function of Fig. 4.23(b). In the case that the two likelihood functions present some parameters of the grid in common, we say that the data are consistent with the model and consequently we decide that there is no fault. This is the case shown in Fig. 4.24, where the measurement likelihood covers all the parameters of the uncertainty region.



(a) **(b)**

Fig. 4.24. Blade 1. Normalized likelihood function corresponding to the uncertainty model and likelihood function corresponding to the sample *k* (no fault case): (a) 3D plot, (b) contour plot.

When the fault occurs, the likelihood function of the faulty measurement is far from the uncertainty region, therefore the product is zero for all the parameter grid values and the fault is decided. See next figure.



(a) **(b)**

Fig. 4.25. Blade 1. Normalized likelihood function corresponding to the uncertainty model and likelihood function corresponding to the sample *k* (fault detected): (a) 3D plot, (b) contour plot.

The fault is detected at the 14.802^{th} sample, 2 samples after the fault has been activated. Hence, the requirement of the benchmark (less that 10 samples) is satisfied.

4.3.3 Second blade. Actuator fault

Fault description: The fault considered here is the Fault #7 of the benchmark. It consists of an abrupt change of the hydraulic power. This pressure drop is modeled by changing the parameters in (86) to $\omega_{n2} = 5.73$ and $\zeta_2 = 0.45$. Also, this fault is

introduced linearly from 350s to 370s, is full active from 370s to 390s, and it linearly outfaces from 390s to 410s. In short, the fault is active from sample 28.000 to sample 32.800. Next figure shows the reference signal, the measured system output and the response of the nominal model when the fault is active. Note that it is not visually clear that the system presents a faulty behavior. This is an indirect hint that this fault is going to be difficult to detect.

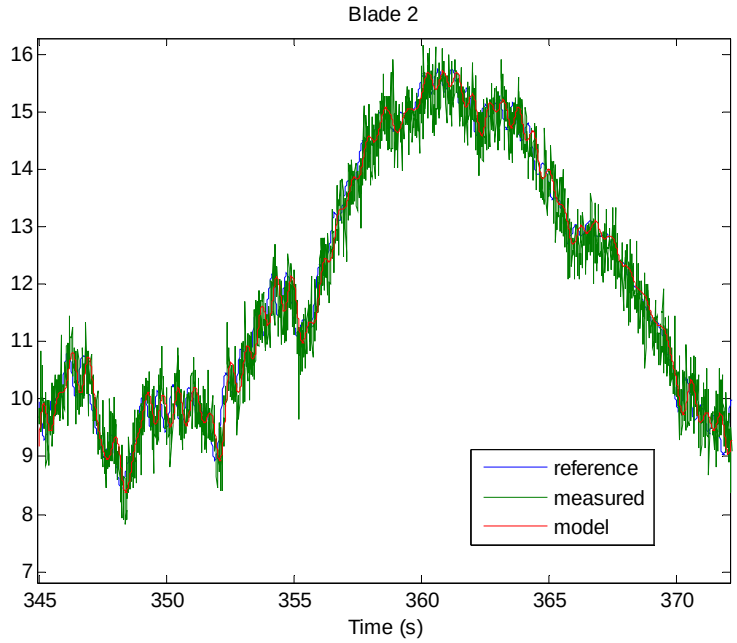


Fig. 4.26. Fault #7 acting on the second blade

The requirement is that this fault must be detected in less than eight samples, that is, the detection time must be $T_D < 8T_s$.

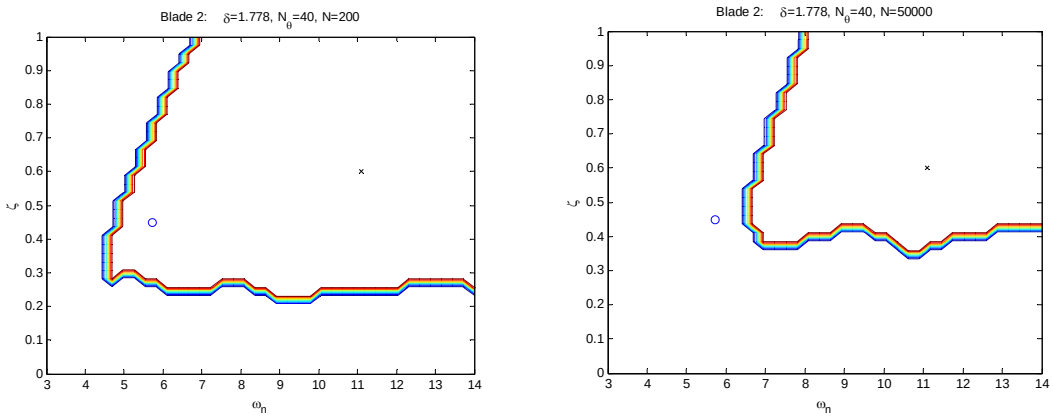
a. Calibration in a fault-free scenario

Again, the Simulink model provided by the benchmark has been used to generate a record of $N = 50.000$ samples in a fault-free scenario.

Error bound: We have proceed as in the Blade 1 and now the maximum values obtained for $|e|$ have been 1.6163 for the forward approximation, 1.6149 for the Tustin approximation, and 1.6125 for the backward approximation. Since they are very similar, we will use the forward differences discrete model again, as in Blade 1.

Uncertainty region: The feasible parameter set has been obtained following the same procedure as for the Blade 1. Here we have selected $\delta = 1.1 \cdot \max|e|$, which in the

case of the forward approximation gives $\delta = 1778$. Fig. 4.27 shows the resulting uncertainty region for (a) N=200 samples and (b) N=50.000 samples, respectively.



(a) **(b)**

Fig. 4.27. Blade 2. Likelihood function contour plot for (a) N=200 samples and (b) N=50.000 samples. The blue little circle corresponds to the perturbed model associated to Fault #7.

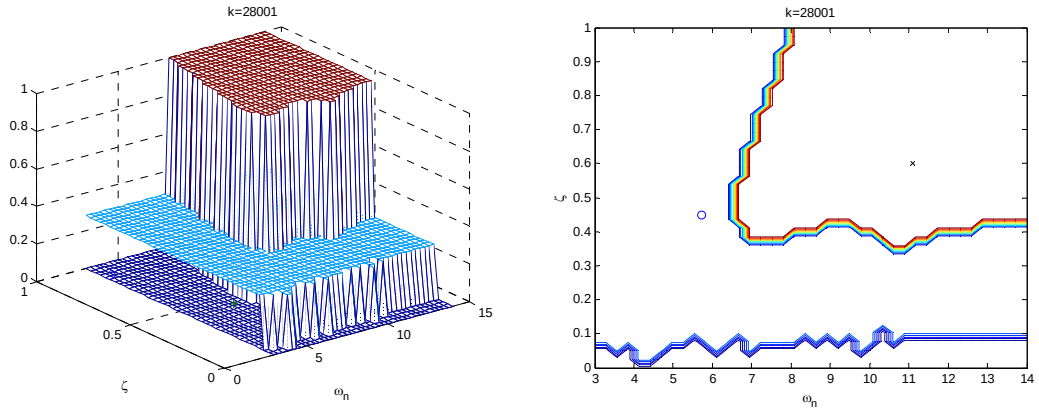
Although the two blades have the same nominal model and the reference signal is the same, the resulting uncertainty regions differ since the measurement noise realization is different. Also, note that since the N=200 region contain the perturbed model, it would be useless to detect the associated fault.

b. Fault detection

The uncertainty region obtained for N=50.000 samples is the one that will be used to perform the fault detection. A new record of measurements has been generated but now the data contain the Fault #7.

Assuming uniform noise: For each new measurement, we compute the likelihood function in a 40×40 grid, assuming that the noise is uniform distributed as $\mathcal{U}(-\delta, \delta)$, and we compare it to the likelihood function of Fig. 4.27(b).

In this case, due to the measurement noise characteristics, the new likelihood functions always cover the likelihood function corresponding to the uncertainty region. When this occurs, no fault can be detected, i.e., the method says that the measurements are always consistent with the uncertainty model. See Fig. 4.28.



(a) **(b)**

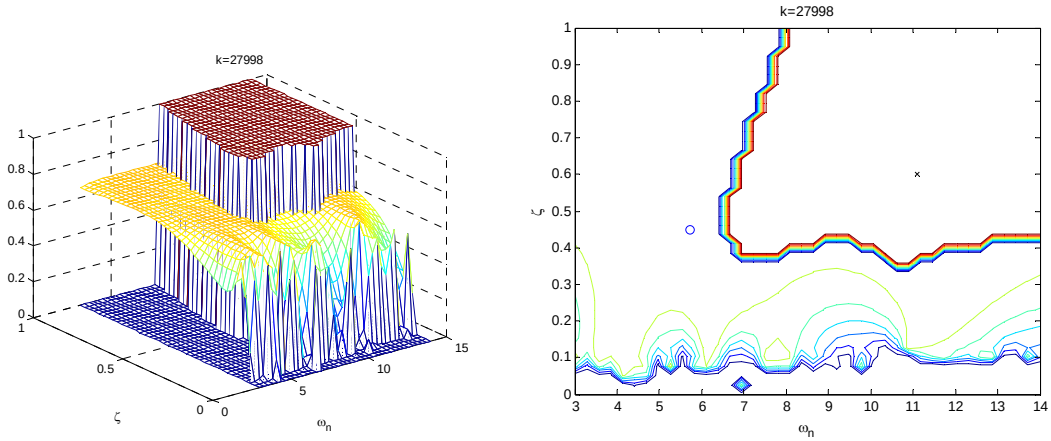
Fig. 4.28. Blade 2. No fault can be detected if we assume uniform noise

This problem can be overcome if another probability distribution for the noise is used. See next section:

Assuming Gaussian noise: Now we assume that the measurement noise is Gaussian distributed $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu = 0$ and standard deviation $\sigma = \delta/3$ (in order to include the 99% of the values of the noise realization, which are associated to the interval $[-3\sigma, 3\sigma]$).

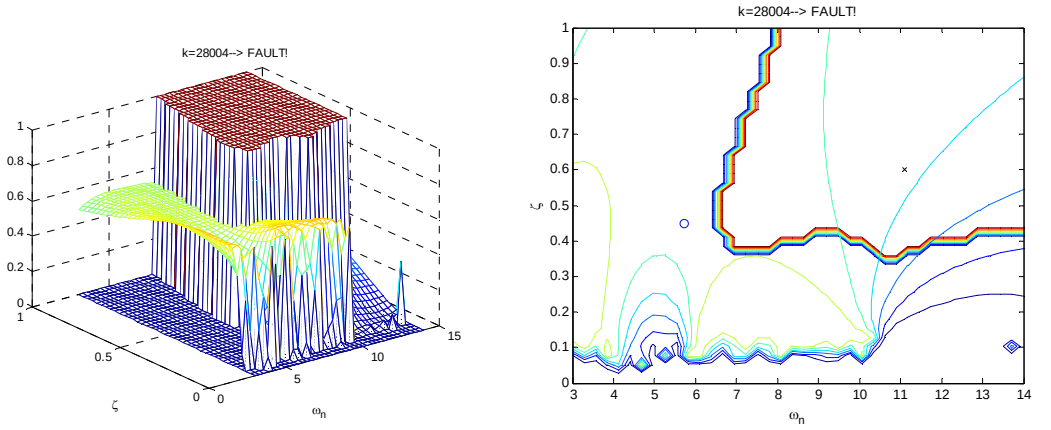
Now, the product of the likelihood function associated to the uncertainty region and the likelihood function of each new measurement will be nonzero even if a fault occurs. Therefore, to decide if the fault has taken place we must define a threshold value such that if the product of the two likelihoods is under this threshold the fault is decided. This value may be associated to a certain probability level. The selection of this value will determine the number of samples until the fault is detected (if it is too low, the number of samples before the detection will be greater) and it will affect to the generation of false alarms (if it is too high) as well. Here we have tuned the threshold value to 0.6. Another alternative is to define a threshold in terms of the volume of the resulting product.

Fig. 4.29 shows the case of no fault detected whereas Fig. 4.30 shows the case of fault detected. Finally, Fig. 4.31 shows the likelihood product for the case of no fault and fault, respectively.



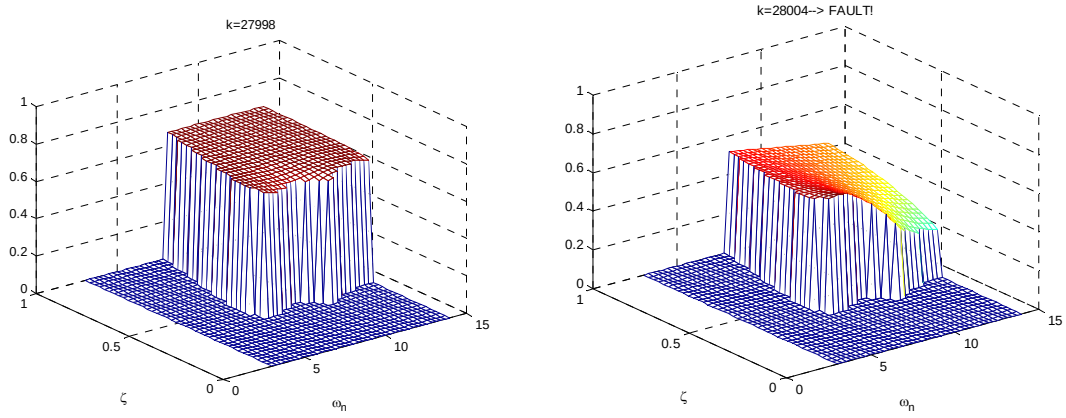
(a) **(b)**

Fig. 4.29. Blade 2. Normalized uniform likelihood function corresponding to the uncentainty model and Gauusian likelihood function for the sample *k* (no fault detected): (a) 3D plot, (b) contour plot.



(a) **(b)**

Fig. 4.30. Blade 2. Normalized uniform likelihood function corresponding to the uncentainty model and Gaussian likelihood function for the sample *k* (fault detected): (a) 3D plot, (b) contour plot.



(a) **(b)**

Fig. 4.31. Blade 2. Product of the normalized uniform likelihood function corresponding to the uncentainty model and the Gaussian likelihood function corresponding to the sample: (a) no fault detected, (b) fault detected.

In this example, the fault has been detected 4 samples after the activation of the fault. Since the benchmark requirement was $T_D < 8T_s$, we have satisfied the problem specifications.

4.3.4 Third blade. Actuator fault

Fault description: The fault considered here is the Fault #8 of the benchmark. It consists of a slow increase of the air content that can be modeled by changing the parameters in (86) to $\omega_{n3} = 3.42$ and $\zeta_3 = 0.9$. This fault is introduced linearly from 440s to 441s, is full active from 441s to 464s, and it linearly outfaces from 464s to 465s. In short, the fault is active from sample 35.200 to sample 37.200. Next figure shows the reference signal, the measured system output and the response of the nominal model when the fault is active. Note that now it is visually clear that the system presents a faulty behavior.

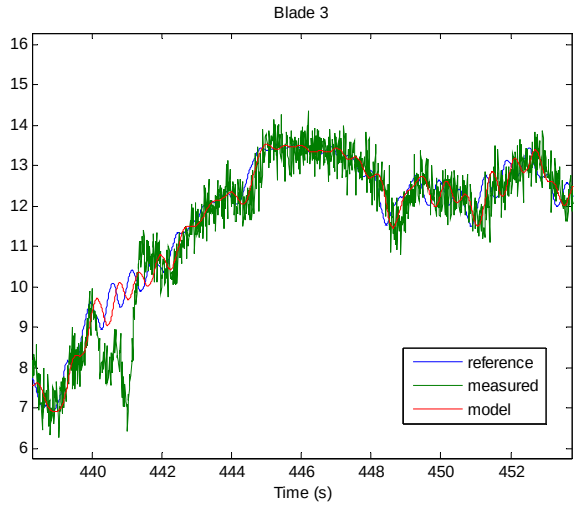


Fig. 4.32. Fault #8 acting on the third blade

The requirement is that this fault must be detected in less than 100 samples, that is, the detection time must be $T_D < 100T_s$.

a. Calibration in a fault-free scenario

Again, the Simulink model provided by the benchmark has been used to generate a record of $N = 50.000$ samples in a fault-free scenario.

Error bound: We have proceed as in Blade 1 and Blade 2 and now the maximum values obtained for $|e|$ have been 1.5255 for the forward approximation, 1.5011 for the Tustin approximation, and 1.4795 for the backward approximation. Since they are very similar, we will use the forward differences discrete model, as in the other two blades.

Uncertainty region: The feasible parameter set has been obtained for a parameters grid of 40×40 and a bound of $\delta = 16781$. Fig. 4.33 shows the resulting uncertainty region for (a) N=200 samples and (b) N=50.000 samples, respectively.

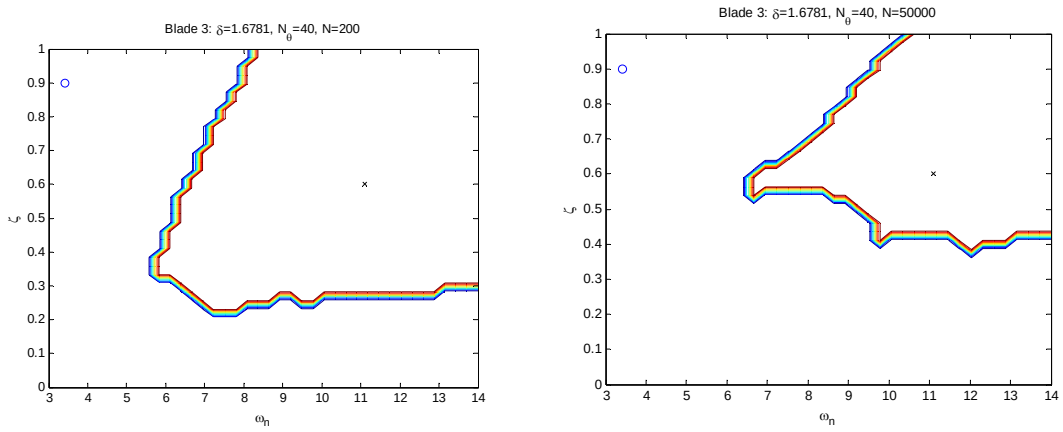


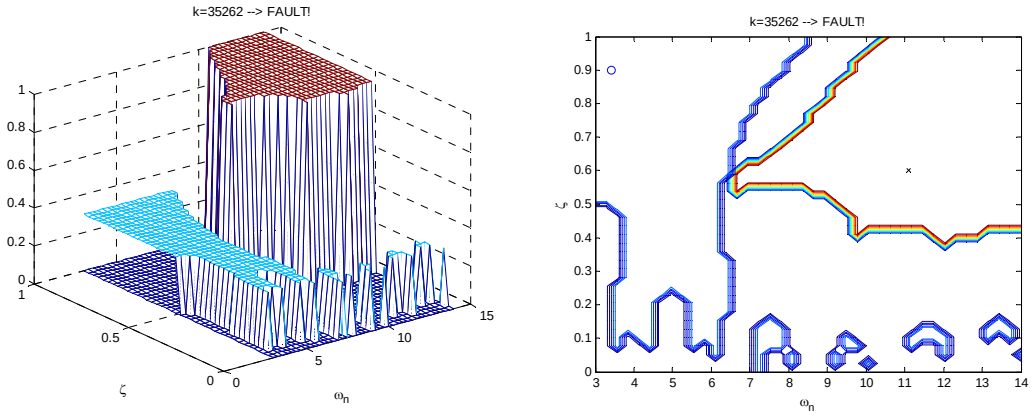
Fig. 4.33. Blade 3. Likelihood function contour plot for (a) N=200 samples and (b) N=50.000 samples. The blue little circle corresponds to the perturbed model associated to Fault #8.

b. Fault detection

The N=50.000 samples uncertainty region will be used to perform the fault detection. A new record of measurements has been generated but now the data contain the Fault #8.

Assuming uniform noise: For each new measurement, we compute the likelihood function in a 40×40 grid, assuming that the noise is uniform distributed as $\mathcal{U}(-\delta, \delta)$, and we compare it to the likelihood function of Fig. 4.33(b).

In this example the fault is detected 62 samples after its activation (see Fig. 4.34); therefore the benchmark requirements are satisfied. However, better results are obtained if Gaussian noise is assumed.



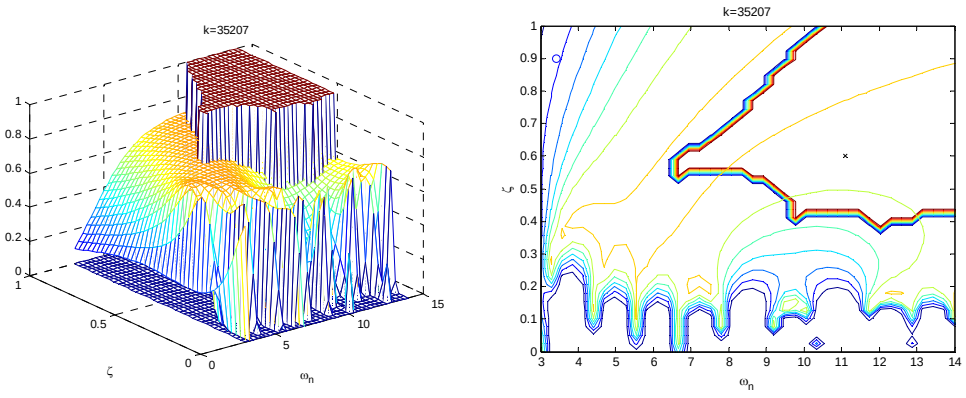
(a) **(b)**

Fig. 4.34. Blade 3. Normalized likelihood function corresponding to the uncertainty model and likelihood function corresponding to the sample *k* (fault detected): (a) 3D plot, (b) contour plot.

Assuming Gaussian noise: Now we perform the fault detection assuming that the measurement noise is Gaussian distributed $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu = 0$ and standard deviation $\sigma = \delta/3$. The threshold value has been selected to 0.6.

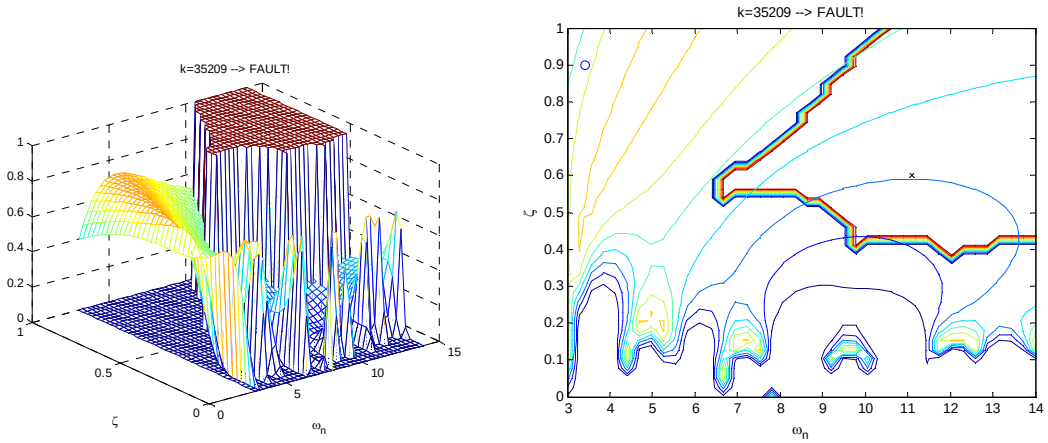
Fig. 4.35 shows the case of no fault detected and Fig. 4.36 shows the case of fault detected. Finally, Fig. 4.37 shows the likelihood product for the case of no fault and fault, respectively.

In this case, the fault is detected 9 samples after the fault is activated.



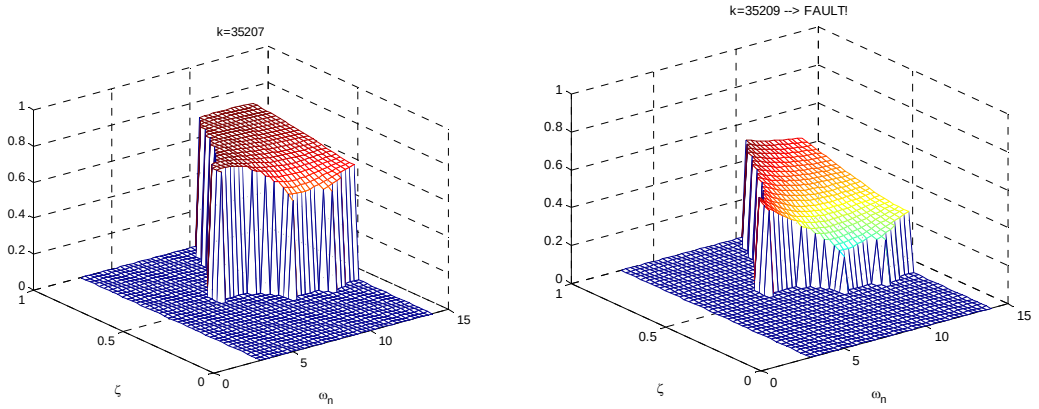
(a) **(b)**

Fig. 4.35. Blade 3. Normalized uniform likelihood function corresponding to the uncertainty model and Gaussian likelihood function for the sample *k* (no fault detected): (a) 3D plot, (b) contour plot.



(a) **(b)**

Fig. 4.36. Blade 3. Normalized uniform likelihood function corresponding to the uncertainty model and Gaussian likelihood function for the sample *k* (fault detected): (a) 3D plot, (b) contour plot.



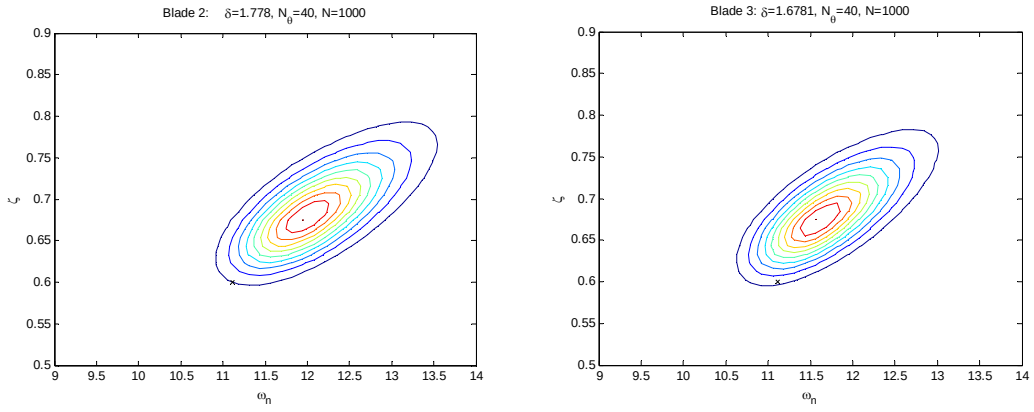
(a) **(b)**

Fig. 4.37. Blade 3. Product of the normalized uniform likelihood function corresponding to the uncertainty model and the Gaussian likelihood function corresponding to the sample: (a) no fault detected, (b) fault detected.

c. Other extensions

Probabilistic uncertainty region: From the previous results, we see that the uncertainty regions obtained assuming uniform distributed noise highly depend of the particular noise realization. Even though the underlying model and the reference signal were the same in the three blades, the resulting feasible parameter set regions were quite different.

More alike uncertainty regions for the three blades can be obtained if we assume that the measurement noise is Gaussian distributed. In this case, the regions will be not hard bounded regions no more, but their shape and size will be similar for the three blades. See Fig. 4.38.



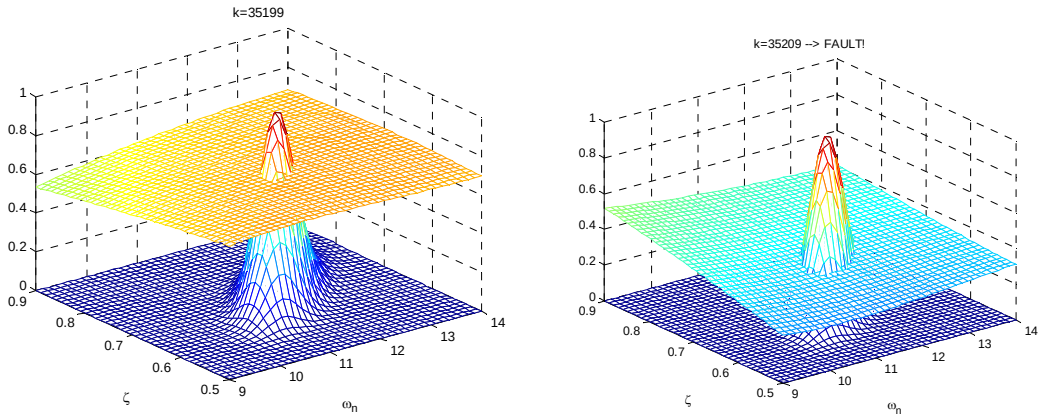
(a) **(b)**

Fig. 4.38. Probabilistic uncertainty regions for (a) blade 2 and (b) blade 3.

Introduction of prior knowledge: The regions shown in Fig. 4.38 are posterior probability regions since they have been obtained by assuming that the noise is Gaussian distributed $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu = 0$ and standard deviation $\sigma = \delta/3$, and assuming that the model parameters are also Gaussian distributed with mean $\boldsymbol{\theta}_0 = (11.11 \quad 0.6)^T$ and covariance matrix $\mathbf{P}_0 = 100 \cdot \mathbf{I}_{2\times 2}$. To obtain the regions of Fig. 4.38 we have perform the product of the (Gaussian) likelihood function of the measurements with the (Gaussian) prior distribution of the parameters. This way the nominal model is nearer the center of the uncertainty region than in the case of uniform distributed noise.

Fault detection: The regions of Fig. 4.38 have been used to perform the fault detection stage and the faults have been detected in 2, 4 and 9 samples after the activation for each blade respectively. This results coincide with the better results obtained in the previous sections. Note however that the number of samples before the detection depends on the selected threshold upon the product between the uncertainty region and the sample likelihood.

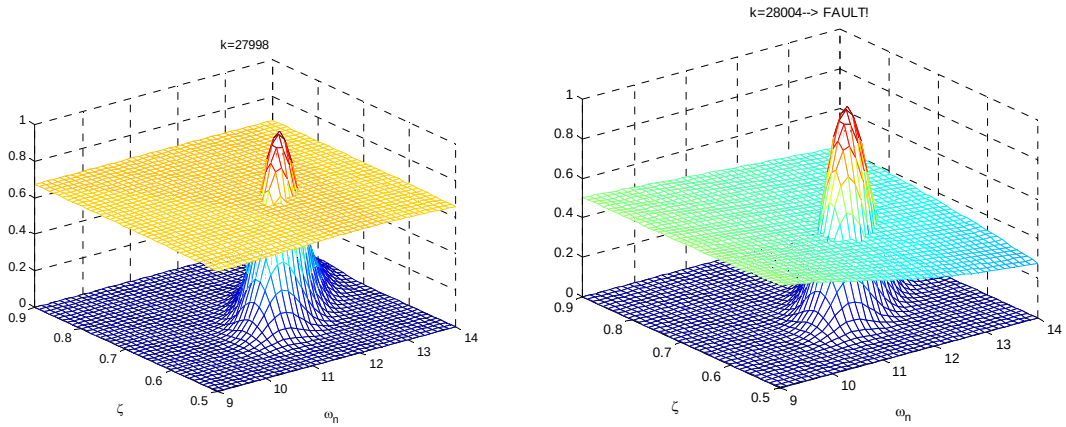
In the Blade 3 case, for a threshold value of 0.4, the fault is detected after 9 samples (see Fig. 4.39) whereas if the threshold is 0.3, the detection is fulfilled in 13 samples.



(a) **(b)**

Fig. 4.39. Blade 3. Fault detection assuming Gaussian noise and Gaussian model parameters (a) no fault detected and (b) fault detected.

In the Blade 2 case, for a threshold value of 0.6, the fault is detected after 4 samples (see Fig. 4.40). If the threshold is 0.4, the detection is attained in 16 samples.



(a) **(b)**

Fig. 4.40. Blade 2. Fault detection assuming Gaussian noise and Gaussian model parameters (a) no fault detected and (b) fault detected.

4.4 Summary and conclusion

In this chapter, we have illustrated the behavior of the Bayesian methodology to the uncertainty modeling oriented to fault detection.

For the quadruple tank process we have obtained the Feasible Parameter Regions (FPS) for the MISO case, the MISO case with observer and the MIMO case, and we have implemented the on-line fault detection algorithm. The FPSs have been obtained by applying the strips set-membership technique explained in Chapter 2 and by computing the parameters likelihood function assuming uniform distributed noise. The likelihood approach is more intensive computationally than the strips technique but it can deal with structures nonlinear in the parameters such as the plant with output observer. In the linear case, either MISO or MIMO, the FPSs obtained by both methods coincide, thus leading to the same behavior in the fault detection stage.

For the wind turbine system we have obtained three discrete models nonlinear in the parameters. Since the relationship between the model parameters is nonlinear, the linear strips set-membership technique cannot be applied. Instead we have used the methodology developed in this dissertation. Firstly we have obtained the uncertainty regions by assuming both uniform noise and Gaussian noise. In the uniform case the uncertainty regions are hard bounded and their shape and size depend on the particular noise realization. This dependence can be minimized if we assume Gaussian distributed measurement noise but, in this latter case, the regions are probabilistic. In the Bayesian framework the uncertainty regions (uniform or Gaussian) can be optionally tuned by means of the introduction of a prior distribution upon the model parameters. This may be interesting when we have some kind of prior knowledge about the model to be

identified, since this way we assign a higher probability where we know the “true” model is. Secondly, the obtained uncertainty regions have been used in the fault detection stage. Different types of faults have been generated in the blades’ pitch angle sensors and actuators. In the fault detection algorithm we have considered uniform and Gaussian distributions for computing the entering samples likelihood functions. In the Blade 2 case, the combination of the hard bounded uncertainty region and uniform distributed sample likelihood has failed to detect faults due to the high dependence to the noise realization of the uniform uncertainty region. This problem can be overcome by simply assuming Gaussian noise in the sample likelihood function and assigning a threshold to the resulting likelihood functions product. Actually, the product of the probabilistic uncertainty regions with the entering samples likelihood functions has successfully detected the faults in all the cases, widely satisfying the requirements of the benchmark problem.

CHAPTER 5

Conclusion and Future Research

In this chapter we summarize and discuss the main contributions of the present dissertation and we point out some lines for future research.

5.1 Robust identification problem

In this thesis, we have proposed a Bayesian methodology to formulate and solve the robust identification problem that takes elements of both stochastic and deterministic robust identification methods. Although parts of the problem have already been solved with Bayesian methods (Sjöberg *et al*., 1995), (Andrieu *et al*., 2001) (Andrieu *et al*., 2010), the novelty here is the definition of the so-called Bayesian Credible Model Set and the aim of establishing a framework in which all the parts of the problem can be solved within a Bayesian viewpoint.

Bayesian Credible Model Set: The key point is the definition of a *Bayesian Credible Model Set* (BCMS). This model set is inspired in the *Feasible Model Set* (FMS) of deterministic methods but it is of stochastic nature and it is obtained by combining (by means the Bayes' rule) the *a priori* information about the system and noise (prior probability distributions) with the *a posteriori* information coming from the measurement data (likelihood function). The BCMS contains all models whose posterior probability distribution conditioned to measurement data is higher than a given threshold.

The BCMS can characterize different types of models. Descriptions in the parameter space and in the frequency domain have been presented. Also, by means of the use of

hierarchical distributions and/or sets of competing models, the method can deal with models where several sources of uncertainty are present (i.e., in the structure and in the parameters). In the simplest case (linear regression models with Gaussian noise and parameters) exact expressions can be derived. For moderately high order models and arbitrary non-conjugate probability distributions, simulation methods based on Markov chain Monte Carlo (MCMC) integration are used instead.

As lines for future research we can include here the extension of the BCMS to the case where the support is a model space, e.g. the spaces ℓ_1 and $\mathcal{H}_\infty$. Since these spaces are closely related to the Robust Control theory, it is expected that a robust identification directly performed over these spaces may lead to better robust control–oriented models.

Credible regions: The model uncertainty is described by means *Highest Posterior Density* (HPD) credible regions. Credible regions are easier to compute than classical confidence regions and they enjoy some desirable properties compared to confidence regions. Credible regions may lead to smaller uncertainty regions (provided the adequate selection of the prior distributions), they can combine hard bounds with soft bounds, they can be disjoint, and they can be computed iteratively as new measurements are available (thus they can be updated on-line and therefore they are useful in fault detection procedures).

Applications of the Bayesian Decision Theory: In this dissertation we have pointed out several applications of the Bayesian Decision Theory, including the selection of a nominal model, the model (in)validation and the optimal design of the experiments.

These issues are strong candidates to future research, especially in the field of fault detection where deciding if a fault has taken place or not can be viewed as a (Bayesian, of course) hypothesis testing problem.

5.2 Interest of the Bayesian viewpoint

At this point we would like to emphasize why the Bayesian viewpoint can constitute a serious alternative to the existing robust identification methods. Actually, the Bayesian viewpoint is especially appealing for several reasons:

1. *Smaller* probabilistic uncertainty regions can be obtained if prior assumptions about plant and noise are formally entered into the modeling procedure by means the Bayes' rule.

2. In absence of *objective a priori* information, *subjective* prior assumptions can be formally entered on the model. Also, Bayesian inference gives tools to modify "erroneous" prior assumptions as new observations enter to the model (Box and Tiao, 1973), (Robert, 2001). However, it has to be said that the selection of subjective priors is, perhaps, the most important issue in the Bayesian methods. Since the computation of the posterior distributions is systematic, the selection of the prior distributions arises as a critical problem.

Although there exist very interesting literature concerning the meaning and selection of subjective priors (Jeffrey, 2004), in the present work we have found that uniform and Gaussian distributions have been sufficient for the considered examples.

3. The model description in terms of *probability distributions* is a very general and flexible one. As we have seen, we can combine in a same model the uncertainty about the model order d and the uncertainty about the parameter vector $\boldsymbol{\theta}$ by means the use of hierarchical priors of the form $p(\boldsymbol{\theta}|d)p(d)$. Moreover, no linearity assumptions are needed. This fact allows the methodology to deal, in the same manner, with structures linear in the parameters and *nonlinear* in the parameters.

4. Regarding the robust control application, an educated selection of the nominal model on the basis of a control-oriented penalty function during the modeling procedure is expected to produce uncertainty models more oriented to robust control.

5. Regarding the fault detection application, the computation of the likelihood functions and the posterior probability distributions can be performed recursively and therefore it can be implemented on-line.

5.3 Comparison to existing methods

Several connections with the existing robust identification methods have been found. On the one hand, some of them can be viewed as particular cases of the Bayesian methodology:

Particular cases: In the case where flat non-informative model priors are used, i.e., if only the likelihood function of the measurements is used, the results of the Bayesian methodology coincide with some of the existing methods. In particular, the Feasible Parameter Set (FPS) of set-membership deterministic methods can be obtained by computing the likelihood function of every set of parameters assuming uniform noise. And the same Model Error Modeling (MEM) uncertainty regions based on conventional system identification can be obtained by assuming Gaussian noise in the likelihood function computation.

On the other hand, the application of the Bayesian ideas can improve the performance of the existing methods:

Bayesian advantages: The suitable selection of the model prior distribution presents some advantages compared to conventional system identification methods and robust identification methods. For instance, in the frequency domain case, increasing the value of the prior precision matrix, $\mathbf{R}_0$, leads to small credible regions in general. This way, the bias/variance trade off of conventional methods can be overcome, i.e., one can take a high order model to reduce the bias error and afterwards select a spiky prior

distribution to compensate the expected increase of the size of the uncertainty region (variance error). Also, compared to the Non Stationary Stochastic Embedding method (NSSE), the Bayesian methodology is the same for the case of real poles than for the case of resonant poles and, again, smaller uncertainty regions may obtained by the adequate selection of the prior distributions.

In summary, we conclude that the Bayesian framework allows a unified treatment of the robust identification problem and additionally presents interesting properties compared to single current methods.

5.4 Application to fault detection

Finally, we present some concluding remarks regarding the fault detection application.

In the linear in the parameters case, the computation of FPS regions by means of the likelihood function has served as a basis for a fault detection procedure of a quadruple tank process. The results have been compared to the ones obtained by the strips intersection set-membership technique explained in Chapter 2. The likelihood approach is slightly more intensive computationally than the strip technique but it can deal with non-linear structures such as the plant with output observer. In the linear case, either MISO or MIMO, the FPSs obtained by both methods coincide, thus leading to the same behavior in the fault detection stage. Also, both strips technique and likelihood technique can be implemented on-line for fault detection purposes, being the computation time similar in both cases.

For the nonlinear in the parameters case, the developed methodology has been successfully tested in the uncertainty modeling and fault detection of a three-bladed wind turbine. In this case study, assuming uniform measurement noise in order to obtain hard bounded uncertainty regions has shown to be a wrong strategy for the fault detection of one of the blades. This bad result has been easily overcome by simply making the assumption of Gaussian noise. Even though in most examples we have assumed flat prior model distribution and uniform noise (in order to make easier the comparison to set-membership techniques), it has to be stressed that the Bayesian approach is a *probabilistic* approach, and that this stochastic nature is an advantage rather than the reverse. In a general case, the adequate selection of the model prior probability distributions may lead to probabilistic uncertainty regions that are tighter than the ones obtained by conventional system identification methods and, as we have seen, this will improve the fault detection based on them.

Future research in this field may consist in developing guidelines for the subjective priors selection and the study of the main features of the probabilistic (Bayesian) fault detection.

APPENDIX A

Optimal Estimation Theory

Classical system identification methods and stochastic robust identification methods deal with the identification of $G(q, \boldsymbol{\theta})$ from experimental data as a standard estimation problem. In this Appendix, we summarize some concepts of Optimal Estimation Theory that are used in this thesis. For more details see the classical textbooks of (Lehman and Casella, 1998) and (Casella and Berger, 2002).

A.1 Estimation problems

In Estimation Theory, three main problems are posed, namely, (1) the *point estimation* problem, (2) the *interval estimation* or *set estimation* problem, and (3) the *hypothesis testing* problem.

System identification relates to all three. To obtain a nominal parameter vector $\boldsymbol{\theta}$ one has to solve a *point estimation* problem. To obtain a confidence region for $\boldsymbol{\theta}$ one has to solve a *set estimation* problem (interval estimation problem if θ is real valued). And the problem of validating a model $\boldsymbol{\theta}$, in the sense of determining if it is inside a particular confidence region, can be viewed as a *hypothesis testing* problem. The fault detection problem can be interpreted as a hypothesis testing problem as well.

Hypothesis testing and set estimation ask the same question, but from a slightly different perspective. Both procedures look for *consistency* between observations and model. The hypothesis test fixes the parameter vector $\boldsymbol{\theta}$ and asks what observation values $\mathbf{y}$ (the acceptance region) are consistent with that fixed value. The confidence set fixes the observed values $\mathbf{y}$ and asks what parameter values $\boldsymbol{\theta}$ (the confidence interval) make this observation value most plausible.

A.1.2 Point estimation

Estimator: An *estimator* $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is a function from the observation space to the parameter space. For real valued observations and parameters we have $\widehat{\boldsymbol{\theta}}: \mathbb{R}^N \rightarrow \mathbb{R}^d$. Since a point estimator is *any* function of an observation, any *statistic* can be a point estimator. But, in general, only *sufficient* statistics are considered as estimators, (Lehmann and Casella, 1998), (Box and Tiao, 1973). Intuitively, a sufficient statistic is a function of the data that summarizes all the available sample information concerning the parameters of the distribution. For instance, for the case of a normal distribution $\mathcal{N}(\mu, \sigma^2)$, a sufficient statistic for the mean and variance (μ, σ^2) is (m, s^2), where $m = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)$.

An estimator $\widehat{\boldsymbol{\theta}}$ is characterised by its probability density function (pdf), usually computed from a N-point observation, $p(\widehat{\boldsymbol{\theta}}|N)$, its expected (mean) value $E[\widehat{\boldsymbol{\theta}}]$, its bias $E[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}$, and its variance-covariance matrix $\mathbf{C} = Cov[\widehat{\boldsymbol{\theta}}] = E\left[(\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}])(\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}])^T\right]$.

Estimator properties: A list of estimator properties can be found in (Schoukens and Pintelon, 1991). A *good* estimator should use *all* the information contained in the measurements and should exhibit *unbiasedness* (accuracy), *consistency*, *sufficiency*, *efficiency* (precision), and *robustness*. Some of these properties may be satisfied only asymptotically, for N tending to infinity. Let us summarize the main estimator properties.

The estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is said to be *unbiased* if and only if its expectation equals to $\boldsymbol{\theta}$, irrespective of sample size, that is, $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$. This implies that there are no systematic errors (bias). Nevertheless, the absolute unbiasedness is a very restrictive condition so, in many times, only *asymptotic unbiasedness* is required, that is, $\lim_{N\rightarrow\infty} E[\widehat{\boldsymbol{\theta}}_N] = \boldsymbol{\theta}$, being $\widehat{\boldsymbol{\theta}}_N$ the *estimate* from N measurements.

For a sample of size N, the estimator $\widehat{\boldsymbol{\theta}}_N$ is said to be *consistent* when it converges in probability to $\boldsymbol{\theta}$ as N tends to infinity, $\lim_{N\rightarrow\infty} \Pr[|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}| > \delta] = 0, \forall\delta > 0$. A more compact notation is $p\lim_{N\rightarrow\infty} \widehat{\boldsymbol{\theta}}_N = \boldsymbol{\theta}$. Consistency is convergence in *probability* and the case of probability one, i.e., not asymptotic, would be the *strong* consistency case. Although it is common to heuristically describe consistency as unbiasedness in large samples (asymptotic unbiasedness), they are not equivalent. An unbiased estimator will always be consistent but the opposite is not necessarily true.

The estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is said to be *efficient* if it possesses small (minimum) variance. The variations on $\widehat{\boldsymbol{\theta}}$ due to measurement noise can be described by means of the covariance matrix, $\mathbf{C} = E\left[(\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}])(\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}])^T | E[\widehat{\boldsymbol{\theta}}]\right]$. The diagonal of $\mathbf{C}$ contains

the variance of individual parameters on $\widehat{\boldsymbol{\theta}}$ whereas the off-diagonal elements contain the covariance between the different parameters on $\widehat{\boldsymbol{\theta}}$.

If the estimator is unbiased, $\mathbf{C}$ has a lower bound; if the estimator is biased it is trivial to generate $\mathbf{C} = \mathbf{0}$. Also, an estimator $\widehat{\boldsymbol{\theta}}_1$ is *relatively efficient* if, for some other estimator $\widehat{\boldsymbol{\theta}}_2$, we have $\mathbf{C}_1 \leq \mathbf{C}_2$.

Finally an estimator is said to be *robust* if (some of) its properties are still valid when the assumptions made in its construction (such as the noise distribution) are no longer applicable.

There exist several methods for finding estimators. Some possibilities are the method of moments, maximum likelihood estimators, Bayesian estimators, and invariant estimators. All these methods are detailed in (Casella and Berger, 1990). Among all, the method of maximum likelihood is by far the most popular technique for deriving estimators.

A.1.3 Hypothesis testing

A *hypothesis* is a statement about a parameter or parameter vector (Casella and Berger, 2002). The goal of a *hypothesis test* is to decide, based on the observation, which of two complementary hypotheses is true. The two complementary hypotheses in a hypothesis testing problem are called the *null hypothesis* and the *alternative hypothesis*. They are usually denoted by H_0 and H_1, respectively.

In a *hypothesis testing problem*, after performing the experiment, we must decide either "to accept H_0 as true" or "to reject H_0 as false" and thus decide H_1 is true. The subset of the sample space for which H_0 will be rejected is called the *rejection region* or *critical region.* The complement of the rejection region is called the *acceptance region.* Note that "rejecting H_0" and "accepting H_1" are not synonymous. Similarly, a distinction can be made with "accepting H_0" and "not rejecting H_0".

A hypothesis test of $H_0: \boldsymbol{\theta} \in \Theta_0$ versus $H_1: \boldsymbol{\theta} \in \Theta_0^c$ might make one of two types of errors. These are summarized in Table A.1. If $\boldsymbol{\theta} \in \Theta_0$ but the hypothesis test incorrectly decides to reject H_0, then the test has made a Type I Error. On the other hand, if $\boldsymbol{\theta} \in \Theta_0^c$ but the test decide to accept H_0, a Type II Error has been made.

		Decision	
		Accept H_0	Reject H_0
True hypothesis	H_0	Correct decision	Type I error
	H_1	Type II error	Correct decision

Table A.1. Hypothesis testing. Type I and Type II errors

There exist several methods of finding test procedures. See for instance invariant tests, union-intersection tests, intersection-union tests in (Casella and Berger, 2002). In system identification the most used is the *likelihood ratio test*. It is a very general method, almost always applicable, and it is also optimal in some cases.

Likelihood ratio test: The likelihood ratio test statistic for testing $H_0: \boldsymbol{\theta} \in \Theta_0$ versus $H_1: \boldsymbol{\theta} \in \Theta_0^c$ is

$$\lambda(\mathbf{y}) = \frac{\sup_{\Theta_0} l(\boldsymbol{\theta}|\mathbf{y})}{\sup_{\Theta} l(\boldsymbol{\theta}|\mathbf{y})} \tag{90}$$

where $l(\boldsymbol{\theta}|\mathbf{y})$ is the likelihood function. A likelihood ratio test is any test that has a rejection region of the form $\{\mathbf{y}: \lambda(\mathbf{y}) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

The numerator of $\lambda(\mathbf{y})$ is the maximum probability of the observed output sequence, the maximum being computed over parameters in the null hypothesis. The denominator is the maximum probability of the observed sample over all possible parameters. The ratio of these two maxima is small if there are parameter points in the alternative hypothesis for which the observed sample is much more likely than for any parameter point in the null hypothesis. In this situation, the criterion says that H_0 should be rejected and H_1 should be accepted as true.

The following theorem states an important asymptotic property of the likelihood ratio test.

Theorem A.1. (Casella and Berger, 2002). Let the system output be distributed as $p(\mathbf{y}|\boldsymbol{\theta})$. Under some regularity conditions on the model $p(\mathbf{y}|\boldsymbol{\theta})$, if $\boldsymbol{\theta} \in \Theta_0$ then the distribution of the statistic $-2\log\lambda(\mathbf{y})$ converges to a chi squared χ_d^2 distribution as the sample size $N \to \infty$. The number of degrees of freedom d of the limiting distribution is the difference between the number of free parameters specified by $\boldsymbol{\theta} \in \Theta_0$ and the number of free parameters specified by $\boldsymbol{\theta} \in \Theta$. □

Remark: The "regularity conditions" needed for the model are general conditions that are satisfied for many reasonable distributions (but not all). These conditions are mainly concerned with the existence and behavior of the derivatives (with respect to the parameter) of the likelihood function, and the support of the distribution (it cannot depend on the parameter).

A.1.4 Set estimation. Interval estimation

In the point estimation problem, the inference is a guess of a single value as the value of $\boldsymbol{\theta}$. The inference in a set estimation problem is the statement that "$\boldsymbol{\theta} \in C$" where $C \subset \Theta$ and $C = C(\mathbf{y})$ is a set determined by the value of the data observed.

An *interval estimate* of a real-valued parameter θ is any pair of functions $L(\mathbf{y})$ and $U(\mathbf{y})$ of the observation. If $\mathbf{y}$ is observed, the inference $L(\mathbf{y}) \leq \theta \leq U(\mathbf{y})$ is made. The *random* interval $[L(\mathbf{y}), U(\mathbf{y})]$ is called an *interval estimator.* It is important to stress that *the interval is the random quantity, not the parameter*. The *coverage probability* (that is, the probability that the random interval covers the true parameter) is a statement on terms of $\mathbf{y}$, not θ. However, the coverage probability can be a variable function of θ.

Set estimators, together with a measure of confidence (usually a confidence coefficient) are known as *confidence sets*. A confidence set with confidence coefficient equal to some value, say 1- α, is simply called a (1-α)-confidence set. Usual values for α are 0.01, 0.05, and 0.1.

In (Casella and Berger, 2002), several methods for finding interval estimators are presented: inverting a test statistic, pivotal quantities, guaranteeing an interval, invariant intervals. The test inversion presented in the next example is very general and relates confidence sets with hypothesis tests.

Example A.1. Relationship between confidence set and acceptance region

This example illustrates how to construct a $1-\alpha$ confidence set for $\boldsymbol{\theta}$, $C(\mathbf{y})$, by inverting an acceptance region, $A(\boldsymbol{\theta})$ (Casella and Berger, 2002).

Let $\{y_n\}_{n=0}^{N-1}$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ and consider testing H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$.

A reasonable *acceptance region* for the hypothesis test, i.e. the set in the sample space for which H_0 is accepted, is given by

$$A(\mu_0) = \left\{(y_0, \dots, y_{N-1}): \quad \mu_0 - \frac{z_{\alpha/2}\sigma}{\sqrt{N}} \leq \bar{y} \leq \mu_0 + \frac{z_{\alpha/2}\sigma}{\sqrt{N}}\right\}$$

where $z_{\alpha/2}$ is the $\alpha/2$-quantile of the standard distribution $\mathcal{N}(0,1)$. This means that H_0 is accepted for sample points with $|\bar{y} - \mu_0| \leq \frac{z_{\alpha/2}\sigma}{\sqrt{N}}$ and that the rejection region is the set $\left\{(y_0, \dots, y_{N-1}): \ |\bar{y} - \mu_0| > \frac{z_{\alpha/2}\sigma}{\sqrt{N}}\right\}$.

Since the test has size α, this means that $\Pr(H_0 \text{ is rejected} | \mu = \mu_0) = \alpha$ or, stated in another way, $\Pr(H_0 \text{ is accepted} | \mu = \mu_0) = 1 - \alpha$. Combining this with the above characterisation of the acceptance region, we can write

$$\Pr\left(\bar{y} - \frac{z_{\alpha/2}\sigma}{\sqrt{N}} \leq \mu_0 \leq \bar{y} + \frac{z_{\alpha/2}\sigma}{\sqrt{N}} | \mu = \mu_0\right) = 1 - \alpha$$

Since this probability statement is true for every μ_0, the statement

$$\Pr\left(\bar{y} - \frac{z_{\alpha/2}\sigma}{\sqrt{N}} \leq \mu_0 \leq \bar{y} + \frac{z_{\alpha/2}\sigma}{\sqrt{N}}\right) = 1 - \alpha$$

is also true. The interval $\left[\bar{y} - \frac{z_{\alpha/2}\sigma}{\sqrt{N}} \;,\; \bar{y} + \frac{z_{\alpha/2}\sigma}{\sqrt{N}}\right]$ obtained by *inverting* the acceptance region of the level α test, is a $(1-\alpha)$-*confidence interval*, i.e., the set in the parameter space with plausible values for μ, and can be expressed as

$$C(\mathbf{y}) = C(y_0, \ldots, y_{N-1}) = \left\{\mu:\ \bar{y} - \frac{z_{\alpha/2}\sigma}{\sqrt{N}} \le \mu \le \bar{y} + \frac{z_{\alpha/2}\sigma}{\sqrt{N}}\right\}$$

These sets are connected to each other by the tautology $\mathbf{y} \in A(\mu_0) \Leftrightarrow \mu_0 \in C(\mathbf{y})$. ■

The quality of set estimators is related to the probability of covering false values. The probability of false coverage indirectly measures the size of a confidence set. Intuitively, smaller sets cover fewer values and, hence, are less likely to cover false values. The probability of coverage of $C(\mathbf{y})$, that is, the probability of true coverage, is the function of $\boldsymbol{\theta}$ given by $\Pr[\boldsymbol{\theta} \in C(\mathbf{y})]$. The probability of false coverage is the function of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}$' defined by the probability of covering $\boldsymbol{\theta}$' when $\boldsymbol{\theta}$ is the true parameter. A (1-α)-confidence set that minimizes the probability of false coverage over a class of $1-\alpha$ confidence sets is called a *uniformly most accurat*e confidence set.

A.2 Maximum Likelihood Estimation

In this section we consider the system identification problem from a Maximum Likelihood Estimation (MLE) viewpoint and present the classical system identification as a particular case of the MLE.

In classical system identification, it is usual practice to model the system by means a parameter vector $\boldsymbol{\theta}$,

$$y_n = G(q, \boldsymbol{\theta})u_n + H(q, \boldsymbol{\theta})v_n, \qquad \textbf{(91)}$$

where q is the forward shift operator, $qu_n = u_{n+1}$, G and H are rational transfer functions in this operator, $\{y_n\}_{n=0}^{N-1}$ and $\{u_n\}_{n=0}^{N-1}$ are respectively the observed output and input samples, and $\{v_n\}_{n=0}^{N-1}$ is a i.i.d. (independent identically distributed) stochastic process with zero mean and finite variance.

> **Remark:** More general system structures can be modeled if we use state-space descriptions.

In this context, the goal of system identification is to obtain an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ given the measurements $\{y_n\}_{n=0}^{N-1}$ and $\{u_n\}_{n=0}^{N-1}$. In particular, classical system identification solves this problem by means the minimization of a cost function V_N depending on the prediction error $\mathcal{E}_n(\boldsymbol{\theta}) = y_n - \hat{y}_{n|n-1}(\boldsymbol{\theta})$.

$\hat{y}_{n|n-1}(\boldsymbol{\theta})$ is the one-step ahead predictor of y_n based on the model parameterised by $\boldsymbol{\theta}$ and the past observations, $Y_{n-1} \triangleq \{y_0, \ldots, y_{n-1}\}$, and it can be expressed as

$$\hat{y}_{n|n-1}(\boldsymbol{\theta}) \triangleq E_{\boldsymbol{\theta}}[y_n|Y_{n-1}] = \int p(y_n|Y_{n-1}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{92}$$

where $E_{\boldsymbol{\theta}}[.]$ is the statistical expectation operator with respect to a pdf (probability density function) dependent on $\boldsymbol{\theta}$.

To compute $\hat{y}_{n|n-1}(\boldsymbol{\theta})$, the steady state Wiener filter may be used

$$\hat{y}_{n|n-1}(\boldsymbol{\theta}) = H^{-1}(q, \boldsymbol{\theta}) G(q, \boldsymbol{\theta}) u_n + [1 - H^{-1}(q, \boldsymbol{\theta})] y_n, \tag{93}$$

provided that $H(q, \boldsymbol{\theta})$ is monic. To obtain (93) one simply has to substitute $v_n = y_n - \hat{y}_{n|n-1}(\boldsymbol{\theta})$ in (91).

A.2.1 Likelihood Function

From an estimation theory viewpoint, since the system output samples $\{y_n\}_{n=0}^{N-1}$ ($\mathbf{y}$ in matrix notation) can be considered a realization of a stochastic process, the system output can be described by means the conditional distribution $p(\mathbf{y}|\boldsymbol{\theta})$ which is called the *sample distribution*.

Consider for instance the linear regression model $\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \mathbf{v}$. If we assume i.i.d. additive Gaussian noise $\mathbf{v} \sim \mathcal{N}(0, \sigma_v^2 \mathbf{I})$, the samples of the system output will be distributed as $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\theta}, \sigma_v^2 \mathbf{I})$.

The conditional distribution $p(\mathbf{y}|\boldsymbol{\theta})$ is interpreted as the likelihood that the system modeled by $\boldsymbol{\theta}$ has generated the observed process $\mathbf{y}$. In this work, the *likelihood function* (LF) will be denoted $l(\boldsymbol{\theta}|\mathbf{y})$.

> **Remark:** Although the mathematical expression is $l(\boldsymbol{\theta}|\mathbf{y}) \equiv p(\mathbf{y}|\boldsymbol{\theta})$, a likelihood function is not a probability density function (pdf) since it is not defined axiomatically. Sometimes, in order to get likelihood functions that integrate to one, it is used the standardized likelihood, $l(\boldsymbol{\theta}|\mathbf{y}) = c \cdot p(\mathbf{y}|\boldsymbol{\theta})$, where $c^{-1} = p(\mathbf{y}) = E_{\boldsymbol{\theta}}[p(\mathbf{y}|\boldsymbol{\theta})] = \int p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the normalizing constant.

Computation of the likelihood function: The computation of the likelihood function can be performed by means the application of the Bayes' rule, $p(A, B) = p(B|A)p(A)$.

$$l(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}) = p(y_0..y_{N-1}|\boldsymbol{\theta}) = p(y_{N-1}|y_0..y_{N-2}, \boldsymbol{\theta}) p(y_0..y_{N-2}, \boldsymbol{\theta})$$

where $p(y_0..y_{N-2}, \boldsymbol{\theta}) = p(y_{N-2}|y_0..y_{N-3}, \boldsymbol{\theta}) p(y_0..y_{N-3}, \boldsymbol{\theta})$ and so on. We proceed iteratively until $p(y_0, y_1, \boldsymbol{\theta}) = p(y_1|y_0, \boldsymbol{\theta}) p(y_0, \boldsymbol{\theta})$. Thus, we have

$$l(\boldsymbol{\theta}|\mathbf{y}) = p(y_0, \boldsymbol{\theta}) \prod_{n=1}^{N-1} p(y_n|Y_{n-1}, \boldsymbol{\theta}) \tag{94}$$

where $Y_n = \{y_0, \ldots, y_n\}$.

The key point is that

$$p(y_n|Y_{n-1}, \boldsymbol{\theta}) = p_\varepsilon(y_n - \hat{y}_{n|n-1}(\boldsymbol{\theta})) = p_\varepsilon(\varepsilon_n) \tag{95}$$

where $\mathcal{E}_n = \mathcal{E}_n(\boldsymbol{\theta})$ are the residual errors and $p_\varepsilon(.)$ is their associated pdf. Moreover, $p_\varepsilon(\varepsilon_n) = p_v(v_n)$ where v_n is the measurement noise.

> **Remark:** Classical Prediction Error Methods (PEM) consider that there is no error in the model structure, therefore the error in the parameter vector estimate and the prediction errors (residuals) are only due to the measurement noise.

As we are considering i.i.d. noise, the joint distribution of the sequence is

$$p_v(\mathbf{v}) = \prod_{n=0}^{N-1} p_v(v_n) \tag{96}$$

Finally, the result is that we can compute the likelihood function using the prediction errors and the measurement noise pdf.

$$l(\boldsymbol{\theta}|\mathbf{y}) = \prod_{n=0}^{N-1} p_v(\varepsilon_n) \tag{97}$$

A.2.2 Maximum Likelihood Point Estimation

The likelihood function has played a fundamental role in the last decades since most parameter estimation techniques rely in the maximum likelihood (ML) paradigm. The estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{y})$ is a maximum likelihood estimator (MLE) if, fixed $\mathbf{y}$, the likelihood function $l(\widehat{\boldsymbol{\theta}}|\mathbf{y})$ attains its maximum (Schoukens and Pintelon, 1991). A list of MLEs can be found in (Gustaffson and Hjalmarsson, 1995) and see (Ljung, 1999a) for a general treatment of the topic.

a. Differentiation

If the likelihood function is differentiable in θ_i, possible candidates for the MLE are the values $\boldsymbol{\theta}$ that solve

$$\frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta}|\mathbf{y}) = 0 \quad , \quad i = 1..d \tag{98}$$

These solutions are only *possible* candidates for the MLE since the first derivative being zero is only a necessary condition for a maximum, not a sufficient condition. Furthermore, the zeros of the first derivative only locate extreme points in the interior of the domain of a function. If the extrema occur on the boundary the first derivative may not be zero. Thus, the boundary must be checked separately for extrema. Points at which the first derivative is zero may be local or global minima, local or global maxima, or inflection points.

There are two inherent drawbacks associated with the general problem of finding the maximum of a function: the first is that of actually finding the *global maximum* and verifying that, indeed, a global maximum has been found; the second problem is that the MLE may be very *sensitive numerically* and sometimes a slightly different sample will produce a vastly different MLE.

Another way to find an MLE is to abandon differentiation and proceed with a direct maximization. This method is sometimes numerically hard to implement.

b. Log-Likelihood Function

In most cases, especially when differentiation is to be used, it is easier to work with the natural logarithm of $l(\boldsymbol{\theta}|\mathbf{y})$,

$$L(\boldsymbol{\theta}|\mathbf{y}) = \log l(\boldsymbol{\theta}|\mathbf{y}) \tag{99}$$

known as the *log-likelihood function*.

The log-likelihood function allows solving the MLE problem as a minimization one:

$$\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}}[-L(\boldsymbol{\theta}|\mathbf{y})] \tag{100}$$

c. Fisher Information Matrix

If the log-likelihood function $L(\boldsymbol{\theta}|\mathbf{y})$ is differentiable twice, one can define the Fisher Information Matrix as:

$$\mathbf{F}_i = E\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}L\right)^T\left(\frac{\partial}{\partial\boldsymbol{\theta}}L\right)\middle|\,\boldsymbol{\theta}\right] = E\left[-\frac{\partial^2 L}{\partial\boldsymbol{\theta}^T\partial\boldsymbol{\theta}}\middle|\,\boldsymbol{\theta}\right] \tag{101}$$

This matrix measures the amount of information present in the measurements $\mathbf{y}$, in relation to the parameters $\boldsymbol{\theta}$.

d. Cramér-Rao Bound

The inverse of the Fisher matrix, $\mathbf{F}_i^{-1} = CR$, is known as the Cramér-Rao bound. This is a lower bound for the covariance matrix $\mathbf{C}$ of an estimator. It can be shown that it is impossible to have an unbiased estimator with $\mathbf{C}$ smaller than the Cramér-Rao bound. In fact, any estimator that reaches Cramér-Rao bound is a MLE.

The existence of this bound is independent of the estimator type: it only needs the measurement noise probability distribution and the exact parameter vector $\boldsymbol{\theta}_{true}$ to be calculated.

Another feature of CR is that, given the same amount of information, the introduction of extra parameters in the model increases the Cramér-Rao bound, that is, it makes larger the uncertainty on the estimates. So, if the order increases, the variance decreases but the bias increases.

A.2.3 Properties of Maximum Likelihood Estimators

a. Properties of MLE

In general, the MLE

$$\boldsymbol{\theta}_o \triangleq \arg\max_{\boldsymbol{\theta}} \lim_{N\to\infty} \frac{1}{N} E[L(\boldsymbol{\theta}|\mathbf{y})] \tag{102}$$

is a good point estimator, possessing some nice properties (Ninness, 2009). If measurement noise is i.i.d. and the log-likelihood function $L(\boldsymbol{\theta}|\mathbf{y})$ is differentiable twice, it can be shown that MLE is unique, asymptotically unbiased

$$\lim_{N\to\infty} E\left[\widehat{\boldsymbol{\theta}}_N\right] = \boldsymbol{\theta}_o$$

strongly consistent,

$$\lim_{N\to\infty} \widehat{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_o \quad \text{w.p.1} \quad \text{(with probability one)} \tag{103}$$

and asymptotically efficient, i.e. its covariance matrix $\mathbf{C}$ approaches the Cramér-Rao bound as N increases,

$$\mathbf{C}^{-1} = \frac{1}{\sigma^2} \lim_{N\to\infty} \frac{\partial^2}{\partial\boldsymbol{\theta}^T \partial\boldsymbol{\theta}} \frac{1}{N} E[L(\boldsymbol{\theta}|\mathbf{y})] \tag{104}$$

where $\sigma^2 = E\left[\boldsymbol{\varepsilon}^2\right]$.

Moreover, the estimates $\widehat{\boldsymbol{\theta}}_N$ present invariance properties and are asymptotically normal distributed,

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_o\right) \to \mathcal{N}(0, \mathbf{P}) \quad \text{as} \quad N \to \infty \tag{105}$$

Finally, these results can still hold even if $p_\varepsilon(\cdot)$ is not equal to the underlying true one, but instead it merely satisfies some mild regularity conditions.

b. Limitations of MLE

Firstly, results (62), (104) and (105) are asymptotic in data length N. But, in practice, one usually assumes that they hold approximately for N finite.

Secondly, these methods assume that the whole dynamics of the system can be explained by means a parameter vector $\boldsymbol{\theta}_o$. And therefore the only error of the model is in the parameters values. So these methods are not suitable for describing the model uncertainty in the way robust control needs.

Thirdly, these results are valid only for the parameter vector estimate $\widehat{\boldsymbol{\theta}}_N$. Sometimes we need to estimate not the parameters but a function of them such as the system frequency response. In these cases one has to form a first order Taylor expansion of the function of interest about $\boldsymbol{\theta}_o$, and then use (105) to obtain the estimate of the function together with error bounds. The result will be accurate if $\left\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_o\right\|$ is small, and this depends on the data length N being large.

A.2.4 Maximum Likelihood Estimators and System Identification

a. Relationship to classical system identification

Classical system identification can be viewed as a particular case of MLE. In these methods the general solution to the problem of the parameter vector identification is:

$$\widehat{\boldsymbol{\theta}}_N \triangleq \arg\min_{\boldsymbol{\theta}} \lim_{N\to\infty} V_N(\boldsymbol{\theta}) \tag{106}$$

where V_N is a cost function depending on the prediction error $\varepsilon_n(\boldsymbol{\theta}) = y_n - \hat{y}_{n|n-1}(\boldsymbol{\theta})$,

$$V_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \ell(\varepsilon_n(\boldsymbol{\theta})) \tag{107}$$

The function $\ell(\cdot)$ is an arbitrary positive mapping and it is usually chosen as $\ell(\mathbf{x}) = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2$. The measurement noise is assumed i.i.d. zero mean Gaussian and thus prediction errors are also assumed $\varepsilon_n \sim \mathcal{N}(\mu_\varepsilon, \sigma_\varepsilon^2)$ with $\mu_\varepsilon = 0$ and independent,

$$p_\varepsilon = \frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\exp\left(-\frac{(\varepsilon-\mu_\varepsilon)^2}{2\sigma_\varepsilon^2}\right)$$

Since

$$\log p_\varepsilon(\mathcal{E}_n(\boldsymbol{\theta})) = -\log\sqrt{2\pi} - \log\sigma_\varepsilon - \frac{\varepsilon_n^2(\boldsymbol{\theta})}{2\sigma_\varepsilon^2}$$

we have $\ell(\mathcal{E}_n(\boldsymbol{\theta})) = \varepsilon_n^2(\boldsymbol{\theta}) = -2\sigma_\varepsilon^2\left[\log p_\varepsilon(\mathcal{E}_n(\boldsymbol{\theta})) + \log\sqrt{2\pi} + \log\sigma_\varepsilon\right]$

On the other hand, the log-likelihood function can be expressed as

$$\log l(\boldsymbol{\theta}|\mathbf{y}) = \log\prod_{n=0}^{N-1} p_v(\varepsilon_n) = \sum_{n=0}^{N-1}\log p_v(\varepsilon_n) = \sum_{n=0}^{N-1}\log p_\varepsilon(\mathcal{E}_n(\boldsymbol{\theta}))$$

Thus, we can write (107) as

$$\begin{aligned} V_N(\boldsymbol{\theta}) &= \frac{-2\sigma_\varepsilon^2}{N}\sum_{n=1}^{N}\left[\log p_\varepsilon(\mathcal{E}_n(\boldsymbol{\theta})) + \log\sqrt{2\pi} + \log\sigma_\varepsilon\right] \\ &= \frac{-2\sigma_\varepsilon^2}{N}\left[\log l(\boldsymbol{\theta}|\mathbf{y}) + N(\log\sqrt{2\pi} + \log\sigma_\varepsilon)\right] \end{aligned}$$

In other words, minimizing $V_N(\boldsymbol{\theta})$ is equivalent to minimize $-\log l(\boldsymbol{\theta}|\mathbf{y})$ and this is equivalent to maximize $l(\boldsymbol{\theta}|\mathbf{y})$. Thus classical system identification methods are a particular case of ML estimation.

b. Computational implementation of the LSE

Regarding the computational implementation of equations

$$\widehat{\boldsymbol{\theta}}_N = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{y},$$

if we define $\mathbf{F} = \boldsymbol{\Phi}^T\boldsymbol{\Phi} = \frac{1}{N}\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n\boldsymbol{\varphi}_n^T$, normal equations are expressed as $\mathbf{F}\widehat{\boldsymbol{\theta}}_N = \frac{1}{N}\sum_{n=0}^{N-1}\boldsymbol{\varphi}_n y_n$ (Ljung, 1999a).

Computation of $\widehat{\boldsymbol{\theta}}_N$ avoids construction of matrix $\mathbf{F}$ since this may be ill-conditioned. Instead, a so-called "square root algorithm" is used and a matrix $\mathbf{M}$ is constructed with the property $\mathbf{M}\mathbf{M}^T = \mathbf{F}$. There exist different possibilities for the construction of $\mathbf{M}$. QR factorizations have been used in the simulation examples of this thesis.

The QR factorization of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as $\mathbf{A} = \mathbf{Q}\mathbf{R}$, $\mathbf{Q} \in \mathbb{R}^{m\times m}$ and $\mathbf{R} \in \mathbb{R}^{m\times n}$,where $\mathbf{Q}$ is orthonormal, $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}\mathbf{Q}$, and $\mathbf{R}$ is upper triangular.

Time domain: Forr the time domain case, $\mathbf{y} = \mathbf{\Phi\theta} + \mathbf{v}$, the problem is solved by choosing $[\mathbf{\Phi} \quad \mathbf{y}] = \mathbf{QR}$ and $\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{R}_2 \\ \mathbf{0}_{1\times d} & \mathbf{R}_3 \\ \mathbf{0}_{(N-d-1)\times d} & \mathbf{0}_{(N-d-1)\times 1} \end{bmatrix}$, where $[\mathbf{\Phi} \quad \mathbf{y}]$ is $N \times (d+1)$, $\mathbf{R}_1$ is a $d \times d$ triangular matrix, $\mathbf{R}_2$ is $d \times 1$, $\mathbf{R}_3$ is scalar.

The LS criterion function V_N is not affected by the orthonormal transformation $\mathbf{Q}$ applied to the vector $\mathbf{y} - \mathbf{\Phi\theta}$, $V_N = |\mathbf{y} - \mathbf{\Phi\theta}|^2 = |\mathbf{Q}^T(\mathbf{y} - \mathbf{\Phi\theta})|^2$. Therefore,

$$V_N = |\mathbf{Q}^T(\mathbf{y} - \mathbf{\Phi\theta})|^2 = \left|\mathbf{Q}^T[\mathbf{\Phi} \quad \mathbf{y}]\begin{pmatrix}-\boldsymbol{\theta}\\ 1\end{pmatrix}\right|^2 = \left|\mathbf{Q}^T\mathbf{QR}\begin{pmatrix}-\boldsymbol{\theta}\\ 1\end{pmatrix}\right|^2 = \left|\mathbf{R}\begin{pmatrix}-\boldsymbol{\theta}\\ 1\end{pmatrix}\right|^2$$
$$= \left|\begin{pmatrix}\mathbf{R}_2 - \mathbf{R}_1\boldsymbol{\theta}\\ \mathbf{R}_3\\ \mathbf{0}_{(N-2)\times 1}\end{pmatrix}\right|^2$$

Finally,

$$V_N = |\mathbf{R}_2 - \mathbf{R}_1\boldsymbol{\theta}|^2 + |\mathbf{R}_3|^2$$

That is, V_N is minimized for $\mathbf{R}_1\widehat{\boldsymbol{\theta}}_{LS} = \mathbf{R}_2$, giving $\min V_N = |\mathbf{R}_3|^2$.

Frequency domain: For the frequency domain case the procedure is analogous. We only have to express the system frequency response in matrix notation, $\mathbf{G} = \mathbf{\Gamma\theta} + \mathbf{w}$, and choose $[\mathbf{\Gamma} \quad \mathbf{G}] = \mathbf{QR}$.

Let us define $\mathbf{G}$ and $\mathbf{\Gamma}$.

The system frequency response can be expressed as

$$G(e^{j\omega_m}) = \sum_{k=0}^{d-1} \theta_k B_k(e^{j\omega_m}) + w_m, \qquad m = 0, 1, \ldots, M-1$$

where $G(e^{j\omega_m})$ is the system frequency response "measured" at frequency ω_m. In fact, frequency response is not directly measured but estimated from time domain input-output data. The estimate is usually optimal in a least squares sense. Therefore the error term w_m is due to both measurement noise corrupting original time domain data and frequency response estimation error.

Assuming that the model is parameterized in terms of basis functions, the model frequency response at frequency ω_m can be expressed in terms of the frequency response of the basis functions as

$$(B_0(e^{j\omega_m}) \quad \ldots \quad B_{d-1}(e^{j\omega_m}))\begin{pmatrix}\theta_0\\ \vdots\\ \theta_{d-1}\end{pmatrix} = \mathbf{B}(\omega_m)\boldsymbol{\theta}$$

For convenience we separate real part and imaginary part of the model frequency response,

$$\begin{pmatrix} \text{Re } \mathbf{B}(\omega_m) \\ \text{Im } \mathbf{B}(\omega_m) \end{pmatrix} \boldsymbol{\theta} = \begin{pmatrix} \mathbf{B}^R(\omega_m) \\ \mathbf{B}^I(\omega_m) \end{pmatrix} \boldsymbol{\theta}$$

Thus, "output vector" $\mathbf{G}$ corresponding to the frequency domain measures is a real-valued vector of length $2\times M$,

$$\mathbf{G}^T = (\text{Re } G(e^{j\omega_1}) \quad \text{Im } G(e^{j\omega_1}) \quad \ldots \quad \text{Re } G(e^{j\omega_M}) \quad \text{Im } G(e^{j\omega_M}))$$

and regression matrix $\boldsymbol{\Gamma}$ is a $2M\times d$ matrix,

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{B}^R(\omega_m) \\ \mathbf{B}^I(\omega_m) \\ \vdots \\ \mathbf{B}^R(\omega_M) \\ \mathbf{B}^I(\omega_M) \end{pmatrix} = \begin{pmatrix} \text{Re } B_0(e^{j\omega_1}) & \cdots & \text{Re } B_{d-1}(e^{j\omega_1}) \\ \text{Im } B_0(e^{j\omega_1}) & & \text{Im } B_{d-1}(e^{j\omega_1}) \\ \vdots & & \vdots \\ \text{Re } B_0(e^{j\omega_M}) & & \text{Re } B_{d-1}(e^{j\omega_M}) \\ \text{Im } B_0(e^{j\omega_M}) & \ldots & \text{Im } B_{d-1}(e^{j\omega_M}) \end{pmatrix}$$

In matrix notation, $\mathbf{G} = \boldsymbol{\Gamma}\boldsymbol{\theta} + \mathbf{w}$.

Computing frequency domain data $\mathbf{G}$ from time domain measurements is a first step in many robust identification techniques such as the non-stationary stochastic embedding.

A.3 Summary of point estimators

Table A.2 summarizes the main point estimators and it is based in the (Eykhoff, 1974) classification of the most used point estimators depending on the (pdf) information required about the measurement noise $\mathbf{v}$, the plant to be identified $\boldsymbol{\theta}_{true}$, and the *cost* of wrong modeling $C(\boldsymbol{\theta}, \boldsymbol{\theta}_{true})$. Note that we assume that the plant can be modelled by $\boldsymbol{\theta}_{true}$ so the system identification problem is reduced to obtain a good estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{N}}$ for $\boldsymbol{\theta}_{true}$.

	Least Squares (LS)	Weighted Least Squares (WLS)	Maximum Likelihood (ML)	Maximum *a Posteriori* (MAP)	Minimum Risk (MR)
A priori info					
about $\mathbf{v}$	none	$E[\mathbf{v}],\ E[\mathbf{v}\mathbf{v}^T]$	$p(\mathbf{v})$	$p(\mathbf{v})$	$p(\mathbf{v})$
about $\boldsymbol{\theta}$	none	none	none	$p(\boldsymbol{\theta})$	$p(\boldsymbol{\theta})$
about cost	none	none	none	none	$C(\boldsymbol{\theta}, \boldsymbol{\theta}_{true})$
Assumptions					
about $\mathbf{v}$	$\mathbf{v}\sim\mathcal{N}(0,\mathbf{I})$	$\mathbf{v}\sim\mathcal{N}(\bar{v},\sigma_v^2\mathbf{I})$	none	none	none
about $\boldsymbol{\theta}$	$p(\boldsymbol{\theta}) \propto ct$	$p(\boldsymbol{\theta}) \propto ct$	$p(\boldsymbol{\theta}) \propto ct$	none	none
about cost	n.c.(1)	n.c	n.c	n.c	$C(\boldsymbol{\theta}, \boldsymbol{\theta}_{true})$ (2)
Objective criterion	$\min_{\boldsymbol{\theta}} \lVert\boldsymbol{\varepsilon}\rVert_2^2$ (3)	$\min_{\boldsymbol{\theta}} \lVert\boldsymbol{\varepsilon}\rVert_2^2$ (3)	$\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}\|\mathbf{y})$	$\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}\|\mathbf{y})$	$\min_{\boldsymbol{\theta}} E_{p(\boldsymbol{\theta}\|\mathbf{y})}[C(\boldsymbol{\theta}, \boldsymbol{\theta}_{true})]$

(1) n.c.: not considered, (2) cost can be either known or assumed, (3) they are equivalent to ML

Table A.2. Summary of point estimators. Prior knowledge vs. arbitrary assumptions

The *Least Squares Estimate* (LSE) is the one that needs less information. In fact no prior information is required about measurement noise, plant, or cost. Instead, LSE *arbitrarily assumes* that noise $\mathbf{v}$ is i.i.d. Gaussian with mean $E[\mathbf{v}] = 0$ and covariance matrix $E[\mathbf{v}\mathbf{v}^T] = \mathbf{I}$. The optimal $\widehat{\boldsymbol{\theta}}_{\boldsymbol{N}}$ is the one that minimises the squared prediction error, where the prediction error is $\boldsymbol{\varepsilon} = \mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}$ in the linear regression case.

On the other hand, the *Bayesian Minimum Risk* (MR) *Estimate* is the one that makes more use of prior information. It is necessary to specify the distribution of noise $p(\mathbf{v})$, the distribution of the parameter vector $p(\boldsymbol{\theta})$ and the cost $C(\boldsymbol{\theta}, \boldsymbol{\theta}_{true})$ to be minimised.

A.4 Example

This example is based on the example of (Ninness, 2009) and (Ninness and Henriksen, 2010).

A.4.1 Experiment

The data generating process (true plant) is $y_n = \frac{0.2}{q-0.8}u_n + v_n,\ n = 1..N$, so the true parameter vector is $\boldsymbol{\theta}^T = (\theta_1, \theta_2) = (0.2, -0.8)$. The experiment consists of only $N = 20$ samples of the excitation signal $\{u_n\}_{n=0}^{N-1}$,

$$u_n = \begin{cases} 1, & n \leq 10 \\ 0, & n > 0 \end{cases}$$

The measurement noise sequence $\{v_n\}_{n=0}^{N-1}$ is i.i.d. uniform with zero mean and variance $\sigma^2 = E[v_n^2] = 0.01$. In the uniform distribution $\mathcal{U}(a, b)$ the mean value is given by $\frac{b+a}{2}$, so to have zero mean we need that $a = -b$. And the variance σ^2 is given by $\frac{(b-a)^2}{12}$ so to have variance equal to 0.01 we need $b = \frac{1}{2}\sqrt{12\sigma^2} = 0.1732$ (Casella and Berger, 2002, p99). Next figure show the system output:

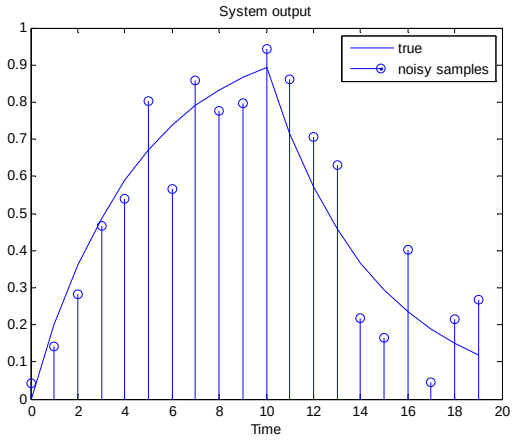


Fig. A.1. Experiment

A.4.2 Maximum Likelihood Point Estimation

The prior information is:

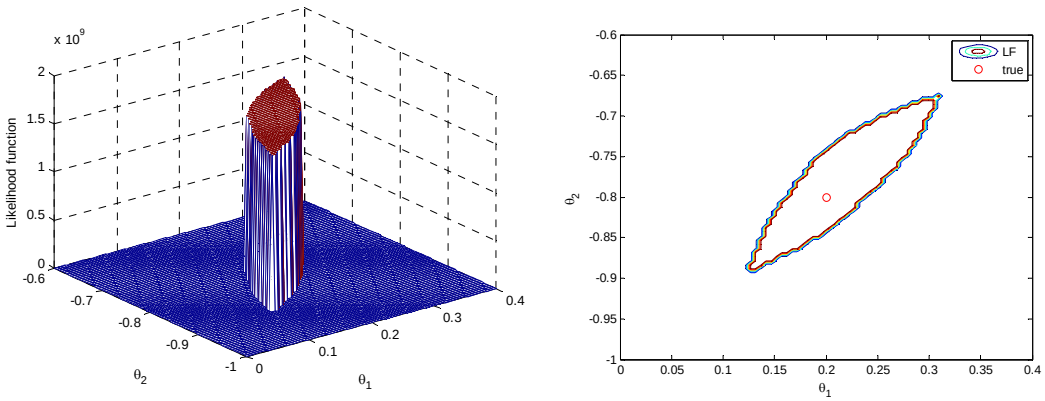
About noise: We *know* the noise pdf. It is i.i.d uniform with zero mean and variance 0.01.

About plant: We *do not know* the parameter vector pdf. Therefore, we *assume* it is constant.

About cost: We *do not know* which is the cost of selecting $\widehat{\boldsymbol{\theta}}_N \neq \boldsymbol{\theta}_{true}$ but we *do not care* about it.

The optimal solution is the one that maximizes the likelihood function, $\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{y})$.

Result: Next figures show the likelihood function and the true parameter vector:



(a) **(b)**

Fig. A.2. MLE: Likelihood function

Comments:
The true parameter vector is effectively at the top of the likelihood function, but we cannot "isolate" it due to the form of the uniform distribution. So, this optimization problem presents no unique solution.

A.4.3 Least Squares Point Estimation

The prior information is:

About noise: We *do not know* the noise pdf. Therefore, we *assume* it is i.i.d normal with zero mean and unit variance.

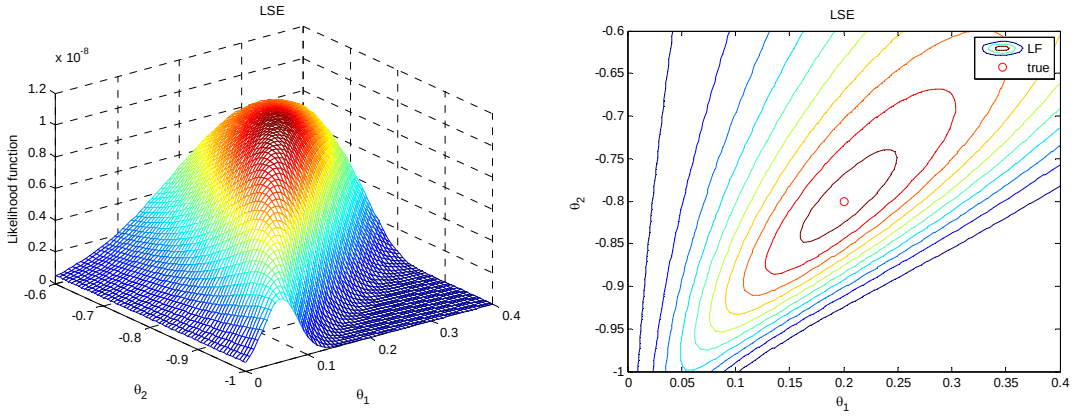
About plant: We *do not know* the parameter vector pdf. Therefore, we *assume* it is constant.

About cost: We *do not know* which is the cost of selecting $\widehat{\boldsymbol{\theta}}_N \neq \boldsymbol{\theta}_{true}$ but we *do not care* about it.

In short, no prior information is required.

The optimal solution is the one that minimizes the square of the prediction errors, $\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}\|_2^2$. But in this example we are going to find the solution by maximizing the likelihood function obtained with the assumption of normal noise, $\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{y})$.

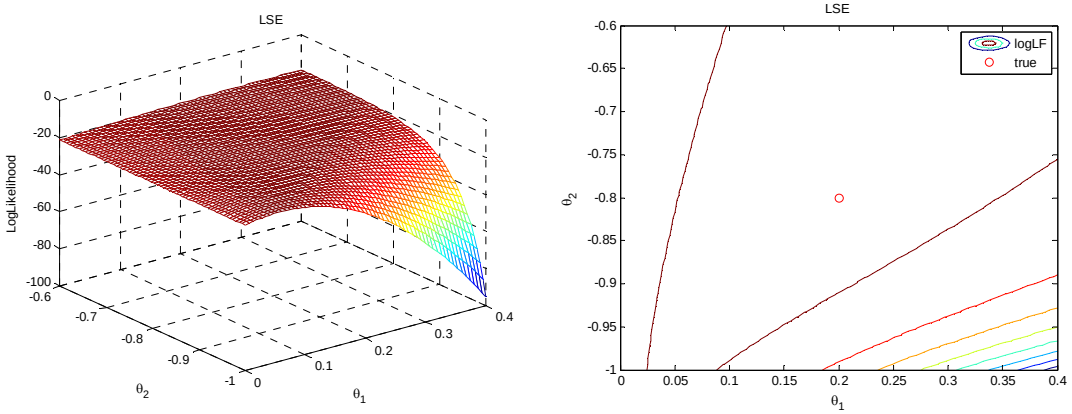
Results: Next figures show the likelihood function and the true parameter vector:



(a) **(b)**

Fig. A.3. LSE: Likelihood function

And the following ones the log-likelihood function and the true parameter vector:



(a) **(b)**

Fig. A.4. LSE: Log-likelihood function

The LS estimate obtained numerically is $\widehat{\boldsymbol{\theta}}_{LS}^T = (0.1898, -0.8169)$. This result can be improved by taking denser parameter vectors.

A.4.4 Weighted Least Squares Point Estimation (Markov Point Estimation)

The prior information is:

About noise: We *do not know* the noise pdf but we *know* its mean μ and variance σ^2. Therefore, we *assume* it is i.i.d normal $\mathcal{N}(\mu, \sigma^2\mathbf{I})$.

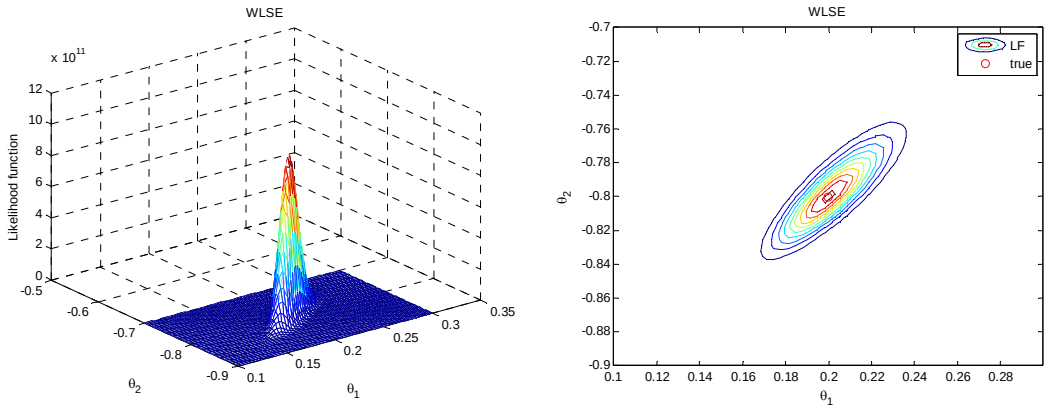
About plant: We *do not know* the parameter vector pdf. Therefore, we *assume* it is constant.

About cost: We *do not know* which is the cost of selecting $\widehat{\boldsymbol{\theta}}_N \neq \boldsymbol{\theta}_{true}$ but we *do not care* about it.

In short, only information about the noise mean and variance is required.

The optimal solution is again the one that minimizes the mean square of the prediction errors, $\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}\|_2^2$. And, again, in this example we are going to find the solution by maximizing the likelihood function obtained with the assumption of normal noise, $\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{y})$.

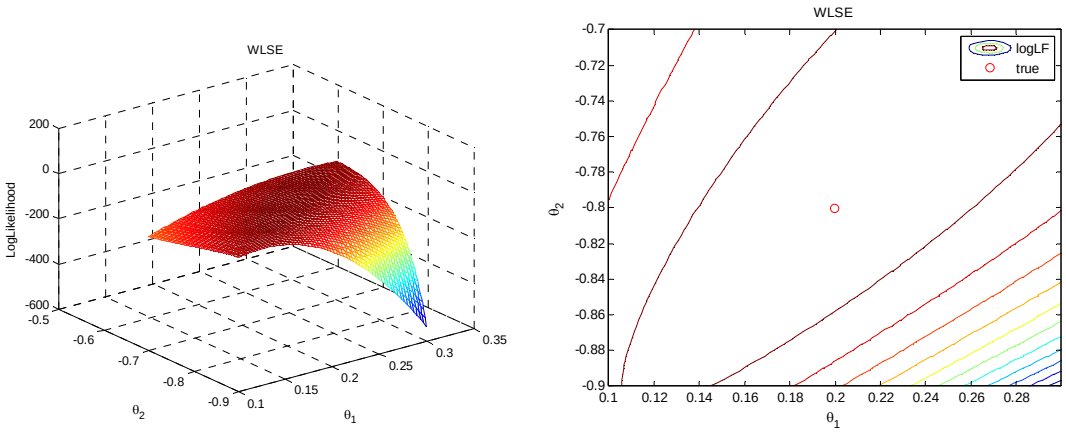
Results: Next figures show the likelihood function and the true parameter vector:



(a) **(b)**

Fig. A.5. WLSE: Likelihood function

And the following ones the log-likelihood function and the true parameter vector:



(a) **(b)**

Fig. A.6. WLSE: Log-likelihood function

The WLS estimate obtained numerically is $\widehat{\boldsymbol{\theta}}_{WLS}^{T} = (0.2017, -0.7983)$. If we had taken the same values for the parameter vector than in the LS case, the result would be equal to the LS.

Compared to the LS, here the likelihood function is spikier. In other words, the uncertainty around the estimate is smaller. This is sensible, since now we have introduced more prior information.

Comments:
Here we have forced the likelihood function be equal to

$$l(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{(2\pi|\mathbf{R}|)^{N/2}} \exp\left(-\frac{\boldsymbol{\varepsilon}^T \mathbf{R}^{-1} \boldsymbol{\varepsilon}}{2}\right)$$

where $\mathbf{R} = E[\mathbf{v}\mathbf{v}^T]$. This choice implies that we are consciously neglecting some prior knowledge about the noise. We *know* the noise is uniform but we prefer to forget this fact and *assume* that the noise is Gaussian. This way, the likelihood function presents one maximum value.

About the name WLSE. If the weighting matrix $\mathbf{R}$ is the covariance of the noise, $\mathbf{R} = E[\mathbf{v}\mathbf{v}^T]$, one speak of "Markov estimator" or "best linear unbiased estimator (BLUE)". If $\mathbf{R}$ is an arbitrary positive definite matrix, then one speaks of "weighted least squares estimator".

To optimize the likelihood $l(\mathbf{y}|\boldsymbol{\theta})$ is the same to optimise the log-likelihood $\log l(\mathbf{y}|\boldsymbol{\theta}) = ct - \frac{\boldsymbol{\varepsilon}^T \mathbf{R}^{-1} \boldsymbol{\varepsilon}}{2}$. Therefore, the cost function to minimise is a quadratic one, $J_{WLS} = \boldsymbol{\varepsilon}^T \mathbf{R}^{-1} \boldsymbol{\varepsilon}$. This explains the name "weighted least squares".

A.4.5 Bayesian estimation

In the Bayesian approach, the Bayes' rule provides a way to combine the prior knowledge (about the system and the measurement noise) with the observations of the system,

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})} \quad \textbf{(108)}$$

where $p(\boldsymbol{\theta}|\mathbf{y})$ is the posterior distribution of the parameter vector $\boldsymbol{\theta}$, $p(\mathbf{y}|\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\mathbf{y})$ is the likelihood function, $p(\boldsymbol{\theta})$ is the prior distribution of the parameter vector $\boldsymbol{\theta}$, and $p(\mathbf{y})$ is a normalising constant $p(\mathbf{y})$,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad \textbf{(109)}$$

Note that in the Bayesian approach the parameter vector $\boldsymbol{\theta}$ is viewed as a random variable. The prior knowledge about the system is contained in $p(\boldsymbol{\theta})$, the prior

knowledge about the measurement noise v is contained in the type of pdf $p_v(\cdot)$ used to construct the likelihood function, and the posterior knowledge due to the observations **y** is contained in the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$.

In this context, the Bayesian maximum a posteriori (MAP) estimate is:

$$\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y}) \tag{110}$$

The posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is the *joint* distribution of all the parameters $\theta_i,\ i = 1..d$. If we want to compute the *marginal* distribution of a particular parameter θ_i, we need to solve the following integral,

$$p(\theta_i|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})\, d\theta_1, \ldots, d\theta_{i-1}, d\theta_{i+1}, \ldots, d\theta_d \tag{111}$$

This integral can be solved numerically only in the simplest cases. As the number of parameters increases one can evaluate it by using Markov Chain Monte Carlo (MCMC) techniques, such as the Metropolis-Hastings sampler. The idea is to construct an ergodic Markov chain with invariant distribution equal to the desired posterior. See Appendix C.

This approach is also interesting because error bounds on estimates are derived from the sampled posterior and thus they do not rely on assumptions of N being large.

A.4.6 Maximum a Posteriori Bayesian Estimation

The prior information is:

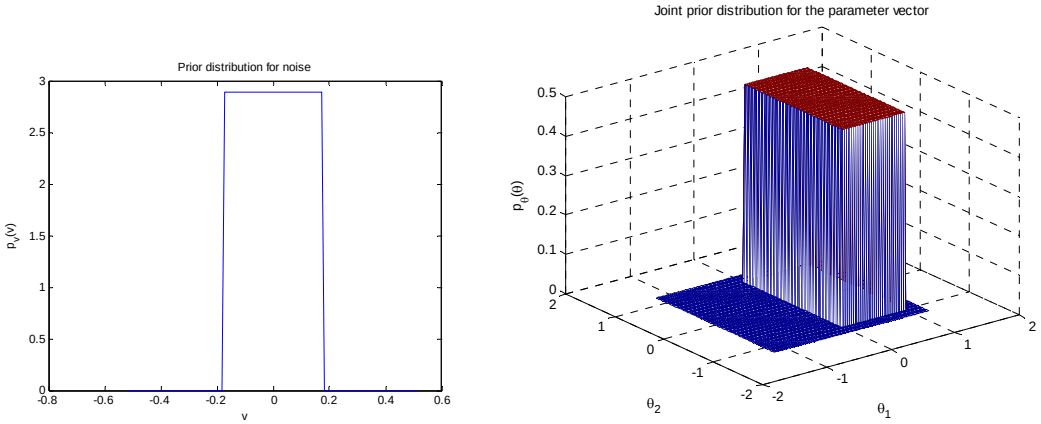
About noise: We *know* the noise pdf $p_v(v)$. It is i.i.d. uniform with zero mean and variance 0.01.

About plant: We *know* the parameter vector pdf $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. In our case, since the plant is stable we know that the parameter in the denominator is such that $|\theta_2| \leq 1$, so we assume that the marginal distribution $p_{\theta_2}(\theta_2)$ is uniform between -1 and 1. And, since the gain is positive, we assume that the parameter on the numerator is $\theta_1 > 0$, and so we take the marginal distribution $p_{\theta_1}(\theta_1)$ uniform between 0 and 1. As θ_1 and θ_2 are independent, we can construct the joint distribution by simply making $p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = p_{\theta_1}(\theta_1)\, p_{\theta_2}(\theta_2)$.

About cost: We *do not know* which is the cost of select $\widehat{\boldsymbol{\theta}}_N \neq \boldsymbol{\theta}_{true}$ but we *do not care* about it.

The optimal solution is again the one that maximizes the joint posterior distribution of the parameters, $\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y})$.

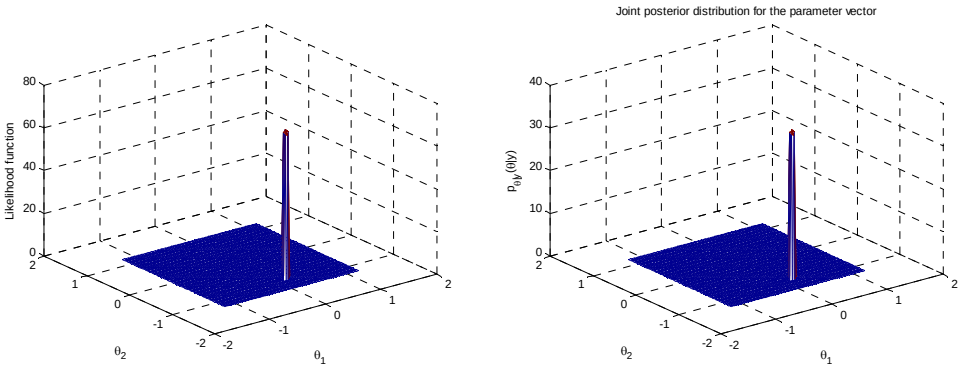
Results: Next figures show the prior knowledge, namely, the noise pdf $p_v(v)$ and the parameter vector prior pdf $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:



(a) **(b)**

Fig. A.7. Prior distributions

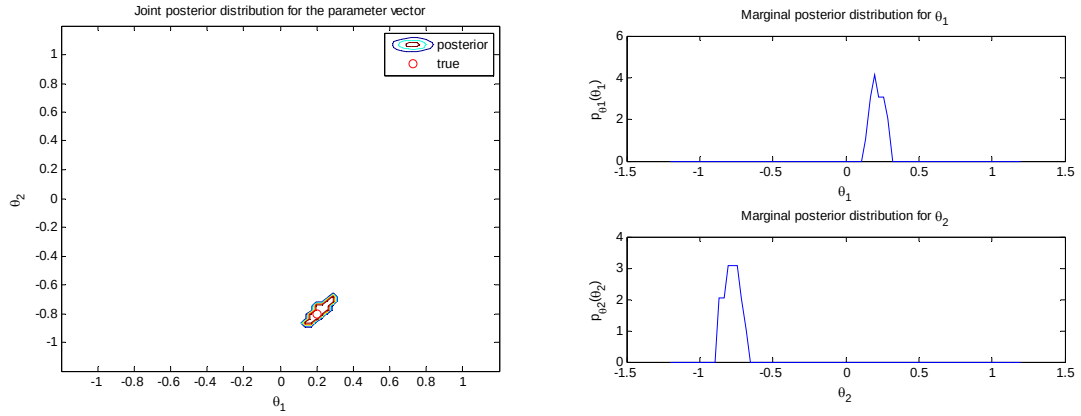
Now, we show the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ and the resulting parameter vector posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$:



(a) **(b)**

Fig. A.8. Likelihood function and posterior distribution

Finally, next figures show the contour plot of the joint posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ and the marginal posterior distributions $p_{\theta_1}(\theta_1|\mathbf{y})$ and $p_{\theta_2}(\theta_2|\mathbf{y})$. These have been obtained by a trapezoidal approximation of the marginalisation integral.



(a) **(b)**

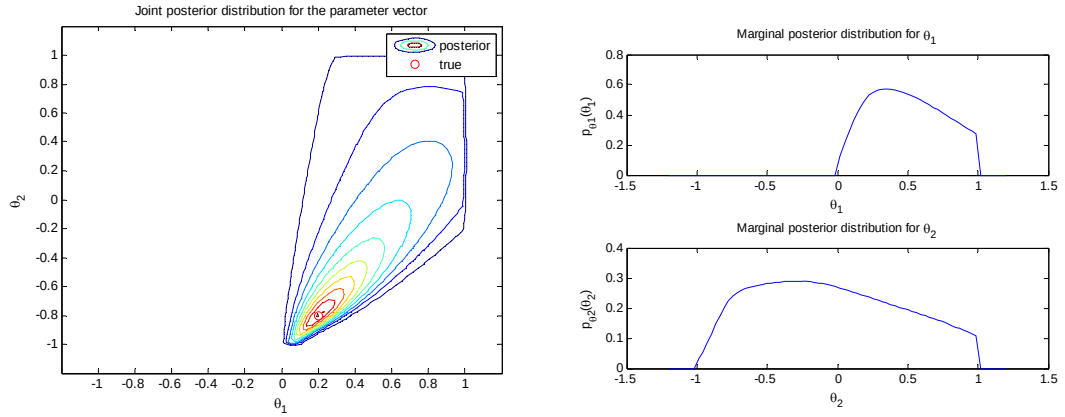
Fig. A.9. Joint posterior distribution and marginal posteriors distributions

The direct maximization of $p(\boldsymbol{\theta}|\mathbf{y})$ does not give a unique value due to the uniform distribution.

Using a normal distribution as prior noise distribution:

To circumvent this problem we can take a normal distribution for the noise $p_v(v)$ (even though we know it is uniform). This way, the posterior will exhibit a unique maximum.

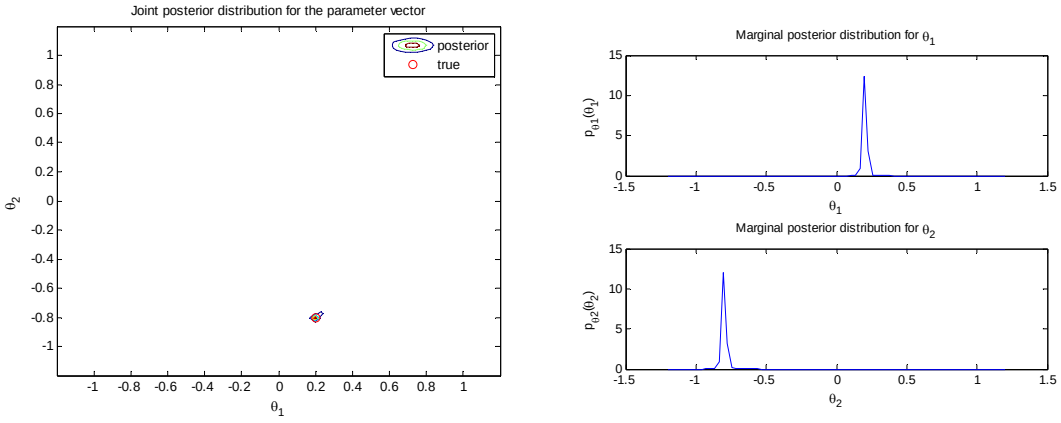
Next figures show the posterior contour and marginal distributions for the case of $v \sim \mathcal{N}(0,1)$. The parameter vector that maximises the posterior is $\widehat{\boldsymbol{\theta}}_{MAP}^{T} = (0.1975, -0.8051)$, and the one that maximizes each of the marginal separately is $\widehat{\boldsymbol{\theta}}_{MAP}^{T} = (0.3494, -0.2886)$.



(a) **(b)**

Fig. A.10. Posterior contour and posterior marginal distributions for variance 1

And finally next figures show the posterior contour and marginal distributions for the case of $v \sim \mathcal{N}(0,0.01)$. The parameter vector that maximizes the posterior is $\widehat{\boldsymbol{\theta}}_{MAP}^{T} = (0.1975, -0.8051)$, and the one that maximises each of the marginal separately is the same, $\widehat{\boldsymbol{\theta}}_{MAP}^{T} = (0.1975, -0.8051)$.



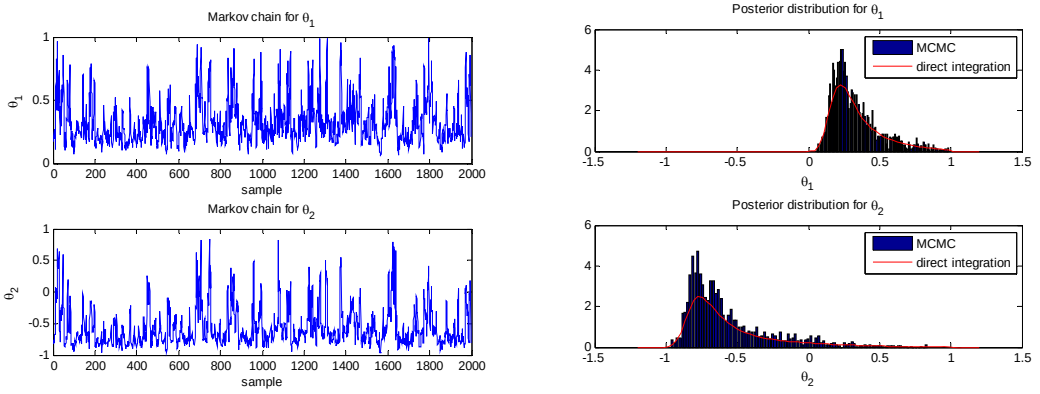
(a) **(b)**

Fig. A.11. Posterior contour and posterior marginal distributions for variance 0.01

Computation of the posterior marginal densities via the Metropolis-Hasting algorithm:

Another way to compute the posterior marginal distributions $p_{\theta_1}(\theta_1|\mathbf{y})$ and $p_{\theta_2}(\theta_2|\mathbf{y})$ is to estimate them by a Markov Chain Monte Carlo (MCMC) sampler. In this example we have implemented the Metropolis-Hastings algorithm with a standard deviation of 0.2 for the random walk process. The underlying distribution for the measurement noise is $p_v \backsim \mathcal{N}(0,0.25)$.

Next figures show the Markov chains and the obtained posterior marginal for both parameters:



(a) **(b)**

Fig. A.12. Markov chains and posterior marginal distributions

APPENDIX B

Orthonormal Bases in System Identification

In the robust identification field it is common practice to use nominal models with fixed denominator. This avoids pole estimation (which is sensitive to measurement noise), separates the nominal model (*B, F; G*) estimation from the noise model (*C, D; H*) estimation (thus, eliminating this source of bias), and allows introducing prior knowledge regarding the plant modes. Basis functions from Laguerre and Kautz expansion series as well as generalized orthonormal bases (GOB) are the most used in the identification of linear models. Polynomials, radial basis functions and wavelets can be used in the nonlinear case.

B.1 Introduction

B.1.1 General input/output models

When selecting the (nominal) model structure, one has to compromise between *parsimony* (simplicity) and enough *flexibility* (to contain the structure that best fits the true system). A common choice is to consider a discrete time description of the system:

$$y_n = G(q)u_n + H(q)v_n \quad , \quad n = 0, \dots, N-1 \qquad \textbf{(112)}$$

where q is the shift operator, $q^{-1}u_k = u_{k-1}$, the dynamic model $G(q)$ and the noise model $H(q)$ are rational functions in q, and $\{y_n\}_{n=0}^{N-1}$, $\{u_n\}_{n=0}^{N-1}$, $\{v_n\}_{n=0}^{N-1}$ are the output, input, and noise sequences respectively.

A general input-output model structure is

$$A(q)y_n = \frac{B(q)}{F(q)}u_{n-k} + \frac{C(q)}{D(q)}v_n \qquad \textbf{(113)}$$

where k is the number of time delays and the polynomials are defined as in (Ljung, 1995): $A(q) = 1 + a_1q^{-1} + \cdots + a_{n_a}q^{-n_a}$, $B(q) = b_0 + b_1q^{-1} + \cdots + b_{n_b}q^{-n_b}$, $C(q) = 1 + c_1q^{-1} + \cdots + c_{n_c}q^{-n_c}$, $D(q) = 1 + d_1q^{-1} + \cdots + d_{n_d}q^{-n_d}$ y $F(q) = 1 + f_1q^{-1} + \cdots + f_{n_f}q^{-n_f}$.

a. Main structures for linear models

From the general structure (113), different sub-structures are derived. Next table lists the most important parameterizations.

Structure		Polynomials	Input/output model
FIR	*Finite Impulse Response*	B	$y_n = B(q)u_{n-k} + v_n$
ARX	*Auto-Regressive with eXogenous input*	A, B	$A(q)y_n = B(q)u_{n-k} + v_n$
ARMAX	*Auto-Regressive Moving Average with eXogenous input*	A, B, C	$A(q)y_n = B(q)u_{n-k} + C(q)v_n$
AR-ARX	*Auto-Regressive - Auto-Regressive with eXogenous input*	A, B, D	$A(q)y_n = B(q)u_{n-k} + \frac{1}{D(q)}v_n$
OE	*Output Error*	B, F	$y_n = \frac{B(q)}{F(q)}u_{n-k} + v_n$
BJ	*Box Jenkins*	B, C, D, F	$y_n = \frac{B(q)}{F(q)}u_{n-k} + \frac{C(q)}{D(q)}v_n$

Table B.1. Common structures for linear models

Regarding the notation, an OE model with 4 *f*-parameters, 3 *b*-parameters, and 2 delays is denoted as an OE(3,4,2)-model. The numbers are presented in "alphabetical order". This is the same convention as in the *System Identification Toolbox* for MATLAB (Ljung, 1995).

Three main identification problems are related to the model structure (113) (Verhaegen and Verdult, 2007). The most general is the one used in Prediction Error Methods (PEM), where the objective is to obtain the one-step prediction $\hat{y}_{n+1}$

$$\hat{y}_{n+1} = a\hat{y}_n + bu_n + l\,(y_n - \hat{y}_n), \quad n = 0, \ldots, N-1 \qquad \textbf{(114)}$$

This is a difficult problem to solve since it involves non-linear optimization but it can be simplified by means the appropriate selection of parameter l.

The selection of $l = 0$ leads to the *simulation problem*, where the output error model is used,

$$\hat{y}_{n+1} = a\hat{y}_n + bu_n, \quad n = 0, \dots, N-1 \tag{115}$$

and again it must be solved by nonlinear optimization techniques.

The selection of $l = a$ leads to the *predictor problem*,

$$y_{n+1} = ay_n + bu_n, \quad n = 0, \dots, N-1 \tag{116}$$

which can be solved by linear least square optimization.

b. ARX and FIR models

In standard system identification ARX and FIR models are extensively used (Tjärnström, 2002). They can be parameterized by means a row regression vector $\boldsymbol{\varphi}_n^T$ as $G(q, \boldsymbol{\theta})u_n = \boldsymbol{\varphi}_n^T\boldsymbol{\theta}$ and thus easily identified by least squares techniques. For the second order case, the parameterization is

$$y_n = \boldsymbol{\varphi}_n^T\boldsymbol{\theta} = (u_{n-1} \;\; u_n \;\; -y_{n-2} \;\; -y_{n-1}) \begin{pmatrix} b_1 \\ b_0 \\ a_2 \\ a_1 \end{pmatrix}, \quad n = 0, \dots, N-1 \tag{117}$$

These models are fast and easy to estimate, no local minima exist, and they are capable of approximating any linear system arbitrarily well, provided that the model order is high enough (this property is useful for model validation purposes). These models are also a useful modelling tool, i.e., one can estimate a high order model and then reduce it to an appropriate order by using some model reduction technique, e.g. based on Hankel norm approximation (see the nearly optimal algorithm in Chapter 3).

Example B.1. ARX and FIR models for the Landau benchmark

Let us illustrate the performance of FIR and ARX models by means of a benchmark example. Data correspond to the measurements of an active suspension system (Landau benchmark example (Landau *et al.*, 2003)) and are available from the website of the *Laboratoire d'Automatique* (EPFL) in Lausanne.

Consider the first experiment (**`data_prim1.mat`**) over the primary path of the active suspension system. The PRBS (pseudo random binary signal) input *u* consists of *N* = 8000 samples of a 10-bit shift register with a clock frequency $f_s = 400\,Hz$.

Fig. B.1(a) shows the spectral analysis between the input and output sequences (we have used the MATLAB function **`psd`**, with 1024 points for the FFT, a 512-Hanning

window and 256-overlap). For these data, Landau has proposed an OE(8,12,0)-model (**`primary_model.mat`**). This model is also shown in Fig. B.1(a).

For the same measurement data, Fig. B.1(b) shows a FIR model of order 50 computed by means the **`gob`** function developed for this thesis. And Fig. B.1(c) and Fig. B.1(d) show an ARX model of order 50 and an AR-ARX model of order 13, respectively. These latter models have been computed with the **`arx`** and **`pem`** functions of the *System Identification Toolbox*.

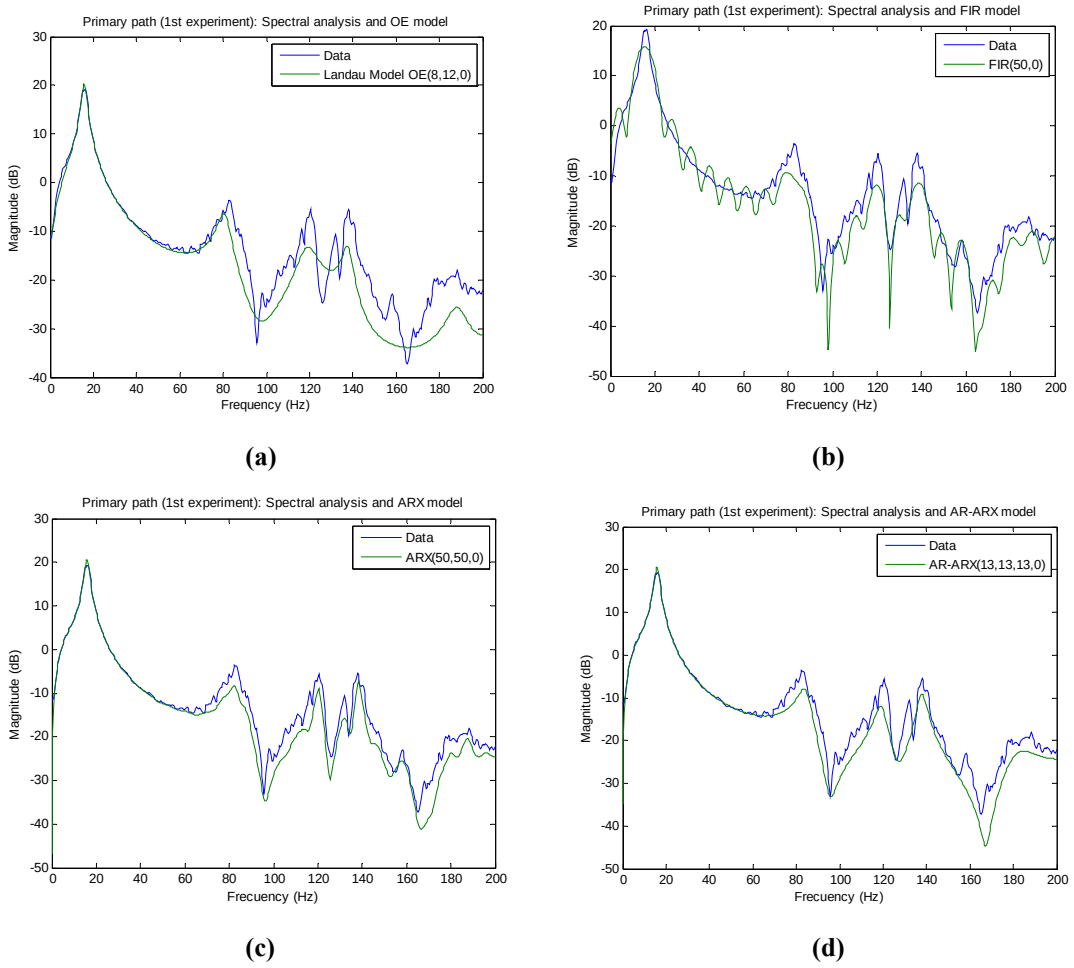


Fig. B.1. Landau benchmark. Spectral analysis of measurement data and (a) OE(8,12,0) model, (b) FIR(50,0) model, (c) ARX(50,50,0) model, and (d) AR-ARX(13,13,13,0) model

High order FIR and ARX models can approximate very well the shape of any frequency response. However, note that this includes the behaviour near the Nyquist frequency which may present an important degree of aliasing, depending to the experiment design. ■

B.1.2 Models for robust identification

Models in Table B.1 (except for the FIR case) are rational models, that is, one has to estimate both the numerator coefficients ($B(q)$, $C(q)$) and the denominator coefficients

$(A(q), F(q), D(q))$. However, in the robust identification field it is preferred to work with fixed pole structures. This is so because rational models present several drawbacks that make them inappropriate for uncertain systems modeling. The main drawbacks are listed below. For a throughout discussion see (Gustafsson and Mäkilä, 1994).

a. Drawbacks of rational models

Pole sensitivity: In uncertain systems, rational models pole estimation may present too much sensitivity to noise. As (Gustafsson and Mäkilä, 1994) point out, identification using a general ARMAX model structure "cannot be guaranteed to result in a *stable* model even if the system is stable". On the other hand, fixed-pole models are guaranteed to produce stable models since the poles location is decided by us.

Numerical issues: Moreover, on the contrary to rational models, the estimation procedure in fixed-pole models is *well-conditioned* and thus robust against measurement data produced by systems outside the identification set (Gustafsson and Mäkilä, 2001), (Gustafsson and Mäkilä, 1996).

Bias due to the noise model: In rational models, the model parameters generally do not appear linearly, and so estimation of them involves the numerical solution of a nonlinear optimization problem. This difficulty can be overcome by recasting the problem in a linear regression form, but in this case the parameters to be estimated affect both the dynamic model $G(q)$ and the noise model $H(q)$ (Ninness and Gustafsson, 1997). This can cause estimates of them to be *biased* (Wahlberg and Ljung, 1986). In the fixed pole structure, the parameter vector $\boldsymbol{\theta} = (\theta_0 \quad \dots \quad \theta_{d-1})^T$ parameterizes only the model for the dynamics, and so $\widehat{\boldsymbol{\theta}}$ is not biased by the noise model $H(q)$ estimate.

Variance estimate: (Ninness and Gustafsson, 1997) In rational models it is difficult to evaluate the variance of the estimated model except in an asymptotic sense, for $N \to \infty$. In the fixed pole structure, since $\boldsymbol{\theta}$ appears linearly, its least squares estimate $\widehat{\boldsymbol{\theta}}$ can be found in closed form and is linear in y_n so that if u_n is not noise corrupted, then finite data variances for $\widehat{\boldsymbol{\theta}}$ can be obtained.

b. Fixed pole models

Thus, in most robust identification problems, the poles in ARX or OE models (e.g. the $A(q)$ and $F(q)$ roots) are not estimated but instead their number and value are *a priori* fixed given the approximate knowledge we have about the system time constants (Ninness and Gustafsson, 1997).

The system is then expressed by means an structure linear in the parameters

$$y_n = \left(\sum_{i=0}^{d-1} \theta_i B_i(q) \right) u_n + v_n \qquad \textbf{(118)}$$

where d is the model order, $\{\theta_i\}_{i=0}^{d-1}$ are the real-valued parameters to be estimated, $\{u_n\}_{n=0}^{N-1}$ is the observed input, $\{y_n\}_{n=0}^{N-1}$ is the observed output, and $\{B_i(q)\}_{i=0}^{d-1}$ is a set of transfer functions rational in the forward shift operator q. Usually the noise sequence $\{v_n\}_{n=0}^{N-1}$ is assumed to be a zero mean i.i.d. (independent identically distributed) Gaussian random sequence (with identity covariance).

c. Choice of poles

The quality of the estimate depends on the selection of the poles in the $\{B_i(q)\}_{i=0}^{d-1}$ functions. The simplest pole choice is to consider FIR-type models in which all poles are located at the origin, i.e. $B_i(q) = q^{-i}$. As it has been shown in the Example B.1, the FIR model is a general model for any stable system but the order d may need to be very large to provide an accurate approximation to the underlying dynamics that have generated the observed data. For example, if the true dynamics have a slow pole, then the model order d will need to be very large for the model structure (118) to provide an accurate approximation to the true dynamics.

To overcome this problem, an alternative strategy is to instead take $B_i(q) = \frac{1}{q-\xi_i}$ where the poles $\{\xi_i\}_{i=0}^{d-1}$ are chosen according to *a priori* knowledge of the dominant modes of the system. For instance, if we know that the system presents a slow pole, we may choose at least one of the $\{\xi_i\}_{i=0}^{d-1}$ near 1. If the poles are well selected the model order can be relatively small, otherwise the estimate could be poor (Ninness and Gustafsson, 1997). Some authors give some guidelines for the pole selection, see e.g. (Gustafsson and Mäkilä, 2001), where fast and slow dynamics are combined to model the behaviour of a distillation column.

Although the selection of the $\{B_i(q)\}_{i=0}^{d-1}$ functions is free, the usual practice is to use functions that constitute an orthonormal basis for some expansion series.

d. Interest of orthonormal basis functions

Let $\mathcal{H}_2(\mathcal{T})$ be the Hardy space of functions that are square integrable on the unit circle $\mathcal{T}$, and analytic outside the unit disk (roughly speaking, $\mathcal{H}_2(\mathcal{T})$ is the space of all stable, causal, discrete-time transfer functions). The orthogonality condition in the $\mathcal{H}_2(\mathcal{T})$ space is the following

$$\langle B_l, B_k \rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} B_l(e^{j\omega})\overline{B_k(e^{j\omega})}d\omega = \delta_{lk} \qquad \textbf{(119)}$$

where δ_{lk} is the Kronecker delta, $\delta_{lk} = 1, l = k$ and $\delta_{lk} = 0, l \neq k$.

Note that the functions that parameterize FIR models, $B_i(q) = q^{-i}$, form an orthonormal basis in the unit circle $\mathcal{T}$. Thus the FIR structure can be interpreted as the Taylor (trigonometric) expansion of the ARX structure being the θ_i parameters the real coefficients of the series. Below, $B(q^{-1})$ is the polynomial in the structure (113).

$$B(q^{-1}) = \theta_0 B_0(q^{-1}) + \cdots + \theta_d B_{d-1}(q^{-1}) = b_0 + b_1 q^{-1} + \cdots + b_{d-1} q^{-(d-1)} \quad \textbf{(120)}$$

This fact suggested the use of orthonormal basis functions $\{B_i(q)\}_{i=0}^{d-1}$ from more sophisticated expansion series, such as Laguerre or Kautz series (Wahlberg, 1991), (Wahlberg, 1994). This way the parameters $\{\theta_i\}_{i=0}^{d-1}$ to be estimated in (118) can be viewed as the real-valued expansion coefficients of the series. And the resulting model sets are spanned by fixed pole orthonormal bases.

Several authors have studied the properties of orthonormal model structures. See for instance (Ninness, Hjalmarsson, and Gustafsson, 1999) and (Gustafsson and Mäkilä, 2001). These structures are interesting because they improve the numerical condition in the coefficients estimation and provide parameterizations that allow decreased variance error while still minimising bias error. They can be used to quantify the asymptotic variability of the estimates as well. Moreover, an orthonormal structure is, under a linear parameter space transform, equivalent to any other equivalently flexible orthonormal structure with the same fixed poles. For a throughout analysis see (Gómez, 1998).

B.2 Main orthonormal bases for robust identification

Laguerre models (in discrete time) were the first proposal to model systems with real poles (Wahlberg, 1991). Later on, for the resonant systems case, Kautz models were proposed (Wahlberg, 1994). Finally, both models were combined in the so called Generalized Orthonormal Basis (GOB) (Heurbeger, Van den Hof, and Bosgra, 1995), (Ninness and Gustafsson, 1997). Continuous time versions appeared in (Akçay and Ninness, 1999).

B.2.1 Laguerre models

a. Discrete time

Estimation using these models was studied in detail in (Wahlberg, 1991).

The functions that form the basis in the Laguerre model are the discrete Laguerre polynomials of order *i*,

$$B_i(q,\xi) = \frac{\sqrt{T_s}\sqrt{1-\xi^2}}{q-\xi}\left(\frac{1-\xi q}{q-\xi}\right)^i \quad , \quad |\xi| < 1, \quad i = 0,1,\ldots,d-1 \quad \textbf{(121)}$$

where T_s is the sampling time, d is the system order (dimension of the parameter vector $\boldsymbol{\theta}$) and ξ the Laguerre parameter, which is selected from the prior knowledge about the system or it is adjusted during the identification procedure. With ξ selected, the model (118) is a fixed-pole ARX model with multiple poles at ξ.

The state-space description of these functions is $\mathbf{A} \in \mathbb{R}^{i\times i}$, $\mathbf{b} \in \mathbb{R}^{i\times 1}$, $\mathbf{C} \in \mathbb{R}^{i\times i}$, $\mathbf{d} \in \mathbb{R}^{i\times 1}$:

$$\mathbf{A} = \begin{bmatrix} \xi & 0 & \dots & \dots & 0 \\ a & \xi & 0 & & 0 \\ (-\xi)a & a & \xi & & \vdots \\ \vdots & & & & 0 \\ (-\xi)^{i-1}a & \dots & (-\xi)a & a & \xi \end{bmatrix} \quad \mathbf{b} = \begin{pmatrix} 1 \\ (-\xi) \\ (-\xi)^2 \\ \vdots \\ (-\xi)^{i-1} \end{pmatrix} \qquad \textbf{(122)}$$

$$\mathbf{C} = \sqrt{aT_s} \times \mathbf{I_{i\times i}} \qquad \mathbf{d} = \mathbf{0_{i\times 1}}$$

where $a = 1 - \xi^2$.

b. Continuous time

In the continuous time, the functions that form the Laguerre model are:

$$B_i(s,\xi) = \frac{\sqrt{2\xi}(s-\xi)^{i-1}}{(s+\xi)^i} \quad , \quad \mathrm{Re}(\xi) < 0 \qquad \textbf{(123)}$$

And the state-space description of the i-th function is $\mathbf{A} \in \mathbb{R}^{i\times i}$, $\mathbf{b} \in \mathbb{R}^{i\times 1}$, $\mathbf{C} \in \mathbb{R}^{i\times i}$, $\mathbf{d} \in \mathbb{R}^{i\times 1}$.

$$\mathbf{A} = \begin{bmatrix} \xi & \mathbf{0_{1\times(i-1)}} \\ \begin{pmatrix} 2 \\ 2 \\ \vdots \\ 2 \end{pmatrix} & \begin{pmatrix} \xi & 0 & \cdots & 0 \\ 2\xi & \xi & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 2\xi & \dots & 2\xi & \xi \end{pmatrix} \end{bmatrix} \quad \mathbf{b} = \begin{pmatrix} \sqrt{2\xi} \\ \mathbf{0_{(i-1)\times 1}} \end{pmatrix} \qquad \textbf{(124)}$$

$$\mathbf{C} = \begin{bmatrix} 1 & \mathbf{0_{1\times(i-1)}} \\ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} & \begin{pmatrix} \xi & 0 & \cdots & 0 \\ \xi & \xi & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \xi & \dots & \xi & \xi \end{pmatrix} \end{bmatrix} \qquad \mathbf{d} = \mathbf{0_{i\times 1}}$$

Example B.2. Laguerre models

We have implemented continuous time and discrete time Laguerre basis functions in a MATLAB function called **`gob.m`**. To check this code, we have used the experiments in (Reinelt, Garulli, and Ljung, 2002). For the first experiment, the results are shown in Fig. B.2.

The continuous time model is of order 4 and its pole is located at -0.2895 ($\xi = -0.2895$). The estimate parameter vector obtained via LSE optimization is $\widehat{\boldsymbol{\theta}}_c = (-8.6912,\ -0.6584,\ 1.0617,\ 0.1939)^T$.

The discrete time model is of 4th order as well and its pole is $\xi = \exp(-0.2895 \cdot T_s)$ where $\cdot\, T_s = 0.04s$. The estimate parameter vector obtained via LSE optimization is $\widehat{\boldsymbol{\theta}}_d = (7.4928,\ 0.6298,\ 1.0792,\ -0.2091)^T$.

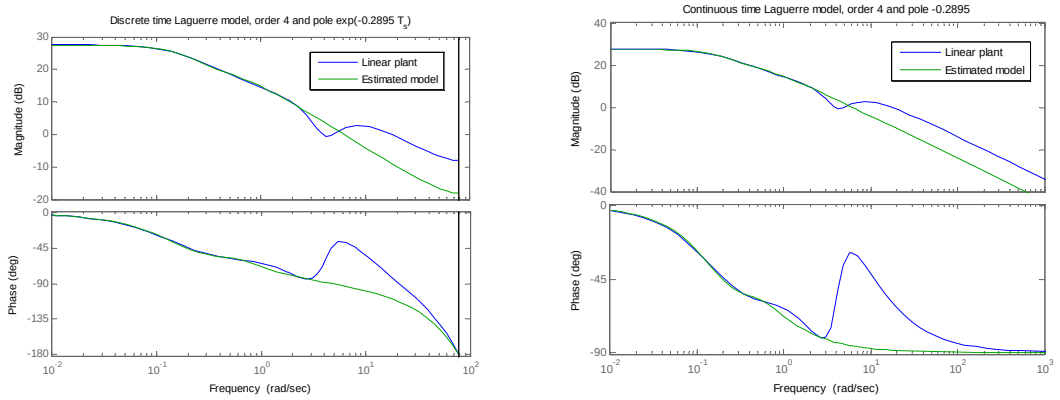


Fig. B.2. Discrete time and continuous time 4th order Laguerre models for the (Reinelt *et al*., 2002) plant

■

c. Properties

Assuming that $\mathcal{B}$ is the space of transfer functions that are discrete time, linear, causal, invariant in q, and BIBO (Bounded Input Bounded Output), the discrete time Laguerre functions $B_i(q) \in \mathcal{B}$ form an orthonormal basis in $\boldsymbol{\ell}_2$ (space of square sumable sequences). Moreover, functions (121) are *dense* in $\mathcal{B}$, i.e., the closure of the linear span of $\{B_i(q)\}_{i=1}^{\infty}$ is $\mathcal{B}$ (see Theorem 3 in (Gustafsson and Mäkilä, 1993)).

Finally, note that this type of model considers only one pole, ξ, and the multiplicity of this pole is the model order d.

B.2.2 Kautz models

a. Discrete time

If the system presents resonant poles, the so-called two-parameter Kautz model or just Kautz model can be used

$$B_i(q) = \begin{cases} \dfrac{\sqrt{(1-b^2)(1-c^2)}}{q^2 + b(c-1)q - c}\left(\dfrac{-cq^2 + b(c-1)q + 1}{q^2 + b(c-1)q - c}\right)^{\frac{(i-1)}{2}}, & i \text{ odd} \\ \dfrac{\sqrt{1-b^2}\cdot(q-b)}{q^2 + b(c-1)q - c}\left(\dfrac{-cq^2 + b(c-1)q + 1}{q^2 + b(c-1)q - c}\right)^{\frac{i}{2}}, & i \text{ even} \end{cases} \qquad \textbf{(125)}$$

where $|b| < 1$, $|c| < 1$, $i = 1,2,\ldots,d$. For an example of the application of Kautz basis functions to a flexible structure, see (Baldelli, Mazzaro, and Sánchez Peña, 2001).

Regarding the construction of the functions that form the basis in the Kautz expansion series, it is convenient to use a balanced minimal realization of the inner function $\frac{-cq^2+b(c-1)q+1}{q^2+b(c-1)q-c}$. In the present work the realization that we have implemented is the one of (Heurberger *et al*., 1995):

$$\mathrm{A} = \begin{bmatrix} b & \sqrt{1-b^2} \\ c\sqrt{1-b^2} & -bc \end{bmatrix} \qquad \mathrm{b} = \begin{pmatrix} 0 \\ \sqrt{1-c^2} \end{pmatrix} \qquad \textbf{(126)}$$

$$\mathrm{c} = \begin{pmatrix} \sqrt{(1-c^2)(1-b^2)} & -b\sqrt{1-c^2} \end{pmatrix} \quad d = -c$$

Example B.3. Kautz models

Let us illustrate the behavior of the Kautz models with the (Wahlberg, 1994) plant. The supposed unknown plant is $ZOH\left\{\frac{1}{s^2+0.2s+1}\right\}$ with a samplig time of $T_s = 0.5s$. The time domain experiment consists of exciting the plant with $N = 1024$ samples of a pseudorandom binary signal and collecting the corresponding output samples.

Fig. B.3 shows the frequency response of the discretized plant along with the responses of a 10th order Laguerre model with the pole located at 0.84 and a 2nd order Kautz model with parameters $b = 0.87$ and $c = -0.9$. The Kautz model obtained is:

$$G_{Kautz}(q) = 0.1389\frac{0.43589(q-0.87)}{(q^2-1.653q+0.9)} + 1.0727\frac{0.21492}{(q^2-1.653q+0.9)}$$

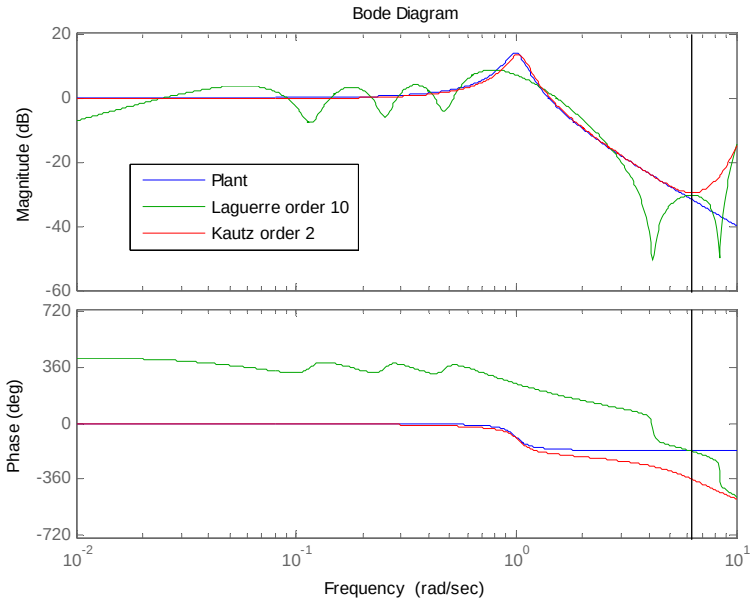


Fig. B.3. Discrete time 10th order Laguerre model and 2nd order Kauth model for the plant of (Wahlberg, 1994)

■

b. Continuous time

The expression of the two-parameter Kautz basis functions for the continuous-time case is (Wahlberg, 1991):

$$B_i(s) = \begin{cases} \dfrac{\sqrt{2c}\cdot s}{s^2+bs+c}\left(\dfrac{s^2-bs+c}{s^2+bs+c}\right)^{i-1}, & i \text{ odd} \\ \dfrac{\sqrt{2bc}}{s^2+bs+c}\left(\dfrac{s^2-bs+c}{s^2+bs+c}\right)^{i-1}, & i \text{ even} \end{cases} \quad \textbf{(127)}$$

where $b > 0$, $c > 0$, $i = 1,2,\dots,d$.

B.2.3 Generalized Orthonormal Basis (GOB)

A criticism to Laguerre and Kautz models from the previous sections is that they consist of only one pole (or one conjugate pair of poles) and the designer must increase the multiplicity in order to get better fit to data. For this reason, several authors such as (Heurberger, Van den Hof, and Bosgra, 1995) and (Ninness and Gustafsson, 1997) proposed a generalized model capable of combining different poles, complex conjugate or real, fast or slow, multiple or not, in a same structure. The result was the so-called Generalised Orthonormal Basis (GOB) and it accounts for FIR, Laguerre, and Kautz basis functions in a unified formulation.

a. Discrete time

Fort the case of real poles or poles in the origin, the functions of the generalized basis are:

$$B_i(q) = \left(\frac{\sqrt{1-|\xi_i|^2}}{q-\xi_i}\right)\prod_{k=0}^{i-1}\left(\frac{1-\bar{\xi}_k q}{q-\xi_k}\right) \quad \textbf{(128)}$$

where the poles $\{\xi_0, \xi_1, \dots, \xi_{d-1}\} \in \mathbb{D}$, $\mathbb{D} = \{z \in \mathbb{C}: |z| < 1\}$, can be different.

Note that if all poles are in the origin ($\xi_k = 0, \forall k$) then (128) is reduced to a FIR model structure, whereas the selection $\xi_k = \xi \in \mathbb{R}$, $|\xi| < 1$ corresponds to the Laguerre model.

The way (128) is defined does not allow to include complex poles, since then the coefficients should be complex, and consequently, the impulse response would be complex too, and thus it would be not useful to describe physical systems.

The solution is to use (128) to obtain the basis $B_i(q)$ corresponding to the complex pole ξ_i and the basis $B_{i+1}(q)$ corresponding to the conjugate complex pole $\xi_{i+1} = \overline{\xi_i}$. These functions are replaced in the model by a linear combination of them designed to maintain the orthonormality and make that the model impulse response be real,

$$\begin{aligned} B_i' &= \alpha B_i + \beta B_{i+1} \\ B_{i+1}' &= \alpha' B_i + \beta' B_{i+1} \end{aligned} \qquad \textbf{(129)}$$

The relation between the coefficients is (Ninness and Gustafsson, 1997), (Gómez, 1998):

$$\begin{pmatrix} \alpha' \\ \beta' \end{pmatrix} = \frac{1}{(\xi_i - \overline{\xi_i})\sqrt{1-\mu^2}} \begin{bmatrix} \xi_i & -1 \\ -\overline{\xi_i} & 1 \end{bmatrix} \begin{bmatrix} \mu & 1 \\ -1 & -\mu \end{bmatrix} \begin{bmatrix} 1 & 1 \\ \overline{\xi_i} & \xi_i \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad \textbf{(130)}$$

where $\mu \equiv \frac{\xi_i + \overline{\xi_i}}{1+|\xi_i|^2}$. If we choose $\alpha = -\beta = \frac{\sqrt{(1-\mu^2)(1+|\xi_i|^2)}}{\overline{\xi_i} - \xi_i}$ the result are the Kautz basis functions.

Example B.4. GOB model

Consider again the Landau benchmark of Example B.1. Fig. B.4 shows a GOB model containing different complex and real poles. The poles position is directly the ones of the OE(8,12,0) model.

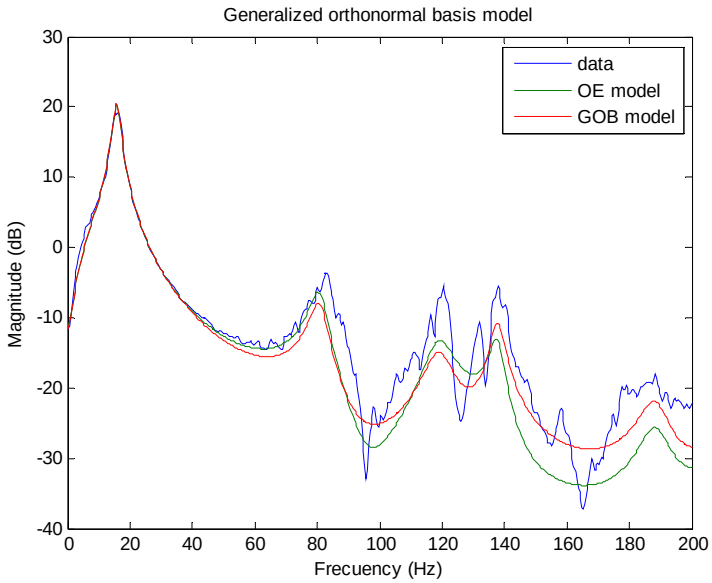


Fig. B.4. GOB modelo for the Landau bechmark

■

b. Continuous time

The generalization of the functions that constitute the basis is analogous to the discrete case,

$$B_i(s) = \frac{\sqrt{2\mathrm{Re}(\xi_i)}}{s+\xi_i}\prod_{k=0}^{i-1}\left(\frac{s-\overline{\xi_k}}{s+\xi_k}\right) \tag{131}$$

where the poles $\{\xi_0, \xi_1, \ldots, \xi_{d-1}\} \in \mathrm{LHP}$, $\mathrm{LHP} = \{s \in \mathbb{C}\colon \mathrm{Re}(s) < 0\}$, can be different. If we take all poles to be real and equal, $\xi_k = \xi \in \mathbb{R}$, (131) corresponds to the Laguerre model, and if we take complex conjugate multiple poles, $\xi_k = \xi \in \mathbb{C}$, then (131) corresponds to the two-parameter Kautz model.

In the continuos case, the functions that constitute the basis are orthonormal in $\mathcal{H}_2(\mathbb{C}_+)$ with respect to the inner product

$$\langle B_k, B_l\rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} B_k(j\omega)\overline{B_l(j\omega)}d\omega = \begin{cases}1, & k = l\\ 0, & k \neq l\end{cases}$$

c. Properties

The generalized orthonormal bases, in the discrete time case and in the continuous time case, are *complete* if their poles satisfy certain conditions.

Since the orthonormal parameterizations are used to approximate functions, let us determine what a good approximation is in the context of systems theory. Assume an element $f(s)$ of a normed linear space of functions $(\mathcal{X}, \|.\|_{\mathcal{X}})$. To obtain an arbitrary good approximation of $f(s)$ consists of obtaining an element $g(s) \in span\{B_i(s)\}_{i=0}^{d-1}$ such as $\|f - g\|_{\mathcal{X}} \leq \varepsilon$ for an arbitrary $\varepsilon > 0$ and for a d value sufficiently large. If the approximation is possible for any $\varepsilon > 0$ arbitrarily small, then one says that $span\{B_i(s)\}_{i\geq 0}$ is complete in $\mathcal{X}$.

B.3 Bases for block-oriented nonlinear models

Nonlinear systems can be parameterized by means basis functions as well. One of the most frequently studied classes of nonlinear models are the so-called block-oriented nonlinear models, which consist of the interconnection of linear time invariant (LTI) systems and static nonlinearities. See (Gómez and Baeyens, 2004).

Within this class, three of the more common model structures are shown in Fig. B.5. The Hammerstein model consists of the cascade connection of a static (memoryless) nonlinearity followed by a LTI system. In the Wiener model the order of the linear and the nonlinear blocks in the cascade connection is reversed. And the feedback block-oriented (FBO) model consists of a static nonlinearity in the feedback path around a LTI system.

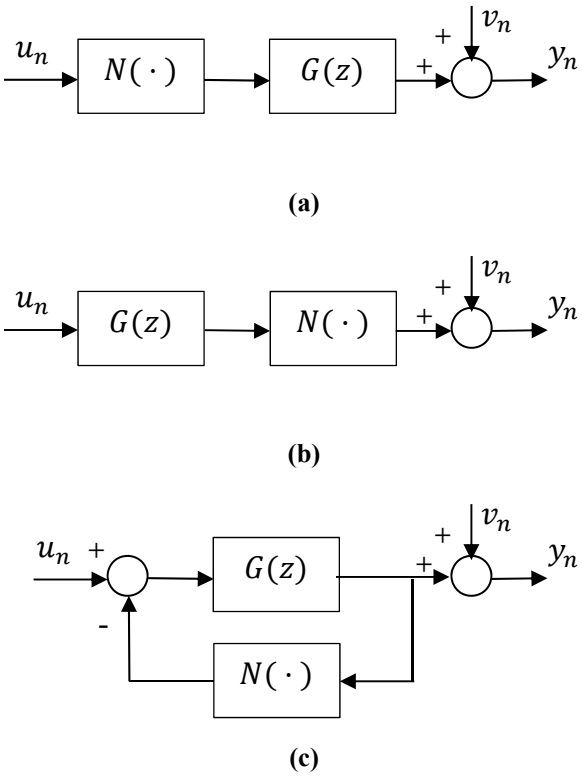


Fig. B.5. Block-oriented nonlinear models: (a) Hammerstein model, (b) Wiener model, (c) Feedback block-oriented model

To illustrate the procedure consider for instance the Hammerstein model.

$$y_n = G(q)N(u_n) + d_n = \left(\sum_{l=0}^{p-1} b_l B_l(q)\right)\left(\sum_{i=1}^{r} a_i g_i(u_n)\right) + v_n$$

The identification problem is to estimate the unknown parameter matrices $a_i \in \mathbb{R}^{n\times n}, i = 1,\ldots,r$ and $b_l \in \mathbb{R}^{m\times n}, l = 0,\ldots,p-1$ characterizing the nonlinear and the linear parts, respectively, from an *N*-point data set $\{u_n, y_n\}_{n=1}^{N}$ of observed input–output measurements.

To solve the problem we start by defining the input-output relation $\mathbf{y} = \boldsymbol{\Phi}_N^T\boldsymbol{\theta} + \mathbf{v}$, where $\mathbf{y} = (y_1,\ldots,y_N)^T$, $\mathbf{v} = (v_1,\ldots,v_N)^T$,$\boldsymbol{\Phi}_N = (\phi_1,\ldots,\phi_N)$, and

$$\boldsymbol{\theta} = \left(b_0a_1,\ldots,b_0a_r,\ldots,b_{p-1}a_1,\ldots,b_{p-1}a_r\right)^T$$

$$\phi_n = \left(B_0(q)g_1^T(u_n),\ldots,B_0(q)g_r^T(u_n),\ldots,B_{p-1}(q)g_1^T(u_n),\ldots,B_{p-1}(q)g_r^T(u_n)\right)^T$$

Then the solution algorithm is the following:

Step 1: Compute the least squares estimate $\widehat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}_N\boldsymbol{\Phi}_N^T)^{-1}\boldsymbol{\Phi}_N\boldsymbol{y}$.

From $\widehat{\boldsymbol{\theta}}$, construct the matrix

$$\widehat{\boldsymbol{\Theta}}_{ab} = \begin{bmatrix} a_1^T b_0^T & a_1^T b_1^T & \cdots & a_1^T b_{p-1}^T \\ a_2^T b_0^T & a_2^T b_1^T & & a_2^T b_{p-1}^T \\ \vdots & & & \vdots \\ a_r^T b_0^T & a_r^T b_1^T & \cdots & a_r^T b_{p-1}^T \end{bmatrix}$$

such that $\widehat{\boldsymbol{\theta}} = \boldsymbol{blockvec}(\widehat{\boldsymbol{\Theta}}_{ab})$.

Step 2: Compute the economy-size Singular Value Decomposition (SVD) of $\widehat{\boldsymbol{\Theta}}_{ab}$ as $\widehat{\boldsymbol{\Theta}}_{ab} = \mathbf{U}_n \boldsymbol{\Sigma}_n \mathbf{V}_n^T = \sum_{i=1}^{n} \sigma_i \, u_i v_i^T$, and the partition of this decomposition as $\widehat{\boldsymbol{\Theta}}_{ab} = [\mathbf{U}_1 \quad \mathbf{U}_2] \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 \\ 0 & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix}$.

Step 3: Compute the estimates of the parameter matrices $\hat{\mathbf{a}} = \mathbf{U}_1$ and $\hat{\mathbf{b}} = \mathbf{V}_1 \boldsymbol{\Sigma}_1$, respectively.

See (Gómez and Baeyens, 2004) for details.

APPENDIX C

Markov Chain Monte Carlo

In Bayesian statistical inference and decision theory the integration operation plays a fundamental role. For example, in computing the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ via the Bayes' rule, $p(\boldsymbol{\theta}|\mathbf{y}) = c^{-1}p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, the constant of proportionality is given by $c = p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$. In a multivariate case, marginal posterior distributions are computed as $p(\theta_i|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathrm{y})\mathrm{d}\boldsymbol{\theta}_{\neg i}$ where $\boldsymbol{\theta}_{\neg i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots \theta_d)^T$. And we might be interested in the minimization of average losses, $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \int L(\boldsymbol{\theta}, \boldsymbol{\theta}_0)p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$, and the computation of summary inferences in the form of posterior expectations, $E[g(\boldsymbol{\theta})|\mathbf{y}] = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$.

In many practical situations, due to a complex and maybe non-standard model structure, posterior probability distributions are not available in a closed form. Moreover optimization and integration of posterior distributions become more difficult as the dimension of the distribution increases. To overcome these drawbacks, simulation techniques such as Monte Carlo Markov chains (MCMC) are used.

The general methodology is reviewed in (Robert and Casella, 1999) and (Bergman, 1999). Other references are (Gelfand et al., 1992), (Tanner, 1996), (Gilks et al., 1996). Application examples are provided in (Girard and Parent, 2004) and (Bergman, 1999), and (Berger and Rios Insua, 1998) present advanced tools for the application of Bayesian methods to models beyond the field of linear regression.

C.1 Monte Carlo integration

Monte Carlo methods for numerical integration consider problems of the form

$$I = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{132}$$

where $\pi(\boldsymbol{\theta})$ is a positive function $\pi(\boldsymbol{\theta}) \geq 0$ that integrates to unity, $\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$. This assumption on the factor $\pi(\boldsymbol{\theta})$ leads to a natural interpretation of $\pi(\boldsymbol{\theta})$ as a probability density function. In the Bayesian context, the density of interest is usually the posterior density of the parameters given the observed data, i.e., $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$.

The Monte Carlo methods rely on the assumption that it is possible to draw a large number N of samples $\{\boldsymbol{\theta}_n\}_{n=1}^N$ distributed according the probability density $\pi(\boldsymbol{\theta})$. The Monte Carlo estimate of the integral (132) is then formed by taking the average over the set of samples

$$\hat{I}_N = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{\theta}_n) \tag{133}$$

where N is assumed to be large, $N \gg 1$.

If the samples in the set $\{\boldsymbol{\theta}_n\}_{n=1}^N$ are independent, $\hat{I}_N$ is an unbiased estimate of I and will almost surely converge to I, $\Pr\left[\lim_{N\to\infty} \hat{I}_N = I\right] = 1$, by the Strong Law of Large Numbers. Moreover, if the variance of $f(\boldsymbol{\theta})$, $\sigma^2 = \int (f(\boldsymbol{\theta}) - I)^2 \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})^2 \pi(\boldsymbol{\theta})d\boldsymbol{\theta} - I^2$, is finite, the error converges in distribution to a zero mean normal distribution, $\lim_{N\to\infty}\sqrt{N}(\hat{I}_N - I) \sim \mathcal{N}(0, \sigma^2)$, by the Central Limit Theorem. These two convergence results are asymptotic, for $N \to \infty$. In practical situations, we usually assume that a large but finite N will lead to a small error.

C.1.1 Comparison to standard numerical integration

The Monte Carlo methods are brute force algorithms but they present two main advantages compared to straightforward numerical integration. Firstly, when applied to high dimensional spaces, standard numerical integration methods generally fail due to their excessive demands for computational resources. Secondly, the error $\varepsilon = \hat{I}_N - I$ of the Monte Carlo estimate is of the order $\varepsilon = O(N^{-1/2})$, independently of the parameter dimension, d.

Standard numerical integration methods generally approximate the integral by a sum over a regular grid on the support set of the integrand. The Monte Carlo methods obtain an adaptive grid since they assume that it is possible to generate N samples from a density given as a factor of the integrand. This, in a sense, is the way these methods solve the curse of dimensionality and is the core difference between straightforward numerical integration and Monte Carlo integration methods.

C.1.2 Optimization

Considering Bayesian maximum *a posteriori* estimators, the sought estimate is the location of the maximum peak of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$, i.e., the mode of the density and, sometimes, what is wanted is the location of the maximum peak of some of the marginal distributions. In any case, the optimization method requires that the function to maximize can be evaluated, at least up to a normalizing factor.

In the cases where this is not possible we will use the Monte Carlo estimate of the density, i.e., the histogram of the Monte Carlo samples. This will yield a discretization of the parameter space and thus require even higher values of N for reliable results.

Minimum risk estimators are obtained in an analogous way. In this case, the Monte Carlo simulation (133) gives the value of the average loss, that is $f(\boldsymbol{\theta}_i) = L(\boldsymbol{\theta}_i, \boldsymbol{\theta}_0)$. The minimization of this loss can then be performed by means standard numerical techniques.

C.2 Sampling methods

The Monte Carlo framework for numerical integration and optimization relays on the assumption that $N \gg 1$ samples from a generic density $\pi(\boldsymbol{\theta})$ can be easily obtained. Methods that get samples from $\pi(\boldsymbol{\theta})$ are known as *sampling methods*. The function $\pi(\boldsymbol{\theta})$ is called the *target distribution*.

For standard distributions (such as uniform, Gaussian, Gamma, Student *t*, etc.) several perfect random sampling algorithms exist.

In the case that more general and higher dimensional distributions, for instance the ones generated by combinations and mixtures of basic distributions (Robert, 2001), it is not possible to directly generate samples of $\pi(\boldsymbol{\theta})$. However, when there is a known upper bound on the density function values, and it is possible to evaluate $\pi(\boldsymbol{\theta})$ everywhere up to a normalising constant, it is still possible to generate samples of $\pi(\boldsymbol{\theta})$.

Rejection sampling and *importance sampling* presented next are useful when the dimension of the state space is less than 10. In high dimensional problems, the approximate shape of the posterior is unknown and many problems arise (slow convergence, low acceptance).

C.2.1 Rejection sampling

The *rejection sampling* procedure is the simplest method. Let $q(\boldsymbol{\theta})$ be a *proposal distribution* from which samples are easily generated and assume that there exists a known constant $M > 1$ such that $\pi(\boldsymbol{\theta}) \leq Mq(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \mathbb{R}^d$.

The procedure is to draw a candidate sample $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta})$ and accept it with probability $1/M$. If $\boldsymbol{\theta}'$ is rejected, the procedure continues to draw samples from $q(\boldsymbol{\theta})$ until an accepted sample is obtained.

Algorithm C.1. Rejection sampling

Step 1. Sample $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta})$ and $u \sim \mathcal{U}(0,1)$.

Step 2. If $u < \frac{\pi(\boldsymbol{\theta}')}{Mq(\boldsymbol{\theta}')}$, $\boldsymbol{\theta}'$ is accepted, otherwise go to Step 1.

□

The finally accepted candidate will be an exact draw from the *target distribution,* $\pi(\boldsymbol{\theta})$. See (Bergman, 1999) for a proof in the scalar case.

C.2.2 Importance sampling

The procedure known as *importance sampling* also deals with a proposal distribution $q(\boldsymbol{\theta})$ which is easy to generate samples from. However, the only general assumption on the *importance function* $q(\boldsymbol{\theta})$ is that its support set covers the support of $\pi(\boldsymbol{\theta})$ i.e., that $\pi(\boldsymbol{\theta}) > 0 \implies q(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$. Under this assumption, any integral on the form (132) can be rewritten

$$I = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})d\boldsymbol{\theta}$$

A Monte Carlo estimate is computed by generating $N \gg 1$ independent samples from $q(\boldsymbol{\theta})$, and forming the weighted sum,

$$f_N = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{\theta_n})w(\boldsymbol{\theta_n}) \tag{134}$$

where $w(\boldsymbol{\theta_n}) = \frac{\pi(\boldsymbol{\theta_n})}{q(\boldsymbol{\theta_n})}$ are the *importance weights*.

Algorithm C.2. Sampling Importance Resampling

Step 1. Generate M independent samples $\{\boldsymbol{\theta}_m\}_{m=1}^{M}$ with common distribution $q(\boldsymbol{\theta})$.

Step 2. Compute the weight $w_m = w(\boldsymbol{\theta_m}) \propto \frac{\pi(\boldsymbol{\theta_m})}{q(\boldsymbol{\theta_m})}$ for each $\boldsymbol{\theta_m}$.

Step 3. Normalize the weights $w_m \equiv \gamma^{-1}w_m$, where $\gamma = \sum_{m=1}^{M} w_m$.

Step 4. Resample with replacement N times from the discrete set $\{\boldsymbol{\theta}_m\}_{m=1}^{M}$ where $\Pr[\text{resampling } \theta_m] = w_m$.

□

The success of applying either the rejection sampling or importance sampling methods relies on determining good proposal distributions and importance functions, respectively. A badly chosen proposal distribution yields a low acceptance rate in the rejection sampling algorithm. Likewise, choosing the wrong importance function yields a large variance of the importance weights with only some samples contributing to the sum (134), and thus a slow convergence of the estimate.

C.3 Markov chain Monte Carlo

An alternative to classical methods are the Markov chain Monte Carlo (MCMC) techniques which generate samples from desired distributions by embedding them as limiting distributions of Markov chains (Andrieu et al., 2001).

The MCMC algorithms are iterative procedures that deliver a sequence of random samples by simulating a Markov chain designed to have a limit distribution given by the density $\pi(\boldsymbol{\theta})$. By discarding an initial *burn in* phase of the Markov chain, ergodic averages of the chain realization can be used to estimate integrals with respect to $\pi(\boldsymbol{\theta})$. Another advantage of using MCMC is that *credible intervals* for any quantity of interest can be formed. See the survey article (Tierney, 1994) for the theoretical foundations of Markov chain Monte Carlo methods.

C.3.1 Markov chain

A *Markov chain* is a sequence of random variables $\{\boldsymbol{\theta}_t\}_{t\geq=0}$ such that

$$\Pr[\boldsymbol{\theta}_t \in A|\boldsymbol{\theta}_0, \dots, \boldsymbol{\theta}_{t-1}] = \Pr[\boldsymbol{\theta}_t \in A|\boldsymbol{\theta}_{t-1}], \quad \forall A \subset \mathbb{R}^d$$

The *transition kernel* of the Markov chain is the conditional density function

$$K(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t) \equiv p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$$

A *time-homogenous* Markov chain is one where the transition kernel is explicitly independent of the time index *t*. In the case of Markov chains over discrete state spaces, the transition kernel is a discrete transition probability matrix. The *p*-step transition kernel is given by

$$K_p(\boldsymbol{\theta}_{t-p}, \boldsymbol{\theta}_t) \equiv p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-p})$$

The initial distribution of the Markov chain is $p(\boldsymbol{\theta}_0)$ and may, in the general case, be a Dirac delta measure indicating that the initial state of the Markov chain is deterministic.

C.3.2 Properties of the Markov chain

e. Invariance

The idea behind Markov chain Monte Carlo methods is to construct a transition kernel such that the limiting, or stationary, distribution of the output of the Markov chain is the desired probability density function $\pi(\boldsymbol{\theta})$.

In order to fulfill this requirement, a condition of *invariance* must hold between the transition kernel $K_p(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)$ of the Markov chain and the target distribution $\pi(\boldsymbol{\theta})$.

Definition C.1. Invariance. The probability density function $\pi(\boldsymbol{\theta})$ is said to be invariant (or stationary) with respect to the transition kernel K if

$$\pi(\boldsymbol{\theta}) = \int K(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)\pi(\boldsymbol{\theta}_{t-1})d\boldsymbol{\theta}_{t-1}\ ,\ \forall \boldsymbol{\theta} \subset \mathbb{R}^d.$$

□

The density $\pi(\boldsymbol{\theta})$ being invariant with respect to the Markov chain implies that if $\boldsymbol{\theta}_t \sim \pi(.)$ for some t, the output of the chain will remain marginally distributed according to $\pi(.)$ for all future time instants. A sufficient condition to ensure π-invariance is to assure that the Markov chain is π-reversible.

f. Reversibility

Definition C.2. Reversibility. A transition kernel K is π-reversible if it satisfies $K(\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}) = K(\mathbf{y}, \boldsymbol{\theta})\pi(\mathbf{y})$.

□

The reversibility condition says that the probability of the Markov chain moving from a region A to a region B is equal to the probability of moving from B to A. This holds whenever the state is in the stationary regime, i.e., under the assumption that it is distributed according to $\pi(\boldsymbol{\theta})$ before the move takes place. Most MCMC algorithms are π-reversible by construction, and therefore $\pi(\boldsymbol{\theta})$ is an invariant distribution of the Markov chain.

g. Irreducibility

Irreducibility defines the regions of the state space which the chain can move around in, but never leave.

Definition C.3. Irreducibility. A Markov chain is φ-irreducible if for any $A \subset \mathbb{R}^d$, if

$\int_A \varphi(\mathbf{y})d\mathbf{y} > 0$, then exists some $p \in \mathbb{Z}_1^\infty$ such that $\int_A K_p(\boldsymbol{\theta}, \mathbf{y})d\mathbf{y} > 0$, for any $\boldsymbol{\theta} \subset \mathbb{R}^d$.

□

A sufficient condition for a kernel K to be φ-irreducible is that for some $p \geq 1$, the kernel $K_p(\boldsymbol{\theta}, \mathbf{y})$ can be factorized by $\varphi(\mathbf{y})$, i.e., that there exists a positive function $f(\boldsymbol{\theta}, \mathbf{y}) > 0$ such that $K_p(\boldsymbol{\theta}, \mathbf{y}) = f(\boldsymbol{\theta}, \mathbf{y})\varphi(\mathbf{y})$. If a chain is irreducible with respect to some density φ and has invariant density π, then the chain is π -irreducible. This leads to the existence of a Strong Law of Large Numbers for Markov chain Monte Carlo methods.

C.3.3 MCMC algorithms

There are many ways of categorizing MCMC methods, but the simplest one is to classify them in one of two groups (Andrieu et al., 2001):

1. The first is used in estimation problems where the unknowns are typically parameters $\boldsymbol{\theta}$ of a model, which is assumed to have generated the observed data $\mathbf{y}$. Examples are the Metropolis-Hastings sampler and the Gibbs sampler.

2. The second is employed in more general scenarios where the unknowns are not only model parameters, but models as well. MCMC methods for the second group allow for generation of samples from probability distributions defined on unions of disjoint spaces of different dimensions. Sampling from such distributions is a nontrivial task. The most representative is the Reversible Jump MCMC.

C.3.4 Metropolis-Hastings algorithm

Most algorithms for Markov chain Monte Carlo estimation are based on the algorithm of Hastings (Hastings, 1970), which is a generalization of the algorithm of Metropolis *et al.* (Metropolis *et al*., 1953).

The Metropolis-Hastings algorithm resembles the previously described sampling methods that a proposal distribution $q(.)$ is used to generate the samples. However, the output of the algorithm is a Markov chain so the proposal density may depend on the current state of the chain.

Let $\boldsymbol{\theta}$ denote the current state of the chain in an iteration of the Metropolis-Hastings algorithm. A candidate sample $\mathbf{z}$ is drawn from the proposal $q(\mathbf{z}|\boldsymbol{\theta})$ and accepted with a probability given by

$$\alpha(\boldsymbol{\theta}, \mathbf{z}) = \min\left(1, \frac{\pi(\mathbf{z})q(\boldsymbol{\theta}|\mathbf{z})}{\pi(\boldsymbol{\theta})q(\mathbf{z}|\boldsymbol{\theta})}\right) \quad \textbf{(135)}$$

If the candidate is accepted the chain moves to the new position, while a rejection of the candidate leaves the chain at the current position in the state space. An interpretation of (135) is that all candidates that yield an increase of π (and is not too unlikely to return from) are accepted.

Algorithm C.3. Metropolis-Hastings Sampler

Step 1. Initialize by setting $t = 0$ and choosing $\boldsymbol{\theta}_0$ randomly or deterministically.

Step 2. Sample $\mathbf{z} \sim q(\mathbf{z}|\boldsymbol{\theta})$.

Step 3. Sample $u \sim \mathcal{U}(0,1)$.

Step 4. Compute the acceptance probability $\alpha(\boldsymbol{\theta}, \mathbf{z})$.

Step 5. If $u \leq \alpha(\boldsymbol{\theta}_t, \mathbf{z})$ accept the move and set $\boldsymbol{\theta}_{t+1} = \mathbf{z}$. Otherwise set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$.

Step 6. Increase t and return to **Step 2**.

□

One very important feature of the Metropolis-Hastings algorithm is that the distributions $\pi(\boldsymbol{\theta})$ only need to be known up to a normalizing constant. The normalizing factor of $\pi(\boldsymbol{\theta})$, $\int \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$, cancels in the expression for the acceptance probability (135) which thus can be evaluated even if it is unknown.

A simplistic way to choose the proposal is to have it fixed, and independent of the current state of the chain. The *independence sampler* (Tierney, 1994) with a proposal distribution $q(\mathbf{z}|\boldsymbol{\theta}) = q(\mathbf{z})$ yields an acceptance probability (135) of

$$\alpha(\boldsymbol{\theta}, \mathbf{z}) = \min\left(1, \frac{w(\mathbf{z})}{w(\boldsymbol{\theta})}\right) \quad \text{where} \quad w(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}$$

In the original algorithm of Metropolis (Metropolis *et al*, 1953), symmetric proposals were considered, i.e., proposal distributions such that $q(\mathbf{z}|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\mathbf{z})$. The acceptance probability then simplifies to

$$\alpha(\boldsymbol{\theta}, \mathbf{z}) = \min\left(1, \frac{\pi(\mathbf{z})}{\pi(\boldsymbol{\theta})}\right)$$

Example C.1. Effect of different proposal distributions

The efficiency of the Metropolis-Hastings algorithm depends on the choice of the proposal distribution. Let us illustrate this effect by an example drawn from (Bergman, 1999).

The pdf to be sampled is a Gaussian mixture consisting of a sum of two Gaussian pdfs, the first with $\mu_1 = -6$, $\sigma_1 = 1$, and the second with $\mu_2 = 3$, $\sigma_2 = 2.5$.

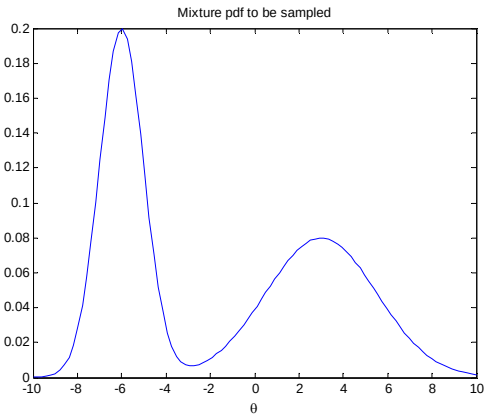


Fig. C.1. Gaussian mixture to be sampled

The chain is deterministically initialized between the modes of $\pi(\boldsymbol{\theta})$,

$$\theta_0 = \frac{\mu_1 + \mu_2}{2} = \frac{-6+3}{2} = 1.5$$

and the proposal $q(\mathbf{z}|\boldsymbol{\theta})$ yields a random walk, i.e., the proposal point is chosen as an independent zero mean addition to the current state of the chain.

$$z_t = \theta_t + x_t \qquad , \qquad x_t \sim \mathcal{N}(0, \sigma_x^2)$$

Different behavior is obtained depending on the average size of the steps proposed by $q(\mathbf{z}|\boldsymbol{\theta})$. With too small steps ($\sigma_x = 0.2$), the chain gets stuck around a local mode of the target distribution. See next figure.

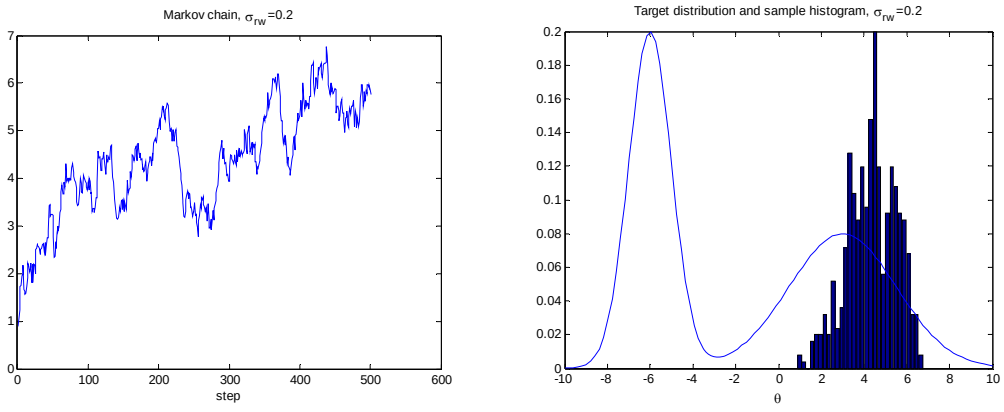


Fig. C.2. Case $\boldsymbol{\sigma_x = 0.2}$. Only one mode is explored

And with too large steps ($\sigma_x = 20$), the proposal will often end up in the tails of $\pi(\boldsymbol{\theta})$ and thus frequently be rejected by the Metropolis-Hastings algorithm.

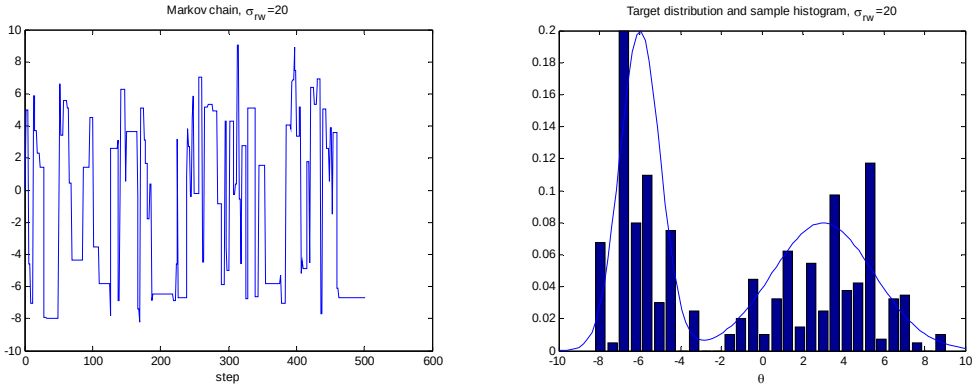


Fig. C.3. Case $\boldsymbol{\sigma_x = 20}$. Many candidates are rejected

The two previous cases are often referred to as slowly mixing chains, while next figure shows a choice of proposal ($\sigma_x = 2$) yielding a good mixing of the chain.

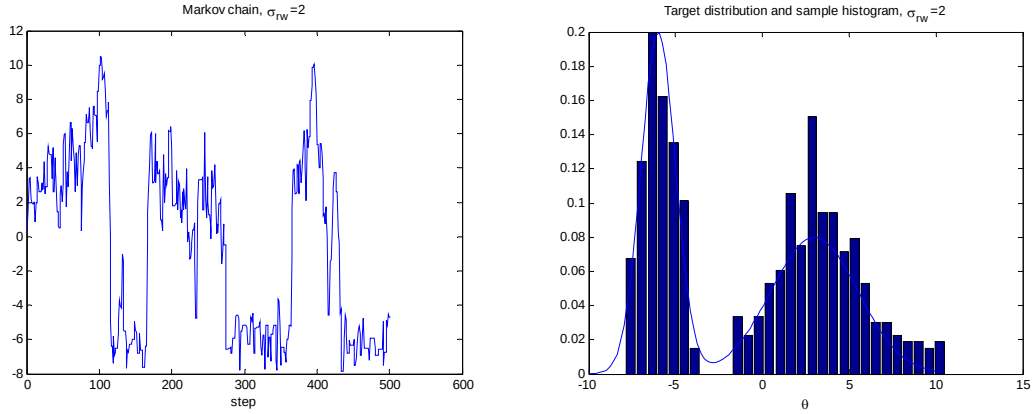


Fig. C.4. Case $\boldsymbol{\sigma_x = 2}$. Both modes are visited (the acceptance probability is high)

■

C.3.5 Gibbs sampling

In the Metropolis-Hastings algorithm, an alternative way to propose a new candidate vector $\mathbf{z}$ is to update scalar or low dimensional subcomponents of $\boldsymbol{\theta}$, in a *blocking* scheme. This is often referred to as *single-component*, or *one-at-a-time* Metropolis-Hastings and it can be a particularly efficient approach in high dimensional problems where it is often hard to choose good proposal distributions.

In single-component Metropolis-Hastings, each component of $\boldsymbol{\theta}$ is updated according to a Metropolis-Hastings step where the invariant distribution is the *full conditional* distribution of that component. The full conditional distribution for the element i of the parameter vector is

$$\pi(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\neg i}) = \frac{\pi(\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}_i}$$

where $\boldsymbol{\theta}_{\neg i}$ is the vector consisting of all elements of $\boldsymbol{\theta}$ except for element number *i*,

$$\boldsymbol{\theta}_{\neg i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots \theta_d)^T$$

A unique proposal $q_i(\mathbf{z}_i|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\neg i})$ can be used for each entry *i*. The algorithm cycles through the entries of $\boldsymbol{\theta}$ sampling from each proposal and accepts the candidate entry $\mathbf{z}_i$ with probability

$$\alpha(\boldsymbol{\theta}_{\neg i}, \boldsymbol{\theta}_i, \mathbf{z}_i) = \min\left(1, \frac{\pi(\mathbf{z}_i|\boldsymbol{\theta}_{\neg i}) q_i(\boldsymbol{\theta}_i|\mathbf{z}_i, \boldsymbol{\theta}_{\neg i})}{\pi(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\neg i}) q_i(\mathbf{z}_i|\boldsymbol{\theta}_{\neg i}, \boldsymbol{\theta}_i)}\right) \quad \textbf{(136)}$$

The newly accepted or rejected entry is then inserted into $\boldsymbol{\theta}$ and the next candidate component is sampled from the proposal distribution of that entry.

The Gibbs sampling algorithm is the most commonly applied MCMC algorithm. The Gibbs sampling algorithm can be seen as a blocking Metropolis-Hastings procedure where proposal samples are drawn directly from the full conditional distributions. Inserting

$$q_i(\mathbf{z}_i|\boldsymbol{\theta}_{\neg i}) = \pi(\mathbf{z}_i|\boldsymbol{\theta}_{\neg i})$$

into (136) yields an acceptance probability of one. Hence, all candidates are accepted and no acceptance probability has to be evaluated.

Algorithm C.4. Gibbs Sampler

Step 1. Initialize by setting $t = 0$ and choose $\boldsymbol{\theta}^{(0)}$ randomly or deterministically.

Step 2. Cycle through the entries of $\boldsymbol{\theta}$ and sample from the full conditionals,

$$\begin{gathered}\boldsymbol{\theta}_1^{(t)} \sim \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{\neg 1}^{(t)}) \\ \boldsymbol{\theta}_2^{(t)} \sim \pi(\boldsymbol{\theta}_2, \boldsymbol{\theta}_{\neg 2}^{(t)}) \\ \ldots \\ \boldsymbol{\theta}_d^{(t)} \sim \pi(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{\neg d}^{(t)})\end{gathered}$$

Step 3. Output $\boldsymbol{\theta}^{(t)}$, increase t and return to **Step 2**.

□

Algorithm C.4 is the deterministic version of the Gibbs sampler. Alternatively, one can cycle through the entries of $\boldsymbol{\theta}$ in a random fashion. Moreover, other partitions of $\boldsymbol{\theta}$ can be used, e.g., one can choose to sample highly correlated entries as one block.

C.3.6 Reversible jump MCMC algorithm

Reversible jump MCMC (Green, 1995) provides a general framework for the case in which the dimension of the parameter space can vary between iterations of the Markov chain. It is the case when we need to calculate posterior probabilities of hierarchical models and when other methods are infeasible because of the large number of possible models (Dellaportas and Forster, 1999). It is also the case when sieve priors are used in which the number of parameters to be sampled depends on another parameter (McVinnish *et al*., 2006). Recent references about the topic are (Green and Hastie, 2009) and (Fan and Sisson, 2010).

In the Bayesian modelling context, suppose that for the observed data $\mathbf{y}$ we have a countable collection of candidate models $\mathcal{M} = \{M_1, M_2, ...\}$ indexed by a parameter k. Each model M_k has a d_k-dimensional vector of unknown parameters, $\boldsymbol{\theta}_k$. Thus, the target distribution is the joint posterior distribution given the observed data $p(k, \boldsymbol{\theta}_k|\mathbf{y})$.

Reversible jump MCMC can be viewed as an extension of the Metropolis-Hastings algorithm onto more general state spaces.

Algorithm C.5. Reversible Jump Sampler

Step 1: Initialize k and $\boldsymbol{\theta}_k$ at iteration $t = 0$.

Step 2: For iteration $t \geq 1$ perform

Step 2.1: Within-model move: With a fixed model k, update the parameters $\boldsymbol{\theta}_k$ according to any MCMC updating scheme.

Step 2.2: Between-models move: Simultaneously update model indicator k and the parameters $\boldsymbol{\theta}_k$ according a reversible proposal/acceptance mechanism.

Step 3: Increment iteration $t = t + 1$. If $t < N$, go to **Step 2.**

□

Step 2.1 can be achieved by a simple random walk Metropolis-Hastings proposal. And one possibility for the Step 2.2 is that the algorithm randomly proposes one of the following move types:

Move 1: Birth move: Move from k to k+1. The "birth" is made by proposing a $\boldsymbol{\theta}_{k+1}$ from $\mathcal{N}(0, \sigma^2)$, where σ^2 is chosen so that the acceptance probability is 1 when $\boldsymbol{\theta}_{k+1} = \mathbf{0}$ is proposed.

Move 2: Death move: Move from k to k-1.

The birth-death moves are based on the zero order centered proposals as defined in (Brooks *et al*., 2003).

The convergence of the resulting Markov chain to its stationary distribution can be assisted by using appropriate starting values, that is by starting the chain in an area of significant probability. In (McVinnish *et al*., 2006), these starting values are obtained using least squares estimates of the impulse response sequence given a moderate value of K.

By the Law of Large Numbers for Markov chains this estimate will converge almost surely to the correct value as the number of samples from the posterior goes to infinity.

APPENDIX D

Bayesian Decision Theory

Many aspects of the system identification, model validation, experiments design and fault detection can be interpreted from a (Bayesian) decision theory viewpoint. In the present appendix, we present the fundamentals of Bayesian modeling and summarize several main concepts of the Bayesian decision theory.

D.1 Fundamentals of Bayesian modelling

In this section we summarize the main concepts of Bayesian statistical analysis and modeling. For more details, the reader is referred to the textbooks (Box and Tiao, 1973), (Berger, 1985) and (Robert, 2001). Early works dealing with the Bayesian approach to classical system identification are (Eykhoff, 1974) and (Perterka, 1981). Also, a survey of Bayesian analysis and its applications can be found in (Berger, 2000). Finally, for recent works in this area, see (Ninness and Henriksen, 2010) and (Schön *et al.*, 2011).

D.1.1 Bayes Theorem

Bayes' Theorem was firstly published in 1763 (Bayes, 1763a), (Bayes, 1763b), after Thomas Bayes' death.

Theorem D.1. Bayes Theorem (1763)

If A and E are events such that $\Pr(E) \neq 0$, then $\Pr(A|E)$ and $\Pr(E|A)$ are related by

$$\Pr(A|E) = \frac{\Pr(E|A)\,\Pr(A)}{\Pr(E|A)\Pr(A) + \Pr(E|A^c)\,\Pr(A^c)} = \frac{\Pr(E|A)\,\Pr(A)}{\Pr(E)} \quad \textbf{(137)}$$

where A^c stands for the complementary event of A in the sense that $\Pr(A) + \Pr(A^c) = 1$. In particular,

$$\frac{\Pr(A|E)}{\Pr(B|E)} = \frac{\Pr(E|A)}{\Pr(E|B)} \quad \textbf{(138)}$$

when $\Pr(B) = \Pr(A)$.

□

One interesting feature of the Bayes' Theorem is that (137) constitutes an *actualization* principle since it describes the updating of the likelihood of A from $\Pr(A)$ to $\Pr(A|E)$ once E has been observed. Also equation (138) expresses the fundamental fact that, for two equal probable causes A and B, the ratio of their probabilities given a particular effect E is the same as the ratio of the probabilities of the effect E given the two causes.

Nowadays, to prove Theorem D.1 is trivial thanks to modern axiomatic probability theory. However, by the time it was formulated it represented a major conceptual step in the history of Statistics, being the first *inversion* of probabilities. Actually, at the end of the XVIII Century, Statistics was often called *Inverse Probability* due to this interpretation[2] (Robert, 2001).

The meaning of inversion here is the following: In *probabilistic modeling*, one characterizes (in a probabilistic way) the behavior of the future observations $\mathbf{y}$ conditional on model parameters $\boldsymbol{\theta}$. By contrast, in a *statistical analysis* the objective is to retrieve the causes (make an *inference* about $\boldsymbol{\theta}$) from the effects (the observations $\mathbf{y}$).

The definition of the likelihood function is an obvious example of the inverting nature of Statistics since, formally, it is just the sample density rewritten in the proper order,

$$l(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}) \quad \textbf{(139)}$$

Bayes proved a continuous version of Theorem D.1 and went further considering that the uncertainty on the parameters $\boldsymbol{\theta}$ of a model could be described through a probability distribution π on Θ, $\pi(\boldsymbol{\theta})$, called *prior distribution*. The inference is then based on the distribution of $\boldsymbol{\theta}$ conditional on $\mathbf{y}$, $\pi(\boldsymbol{\theta}|\mathbf{y})$, called *posterior distribution* and defined by

[2] There exist many classical books about Philosophy of Science which include very interesting historical examples related to Bayes' ideas. See, e.g. (Earman, 1992), (Horwich, 1982), (Rosenkrantz, 1977), (Howson and Urbach, 1989).

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y},\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{140}$$

where $p(\mathbf{y},\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is the joint probability of the observation and parameters. Note that the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ is actually proportional to the likelihood (the distribution of $\mathbf{y}$ conditioned upon $\boldsymbol{\theta}$), multiplied by the prior distribution of $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto l(\boldsymbol{\theta}|\mathbf{y}) \cdot \pi(\boldsymbol{\theta}) \tag{141}$$

Let us illustrate the updating of prior beliefs by means the following example (Box and Tiao, 1973):

Example D.1. Bayes' rule: How to model the knowledge gained from experience

Two physicists, Mr. A and Mr. B, are concerned with estimating some physical constant θ, previously known only approximately.

Prior distributions: Physicist A, being very familiar with this area of study, can make a moderately good guess of what the answer will be, and his prior opinion about θ can be approximately represented by a normal distribution centered at 900, with a standard deviation of 20, that is $\theta \sim \mathcal{N}(900, 20^2)$,

$$p_A(\theta) = \frac{1}{\sqrt{2\pi}20} \exp\left[-\frac{1}{2}\left(\frac{\theta - 900}{20}\right)^2\right]$$

By contrast, Mr. B has had little previous experience in this area, and his rather vague prior beliefs are represented by the normal distribution $\theta \sim \mathcal{N}(800, 80^2)$. That is, he centers his prior at 800 and is considerably less certain about θ than A (his standard deviation is 80),

$$p_B(\theta) = \frac{1}{\sqrt{2\pi}80} \exp\left[-\frac{1}{2}\left(\frac{\theta - 800}{80}\right)^2\right]$$

Fig. D.1(a) shows the prior distributions $p_A(\theta)$ and $p_B(\theta)$.

Likelihood function: Suppose now that an unbiased method of experimental measurement is available. Any observation y made by this method follows a Normal distribution where the mean value is the real value of θ and the standard deviation is 40. Hence, the standardized likelihood function can be represented by a normal curve centered at y with standard deviation 40. Let us suppose that the result of the single observation is $y = 850$, then the likelihood function is shown in Fig. D.1(b).

Posterior distributions: Now we apply the Bayes' theorem to show how each man's opinion regarding θ is modified by the information coming from that piece of data. Since the prior distributions and the likelihood function are Gaussian, the posterior distributions will be Gaussian as well[3].

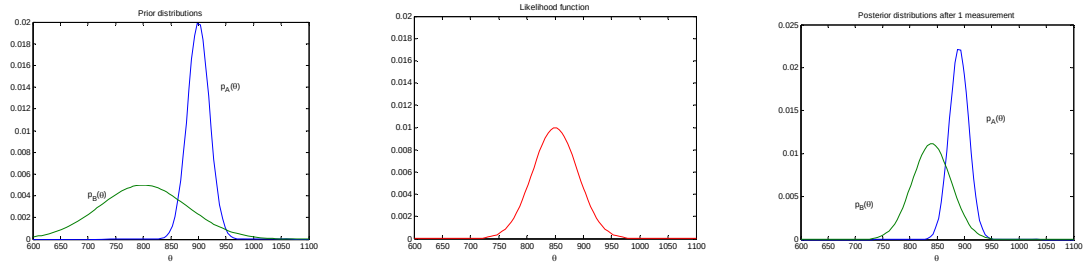
The combination of a prior $\mathcal{N}(\theta_0, \sigma_0^2)$ and a standardized likelihood function $\mathcal{N}(y, \sigma^2)$ leads to a posterior $\mathcal{N}(\bar{\theta}, \bar{\sigma}^2)$ where the parameters are given by

$$\bar{\theta} = \frac{w_0\theta_0 + w_1 y}{w_0 + w_1} \quad \text{and} \quad \bar{\sigma}^2 = \frac{1}{w_0 + w_1}$$

where $w_0 = \frac{1}{\sigma_0^2}$ and $w_1 = \frac{1}{\sigma^2}$.

The posterior mean $\bar{\theta}$ is a weighted average of the prior mean θ_0 and the observation y, the weights being proportional to w_0 and w_1 which are, respectively, the reciprocal of the variance of the prior distribution of θ and that of the observation y.

Physicist A's posterior opinion now is represented by the normal distribution $p_A(\theta|y)$ with mean 890 and standard deviation 17.9, while that for B is represented by the normal distribution $p_B(\theta|y)$ with mean 840 and standard deviation 35.78. These posterior distributions are shown in Fig. D.1(c).



(a) **(b)** **(c)**

Fig. D.1. (a) Prior distributions, (b) likelihood function and (c) posterior distributions

After this single observation, we see that the ideas of A and B about θ, as represented by the posterior distributions, are much closer than before, although they still differ considerably. We see that A, relatively speaking, did not learn much from the experiment, while B learned a great deal. The reason is that to A, the uncertainty in the measurement, as reflected by $\sigma = 40$, was larger than the uncertainty in his prior ($\sigma_{0,A} = 20$). On the other hand, the uncertainty in the measurement was considerably smaller than that in B's prior ($\sigma_{0,B} = 80$). For A, the prior has a stronger influence on the posterior distribution than has the likelihood, while for B the likelihood has a stronger influence than the prior.

[3] In Bayesian probability theory, prior and posterior distributions are called *conjugate distributions* if they belong to the same family. In particular, the Gaussian family is conjugate to itself (or *self-conjugate*) with respect to a Gaussian likelihood function: if the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian.

Now suppose that 99 further independent measurements are made and that the sample mean $\bar{y} = \frac{1}{100}\sum_{i=1}^{100} y_i$ of the entire 100 observations is 870. Fig. D.2(a) shows the new likelihood function and Fig. D.2(b) shows the new posterior distributions. After 100 observations, A and B would be in almost complete agreement. This is because the information coming from the data almost completely overrides prior differences.

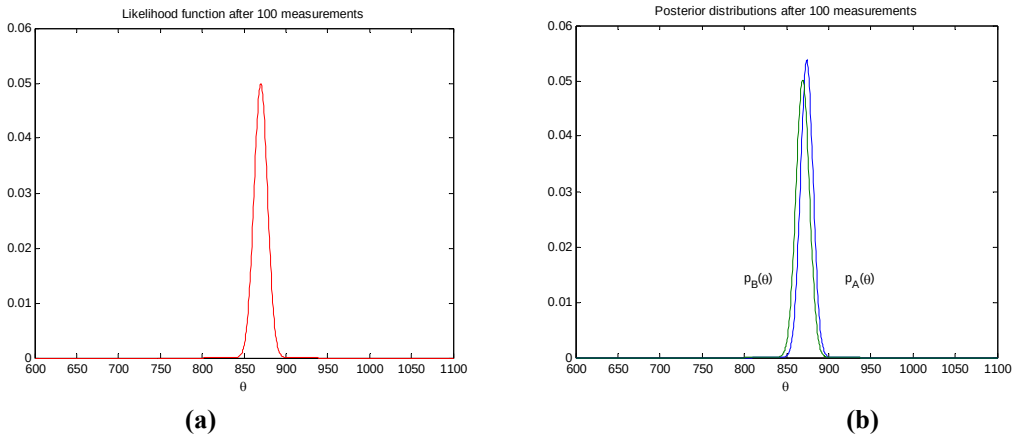


Fig. D.2. (a) Likelihood function and (b) posterior distributions after 100 measurements

This example shows how the contribution of the prior in the posterior computation depends on its sharpness or flatness in relation to the sharpness or flatness of the likelihood with it has to be combined.

After a single observation, the priors were very influential in deciding the posterior distributions, since the likelihood was not sharply peaked relative to either of them. For this reason the posterior distributions were so much different.

But after 100 observations, the priors were dominated by the likelihood (both the priors were rather flat compared with the likelihood function), and for this reason the posteriors are so much closer.

■

D.1.2 Consistency and efficiency

There exist several results concerning the consistency and the efficiency of Bayesian inference (Robert, 2001). In a general context, Bayes estimators are asymptotically *consistent*, that is, they almost surely converge to the true value of the parameter when the number of observations N goes to infinity. This is the case with estimators $\widehat{\boldsymbol{\theta}}$ that minimize the posterior loss associated with the loss function $L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \left|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right|^{\alpha}$, $\alpha \geq 1$, under weak constraints on the prior distribution $p(\boldsymbol{\theta})$ and the sampling density $p(\mathbf{y}|\boldsymbol{\theta})$.

The consistency can also be defined in terms of the Hellinger distance (Barron *et al.*, 1999). The Hellinger distance between two probability distributions p_1 and p_2 is defined as

$$d(p_1, p_2) = \int \left(p_1(\boldsymbol{\theta})^{1/2} - p_2(\boldsymbol{\theta})^{1/2}\right) d\boldsymbol{\theta} \tag{142}$$

A general condition for consistency of a posterior distribution is that in the Hellinger neighborhood of the true distribution, the posterior probability tends to the true almost surely when the sample size N goes to infinity. The basic assumption needed on the prior distribution $p(\boldsymbol{\theta})$ is that it gives positive mass to every Kullback-Leibler neighborhood of the true distribution. The Kullback-Leibler pseudo-distance is

$$d(p_1, p_2) = \int \ln \frac{p(\boldsymbol{\theta}|\mathbf{y})}{p(\boldsymbol{\theta})} p(\mathbf{y}, \boldsymbol{\theta}) d\boldsymbol{\theta}\, d\mathbf{y} \tag{143}$$

Sufficient conditions for convergence of the posterior distribution around the true system (and not only for Gaussisan noise) can be found in (Ghosal and van der Vaart, 2007).

Finally, regarding the asymptotic *efficiency* of some Bayes estimates, the posterior distribution converges towards the true value at the rate $N^{-1/2}$.

D.1.3 Selection of the prior distributions

The novelty of Bayesian inference compared to classical system identification methods is that expressing prior knowledge about the model by means a probability distribution puts in the same conceptual level *stochastic measurement noise* and *model*.

In the Bayesian view, any quantity the true value of which is not known is a *random variable*. Therefore not only experimental data are realizations of a random process, models are considered as random entities as well. Any unknown or uncertain constant (like model parameters) is a random variable.

a. Subjective priors

In order to apply the Bayes' rule and hence make posterior inferences about the model, prior distributions on the model and measurement noise must be defined before any experiment has been carried out. The purpose of them is to reflect the prior knowledge about the plant and noise. Sometimes there is no prior knowledge for sure and it must be substituted by arbitrary assumptions, more or less educated, or personal beliefs. Hence the term *subjective* appears.

The choice of the subjective prior is the main issue in the Bayesian paradigm and a great amount of works deal with this topic. See, for instance (Jeffrey, 2004). The selection of the prior is a critical point since, once selected, the posterior is computed in a systematic way and the inferences almost too.

However, it is not easy to specify numerically and uniquely one's own state of mind in terms of prior probability distribution, especially in multivariate problems. Also, the selection will be always arbitrary. Many candidates for prior distributions have been proposed. See a survey in (Robert, 2001, Ch.3). And general guidelines to their choice are given by (Berger, 1985), (Box and Tiao, 1973) and (Robert, 2001).

b. Types of priors

Conjugate distributions: Conjugate distributions are used when the prior information about the model is too vague or unreliable. However, they are not necessarily non-informative. They are defined as follows (Robert, 2001):

Definition D. 1. A family $\mathcal{F}$ of probability distributions on Θ is said to be conjugate (or closed under sampling) for a likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ if, for every prior belonging to $\mathcal{F}$, the posterior distribution also belongs to $\mathcal{F}$. □

Conjugate prior distributions are usually associated with exponential sampling distributions. Many common distributions (such as the Normal, Poisson, Gamma, Binomial, Beta) are exponential distributions. For instance, the Normal distribution is conjugate with respect to Normal likelihood, the Gamma distribution is conjugate with respect to Normal, Gamma and Poisson likelihoods, etc.

Improper distributions: Improper priors have distributions which integrate to infinity and arise when the support of the distribution is unbounded and a uniform distribution is used. There are several examples of paradoxes arising when improper priors are used. Their use is motivated by the fact that in many cases the posterior distribution is still proper.

Non-informative distributions: If we are ignorant about the process and noise we can model this state of knowledge by means of an uninformative prior, relatively flat compared to the information coming from the data, i.e. compared to the likelihood function. This type of prior is chosen also in situations where we want that the data "speak from themselves" without any prejudice introduced by the prior.

Non-informative prior distributions are purely subjective distributions. However, a completely unprejudiced prior is very difficult to obtain. In fact, it is impossible "knowing nothing" about a model and it is impossible to describe the state of "absolute ignorance" by means a model. Next example from (Hjalmarsson and Gustafsson, 1995) illustrates this point.

Example D.2. Non-informative uniform prior

Consider the case where the random variable X is known to have the range $[-5,5]$. It might appear as if a uniform distribution on this interval is non-informative. The

knowledge that $X \in [-5,5]$ is exactly the same as the knowledge that $X^2 \in [0,25]$. Hence a non-informative prior is that X^2 is uniformly distributed on $[0,25]$. But these two priors are completely different,

$$P_1 = \Pr(X^2 < 12.5 | X^2 \sim \mathcal{U}(0,25)) = \Pr(X < 0 | X \sim \mathcal{U}(-5,5)) = 0.5$$
$$P_2 = \Pr(X^2 < 12.5 | X \sim \mathcal{U}(-5,5)) = \frac{\sqrt{12.5}}{5} = 0.7071 \neq P_1$$

This example illustrates the fact that any attempt to translate very imprecise knowledge of the type "$X \in [-5,5]$" into a probabilistic prior will always introduces a prejudice. ■

The expression "non-informative" (as well as the concept of information in general) always has only a relative meaning and all what can be done is to suggest a reasonable mathematical model of the situation when "little is known *a priori*".

Many authors have proposed non-informative priors. The most representative are the Jeffrey's prior, maximum entropy priors, and reference priors. In (Kass and Wasserman, 1996) methods for selecting non-informative priors are reviewed.

Example D.3. Reference prior

Noise and model prior distributions can be combined in several ways. For instance, consider the case of normal sampling distribution and non-informative priors with $\boldsymbol{\theta}$ and $\ln \lambda$ approximately independent and locally uniform so that a non-informative *reference* prior is

$$p(\boldsymbol{\theta})p(\lambda) \propto \lambda^{-1}$$

where λ is the noise variance. If the noise variance λ is not known, information about it coming from the sample can be used.

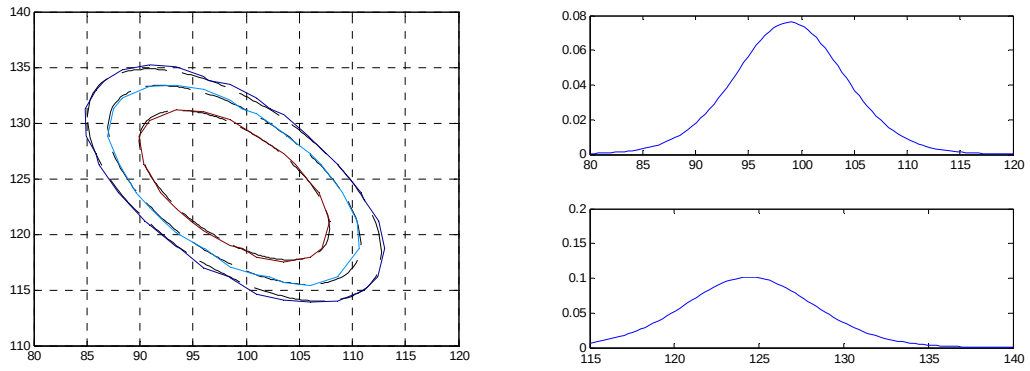
Remark: In the Bayesian viewpoint the independence is defined rather in terms of conditional probabilities, that is "*a*" is independent of "*b*" if the knowledge of the true value of "*b*" does not bring any information about "*a*" and therefore $p(a|b)=p(a)$. Obviously, this leads to the traditional (parametric, Fisherian) definition of independence, $p(a,b)=p(a)p(b)$.

The resulting posterior is the multivariate *t* distribution (Box and Tiao, 1973).

$$p(\boldsymbol{\theta}|y) = \frac{\Gamma\left(\frac{\upsilon + d}{2}\right) \cdot |\boldsymbol{\Phi}^T\boldsymbol{\Phi}|^{1/2} \cdot s^{-d}}{\Gamma\left(\frac{1}{2}\right) \cdot \Gamma\left(\frac{\upsilon}{2}\right) \cdot \upsilon^{d/2}} \cdot \left[1 + \frac{\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_N\right)^2 \boldsymbol{\Phi}^T\boldsymbol{\Phi}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_N\right)}{\upsilon \cdot s^2}\right]^{-\frac{\upsilon+d}{2}}$$

where $\upsilon = N - d$ and $s^2 = \frac{1}{\upsilon}(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^T(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta})$.

Regions in the parameter space for the two dimension case corresponding to the 75%, 90% and 95% levels of this distribution are shown in Fig. D.3(a) as well as the marginal distributions (see Fig. D.3(b)).



(a) **(b)**

Fig. D.3. (a) Posterior credible regions for a reference prior, (b) marginal posteriors

■

Hierarchical priors: Several simultaneous sources of uncertainty (model order, noise variance, …) can be modeled by means *mixture* and *hierarchical* priors. Let us define and illustrate these concepts.

While, in general, a random variable can have only one distribution, it is often easier to model a situation by thinking in terms of a hierarchy. The advantage of the hierarchy is that complicated processes may be modeled by a sequence of relatively simple models placed in a hierarchy. For example, the non-central chi squared distribution with p degrees of freedom and non-centrality parameter λ presents a pdf given by

$$p(x|\lambda,p)=\sum_{k=0}^{\infty}\frac{x^{\frac{p}{2}+k-1}e^{-x/2}}{\Gamma(\frac{p}{2}+k)2^{\frac{p}{2}+k}}\cdot\frac{\lambda^k e^{-\lambda}}{k!}$$

This is a *mixture* distribution, made up of central chi squared densities and Poisson distributions. The *hierarchy* is $X|K\sim\chi^2_{p+2K}$ and K~Poisson(λ). Analogously, by combining normal and gamma models, the final result is a shifted t distribution. In (Igusa *et al*, 2002) such hierarchical models are used in order to differentiate between random and epistemic uncertainties within the structural engineering context.

Dealing with the hierarchy is no more difficult than dealing with conditional and marginal distributions. A useful result is $E[X]=E\big[E[X|K]\big]$. In the previous example, $E[X|K]=p+2K$, so the global mean is $E[X]=p+2\lambda$.

Example D.4. Hierarchical prior for the measurement noise

Most times we deal with normal distributed measurement noise of unknown variance. One possibility is to estimate the noise variance from the experimental data, as (Ljung, 1999a) and (Goodwin *et al*., 2002) do. Another possibility is to assume a probability

distribution for the noise variance itself. Hence, the disturbance class is described by means a hierarchy. For instance,

$$\mathcal{V} = \{\mathbf{v} \in \mathbb{R}^N: \quad v \sim \mathcal{N}(0, \lambda), \;\; \lambda \sim \mathcal{W}^{-1}(m, s)\}$$

where $\mathcal{W}^{-1}(m, s)$ stands for the inverse Wishart distribution, which has the pdf

$$p(\lambda) = \frac{\sigma^{\frac{m}{2}} \cdot e^{-\frac{\sigma}{2\lambda}}}{2^{\frac{m}{2}} \cdot \Gamma\left(\frac{m}{2}\right) \cdot \lambda^{\frac{m+2}{2}}}$$

and it has mean $E[\lambda] = \frac{\sigma}{m-2}$ and variance $Var[\lambda] = \frac{2\sigma^2}{(m-2)^2(m-4)}$.

Remark: In Bayesian statistics the inverse Wishart distribution is widely used since it is the conjugate prior for the covariance matrix of a multivariate normal distribution.

The Wishart distribution $W(\mathbf{\Sigma}, \nu)$ is a generalization of the univariate chi-square distribution $\chi^2(\nu)$ to the multivariate case. The chi-squared is obtained by squaring a random variable distributed as standard normal while the Wishart distribution is obtained analogously from multivariate normal random vectors. The Wishart distribution is often used as a model for the distribution of the sample covariance matrix for multivariate normal random data, after scaling by the sample size.

The inverse Wishart distribution $W^{-1}(\mathbf{\Sigma}, \nu)$, which is based in the Wishart distribution, is used as the conjugate prior for the covariance matrix of a multivariate normal distribution.

■

Nonparametric priors: When the support of the prior model distribution is a high dimensional or infinite dimensional space, such as the spaces ℓ_1 and $\mathcal{H}_\infty$, one must use *nonparametric* priors. They are typically constructed so than the posterior distribution possesses some desirable asymptotic properties such as strong consistency.

For the case the number of parameters (samples of the impulse response h) is finite but grows to infinity with the number of observations N, (McVinnish *et al*., 2006) propose the use of the kernel estimation, where a density is approximated by a mixture

$$\frac{1}{N\sigma} \sum_{i=1}^{N} K\left(\frac{h - h_i}{\sigma}\right)$$

where K is a density function and σ can be estimated by many methods, including wavelet bases and Dirichlet distributions (Robert, 2001).

Nonparametric priors are also used when the set of possible models M_k contain models of varying dimensions. In this case *sieve* priors are useful. A sieve prior is a mixture prior on $\mathcal{F}$ of the form

$$p = \sum_{j=1}^{J} a_j p_j \quad \text{where} \quad a_j \geq 0, \quad \sum_{j=1}^{J} a_j = 1$$

and each p_j is a prior defined on $\mathcal{F}$ but supported on $\mathcal{F}_j$. This is, for instance, the case

$$\sum_{i=1}^{k} p_{ik} \cdot p(M_k|\theta_{ik})$$

where $p(\cdot\,|\theta_{ik})$ is a parameterized density, the sum of the weights p_{ik} sum up to 1, and the number of components k is unknown. This situation is usual in hidden Markov models and other dynamic models, as well as neural networks. The numerical simulation of this kind of models relies on computational tools as the reversible jump Markov Chain Monte Carlo.

D.2 Decision problems

There exist strong connections between *statistical inference* and *decision theory*. Often, they are not distinguished clearly enough from each other. Statistical inference is only a part of the decision making. Statistical inference provides probability distributions conditional on data as a rational basis for decisions. Then, decision theory adds the utility (or risk), calculates expectations, and performs maximization (or minimization). Decision theory also defines formally all parts of the inference problem and the decision making process including desired optimality criteria. These criteria are then used to compare alternative decision procedures. See (Casella and Berger, 2002) and (Robert, 2001) for more insight.

One has to solve a decision problem,

- when one has a reason to choose some single value from the set of possible values of an uncertain quantity. This is the case of the *nominal model* identification.
- when one has to accept as true a single hypothesis from the set of mutually exclusive hypotheses none of which is known to be certainly true. This is the case of *model validation* and *fault detection*.
- when one has to design the data collection procedure. This is the *design of experiments*.

The specific characteristic of Bayesian Decision Theory is that one must start by determining three factors:

1. The distribution family for the observations $p(\mathbf{y}|\boldsymbol{\theta})$,
2. The prior distribution for the parameters $p(\boldsymbol{\theta})$,
3. The loss associated with the decisions $L(\boldsymbol{\theta}, \delta)$ where δ is the decision.

Note that different choices of prior distributions can result in different decisions. This fact is viewed as a drawback by non-Bayesian practitioners. However, for Bayesians,

making explicit the dependence of the decision on the choice of what is believed to be true is an advantage of Bayesian analysis rather than the reverse. Another feature of Bayesian decision problems is that it is often not true that the prior is dominated by the likelihood.

D.2.1 Actions and decisions

To formulate the statistical inference problem as a decision problem and we need some definitions and notation. For simplicity we consider the parametric case. Suppose we perform an experiment and, as a result, we end up with some measured input/output *data*. Output data **y** are random variables belonging to a *sample space*, say $\mathcal{Y}$.

We are interested in finding simple linear time invariant *models* fitting as best as possible the given data. The models will be characterized by a *parameter vector* $\boldsymbol{\theta}$ belonging to a certain *parameter space* Θ.

Action space: Once the data is observed a *decision* regarding $\boldsymbol{\theta}$ is to be made. The set of all allowable decisions is the *action space* $\mathcal{A}$. The action space determines if the problem is a point estimation problem, a set estimation problem, or a hypothesis testing problem:

- When identifying a nominal model, we are performing a point estimation of the parameter vector $\boldsymbol{\theta}$. Actions are guesses at the value of $\boldsymbol{\theta}$. Hence the allowable action space is directly the parameter space, $\mathcal{A} = \Theta$.

- When obtaining an uncertainty region, we are performing a set estimation (interval estimation in the scalar case). In this case actions are intervals or subsets in Θ, that is, the action space $\mathcal{A}$ is formed by all subsets in Θ.

- When validating a model, we are performing a hypothesis testing where the null hypothesis H_0 is the membership of a particular model in the interval estimated at the previous point. Action space is then composed by two elements, $\mathcal{A} = \{a_0, a_1\}$, where a_0 is the action of accepting H_0 and a_1 is the action of rejecting H_0.

Decision rule: A *decision rule* δ is a function from $\mathcal{Y}$ to $\mathcal{A}$ that specifies, for each $\mathbf{y} \in \mathcal{Y}$, what action $a \in \mathcal{A}$ will be taken if **y** is observed. Thus, in a hypothesis testing setup the decision $\delta(\mathbf{y})=a_0$ will be taken for each **y** that is in the acceptance region of the test. All the allowable δ form the *decision space* $\mathcal{D}$ and the selection of a particular δ will be decided regarding optimality properties in some sense.

In Bayesian decision analysis, it is supposed that a choice has to be made from a set of available actions $(a_1,\ldots,a_r)$, where the payoff or utility of a given action depends on a state of nature $\boldsymbol{\theta}$ which is unknown. The decision maker's knowledge of $\boldsymbol{\theta}$ is represented by a posterior distribution which combines prior knowledge of $\boldsymbol{\theta}$ with the

information provided by an experiment, and he is the supposed to choose that action which maximizes the expected payoff over the posterior distribution.

D.2.2 Conditional Bayes principle

The decision a, $a \in \mathcal{A}$, may be correct, incorrect but not too wrong, or grossly incorrect. A way to quantify the correctness is the *loss function*, $L(\boldsymbol{\theta}, a)$, which will be greater as a become more incorrect.

Loss function: Suppose that the true model is G_{true}. The identification of a nominal model can be viewed as a decision on the basis of data $\delta(\mathbf{y})$. The loss $L(G_{true}, \delta(\mathbf{y}))$ is usually a function of the estimation error, e.g.,

$$L(G_{true}, G) = \|G_{true} - G\|_S^p \tag{144}$$

where $\|\cdot\|_S$ is some suitable norm in the system space.

Risk function: The risk or conditional risk $R(G_{true}, G)$ associated to an estimator G is defined as the average loss with respect to data $\mathbf{y}$, that is,

$$R(G_{true}, G) = E_y[L(G_{true}, G)] = \int_y L(G_{true}, G)p(\mathbf{y}|G)d\mathbf{y} \tag{145}$$

The conditional risk R is preferred to the loss L since it accounts for the likelihood of observations $\mathbf{y}$. In other words, we will not be especially concerned with large values of L if the likelihood of occurrence of $\mathbf{y}$ is small.

Bayesian risk: The Bayesian risk $\mathcal{R}(G_{true}, G)$ is defined as the average risk with respect to the prior $p(G)$

$$\mathcal{R}(G_{true}, G) = E_G[R(G_{true}, G)] = \int_{\mathcal{G}} R(G_{true}, G)p(G)dG \tag{146}$$

Substituting (145) in (146) , we have

$$\mathcal{R}(G_{true}, G) = E_G[R(G_{true}, G)] = \int_{\mathcal{G}} \int_y L(G_{true}, G)p(\mathbf{y}|G)p(G)d\mathbf{y}\, dG$$

Now, since $p(\mathbf{y}|G)p(G) = p(G|\mathbf{y})p(\mathbf{y})$, we can split the previous integration into two parts,

$$\mathcal{R}(G_{true}, G) = E_G[R(G_{true}, G)] = \int_y \left(\int_{\mathcal{G}} L(G_{true}, G) p(G|\mathbf{y}) dG \right) p(\mathbf{y}) d\mathbf{y}$$

Since both $L(G_{true}, G)$ and $p(G|\mathbf{y})$ are positive, the inner integral is positive for any **y**. Furthermore, since $p(\mathbf{y})$ also is positive, the value of G that minimizes the risk is the value that minimizes the inner integral,

$$\hat{G} = \arg\min_G \int_{\mathcal{G}} L(G_{true}, G) p(G|\mathbf{y}) dG \tag{147}$$

This is the optimal choice for $\delta(\mathbf{y})$. A necessary condition for such a minimum is $\frac{\partial}{\partial G} \int_{\mathcal{G}} L(G_{true}, G) p(G|\mathbf{y}) dG \Big|_{G=\hat{G}} = 0$. (Eykhoff, 1974)

Conditional Bayes Principle: Bayesian decision theory attempts to minimize the Bayes risk. The conditional Bayes principle states that an action should be chosen which minimizes the Bayesian expected loss (Berger, 1985). The *Bayes rule* with respect to a prior $p(G)$ is the decision rule $\hat{\delta}(\mathbf{y})$ that minimizes $\mathcal{R}$ among all possible δ*'s*.

$$\mathcal{R}\left(p(G), \hat{\delta}\right) = \inf_{\delta \in \mathcal{D}} \mathcal{R}(p(G), \delta) \tag{148}$$

Typically one may find a unique Bayes rule, but there may be no one or many. Bayes rules can be found by the application of some useful theorems (see (Casella and Berger, 2002)).

Finally, note the role of the prior distribution on the definition of the Bayesian risk $\mathcal{R}$: From a subjective point of view, this prior reflects the beliefs of the experimenter about the value of $\boldsymbol{\theta}$ prior to data collection. From a decision theoretic point of view, subjective prior $p(\boldsymbol{\theta})$ is just a weight function:

- selecting $\boldsymbol{\theta}$ such as $p(\boldsymbol{\theta})$ is large, the experimenter would like to have particularly small risk;
- selecting $\boldsymbol{\theta}$ such as $p(\boldsymbol{\theta})$ is small, the experimenter is not concerned about the risk.

References and Bibliography

(Akçay and Ninness, 1999) Akçay, H. and B. Ninness, "Orthonormal basis functions for modelling continuous-time systems", *Signal Processing*, **77**, 261-274, 1999.

(Akçay, Hjalmarsson, and Ljung, 1996) H. Akçay, H. Hjalmarsson, and L. Ljung, "On the choice of norms in system identification", *IEEE Transaction on Automatic Control*, **41**(6):1367-1372, 1996.

(Alamo, Bravo and Camacho, 2005) T. Alamo, J. Bravo, and E. Camacho, "Guaranteed state estimation by zonotopes", *Automatica*, 41(6), 1035–1043, 2005.

(Andrieu *et al*., 2001) C. Andrieu, P.M. Djuric, and A. Doucet, "Model selection by MCMC computation", *Signal Processing*, **81**, 19-37, 2001.

(Andrieu et al., 2010) C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods", *Journal of the Royal Stats. Society*, Series B, **72**, Part 3, pp. 269–342, 2010.

(Apeland *et al*., 2002) S. Apeland, T. Aven, and T. Nilsen, "Quantifying uncertainty under a predictice, epistemic approach to risk analysis", *Reliability eng and syst safety*, **75**, (2002), 93-102.

(Balas, 1982) M. J. Balas, "Trends in Large Space Structures Control Theory: Fondest Hopes, Wildest Dreams", *IEEE Trans. Automat Contr*., Ac-27, pp.522-535, June 1982.

(Baldelli, Mazzaro, and Sánchez Peña, 2001) D.H. Baldelli, M.C. Mazzaro, and R. S. Sánchez Peña, "Robust identification of lightly damped flexible structures by means orthonormal bases", *IEEE Transactions on Control Systems Technology*, **9**(5), 2001.

(Barron, Schervish, and Wasserman, 1999) A. Barron, M.J. Schervish and L. Wasserman, "The consistency of posterior distributions in nonparametric problems", *Ann. Statist*., vol.27, no.2, pp.536-561, 1999

(Bayes, 1763a) T. Bayes, "An essay towards solving a problem in the doctrine of chances", (1763), *Philosophical Transactions of the Royal Society*, **53**: 370-418, 1973.

(Bayes, 1763b) T. Bayes, "A letter from the late Reverend Mr. Thomas Bayes to John Canton", (1763), *M.A. & F.R.S. Philosophical Transactions of the Royal Society*, **53**: 269-271, 1973.

(Berger and Rios Insuna, 1998) Berger, J.O., Rios Insua, D., "Recent developments in Bayesian inference with applications in hydrology". In: *Statistical and Bayesian Methods in Hydrological Sciences*. Selected papers from the International

Conference in honor of Professor Jacques Bernier, UNESCO IHP-V, Technical Document in Hydrology No. **20**, 11–13 September 1995, pp. 43–61, 1998.

(Berger, 1985) J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer-Verlag, 1985.

(Berger, 2000) J. O. Berger, "Bayesian analysis: A look at today and thoughts of tomorrow*", Journal of the American Statistical Association*, **95**(452), Dec. 2000.

(Bergman, 1999) Bergman, N., *Recursive Bayesian Estimation: Navigation and Tracking Applications*, Linköping University, 1999. Sup: L. Ljung and Gustafsson

(Blanke *et al*., 2003) M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Diagnosis and Fault-Tolerant Control*, Springer, 2003.

(Blesa *et al*., 2011b) J. Blesa, V. Puig, and J. Saludes, "Identification for passive robust fault detection using zonotope-based set-membership approaches", *International Journal of Adaptive Control and Signal Processing*, **25**(9):788–812, 2011.

(Blesa, 2011a) J. Blesa, *Robust Identification and fault Diagnosis using Set-membership Approaches*, Ph. D Thesis, Supervisors: V. Puig and J. Saludes, Upc, 2011.

(Blesa, Puig, and Saludes, 2013) J. Blesa, V. Puig, and J. Saludes, "Robust Fault Detection using Polytope-based Set-membership Consistency Test", *IET Control Theory and Applications*, To appear, 2013

(Bolstad, 2010) W. M. Bolstad, *Understanding Computational Bayesian Statistics*, John Wiley, 2010.

(Borsuk *et al*., 2004) Mark E. Borsuk, Craig A. Stow and Kenneth H. Reckhow, "A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis", Ecological Modelling, Volume 173, Issues 2-3, 1 April 2004, Pages 219-239.

(Box and Tiao, 1973) Box, G.E.P., Tiao, G.C., *Bayesian Inference in Statistical Analysis*, John Wiley, 1973.

(Brooks *et al*., 2003) S.P. Brooks, P. Giudici, and G.O. Roberts, "Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions", *J. R. Statist. Soc. B*, **65**, Part 1, pp. 3-55.

(Brooks, 1998) S. P. Brooks, "Markov chain Monte Carlo method and its application", *The Statistician*, **47**, Part 1, pp.69-100, 1998.

(Cairns, 2000) Cairns, A.J.G., "A discussion of parameter and model uncertainty in insurance", *Insurance: Mathematics and Economics*, **27**, 313-330, 2000.

(Casella and Berger, 2002) Casella, G., Berger, R.L., *Statistical inference*, Duxbury Thomson Learning, 2nd ed., 2002.

(Chaloner and Verdinelli, 1995) Chaloner, K., Verdinelli, I., "Bayesian Experimental Design: A Review", *Statistical Science*, vol.**10**, no.3, 273-304, 1995.

(Chen and Gu, 2000) J. Chen and G. Gu, *Control-Oriented System Identification. An $\mathcal{H}_\infty$ Approach*, John Wiley, 2000.

(Chen and Patton, 1999) J. Chen and R. Patton, *Robust Model-Based Fault Diagnosis for Dynamic Systems*, Kluwer Academic Publishers, 1999.

(Chen, Shao, and Ibrahim, 2000) M.-H. Chen, Q.-M. Dhao, and J. G. Ibrahim, *Monte Carlo Methods in Bayesian Computation*, Springer Series in Statistics, Springer, 2000.

(Chiang *et al*., 2008) R. Chiang, M.G. Safonov, G. Balas, and A. Packard, *Robust Control Toolbox*, 3rd ed. Natick, MA: The Mathworks, Inc., 2007

(Chisci *et al*., 1998) L. Chisci, A. Garulli, A. Vicino, and G. Zappa, "Block recursive parallelotopic bounding in set-membership identification", *Automatica*, **34**(1):15-22, 1998.

(Cooley and Lee, 1998) B. L. Cooley and J. H. Lee, "Integrated identification and robust control", *J. Proc. Control*., vol. **8**, 1998.

(Cornford, 2004) D. Cornford, "A Bayesian state space modelling approach to probabilistic quantitative precipitation forecasting", Journal of Hydrology, Volume 288, Issues 1-2, 20 March 2004, Pages 92-104.

(Dearden, 2010) R. Dearden, "Bayesian Fault Diagnosis: Common Approaches and Challenges", *2nd International Workshop on Cognitive Information Processing*, 2010

(Dellaportas and Forster, 1999) P. Dellaportas and J.J. Forster, "Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models", *Biometrika*, **86**(3):615-633, 1999.

(Ding, 2008) S. Ding, *Model-Based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools*, Springer, 2008.

(Douma and Van den Hof, 2005) S. G. Douma and P. M. J. Van den Hof, "Relations between uncertainty structures in identification for robust control", *Automatica*, **41**, March 2005.

(Earman, 1992) Earman, J. *Bayes or Bust? A Critical examination of Bayesian confirmation theory*, MIT Press, Cambridge, MA, 1992.

(Edwards, Lindman, and Savage, 1963) Edwards, W., H. Lindman, and L. J. Savage. "Bayesian statistical inference for psychological research". *Psychological Review* 70:193–242, 1963.

(Esmaeilsabzali *et al*., 2006) H. Esmaeilsabzali *et al*., "Robust identification of a lightly damped flexible beam using set-membership and model error modelling techniques", *Proc. of the 20006 IEEE Int. Conf on Control App.*, Germany, October, 2006.

(Eykhoff, 1974) P. Eykhoff, *System Identification. Parameter and State Estimation*, John Wiley, 1974.

(Fan and Sisson, 2010) Y. Fan and S. A. Sisson, "Chapter 1. Reversible Jump Markov chain Monte Carlo", Arxiv num. 1001.2055v1, Stat.ME, 2010.

(Fernández-Cantí, Blesa, and Puig, 2013) R.M. Fernández-Cantí, J. Blesa, and V. Puig, "Set-membership Identification and Fault Detection using a Bayesian Framework", submitted to the *European Control Conference* (ECC 2013), Zurich, Switzerland, 2013.

(Fogel and Huang, 1982), E. Fogel and F. Huang, "On the value of information in system identification –bounded noise case", *Automatica*, **18**:229-238, 1982.

(Garulli and Reinelt, 2000) A. Garulli and W. Reinelt, "On model error modelling in set-membership identification", *Proc of the SYSID*, 2000.

(Garulli, Vicino, and Zappa, 2000) Garulli, A., A. Vicino, G. Zappa, "Conditional central algorithms for worst-case set-membership identification and filtering", *IEEE Transactions on Automatic Control*, vol.**45**, no.1, pp.14-23, 2000.

(Gelfand et al, 1992) Gelfand, A., Dey, D., Chang, H., "Model determination using predictive distributions with implementation via sampling-based methods". In: Bernardo, J.M. et al. (Eds.), *Bayesian Statistics*, Vol. **4**. Oxford University Press, Oxford, pp. 147–158, 1992.

(Ghosal and van der Vaart, 2007) S. Ghosal and A.W. van der Vaart, "Convergence rates of posterior distributions for noniid observations", *Annals of Statistics*, vol. 35, no. 1, pp192-223, 2007.

(Gilks et al., 1996) Gilks, W., Richardson, S., Spiegelhalter, D., *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

(Girard and Parent, 2004) P. Girard and E. Parent, "The deductive phase of statistical analysis via predictive simulations: test, validation and control of a linear model

with autocorrelated errors representing a food process", *Journal of Statistical Planning and Inference*, **124**: 99-120, 2004.

(Gómez and Baeyens, 2004) J.C. Gómez and E. Baeyens, "Identification of block-oriented nonlinear systems using orthonormal bases", *Journal of Process Control*, **14**, 685-697, 2004.

(Gómez, 1998) J. C. Gómez, *Analysis of Dynamic System Identification using Rational Orthonormal Bases*, Ph. D. Thesis, The University of Newcastle, Australia, 1998.

(Goodwin and Payne, 1977) G.C. Goodwin and R.L. Payne, *Dynamic System Identification*, Academic Press, 1977.

(Goodwin, Braslavsky, and Seron, 2002) G.C. Goodwin, J.H. Braslavsky, and M.M. Seron, "Non-stationary stochastic embedding for transfer function estimation", *Automatica*, **38**: 47-62, 2002.

(Green and Hastie, 2009) P. J. Green and D. I. Hastie, "Reversible jump MCMC", Draft, June 13, 2009.

(Green, 1995) Green, P., "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", *Biometrika*, **82**:711-732, 1995

(Gu and Khargonekar, 1992) G. Gu and P. P. Khargonekar, "Linear and nonlinear algorithms for identification in H∞ with error bounds", *IEEE Transactions on Automatic Control*, **37**: 953-963, 1992.

(Gustafsson and Hjalmarsson, 1995), F. Gustafsson, and H. Hjalmarsson, "Twenty-one ML estimators for model selection", *Automatica*, 31(10):1377-1392, 1995.

(Gustafsson and Mäkilä, 1993) T.K. Gustafsson and P.M. Mäkilä, "On system identification and model validation via linear programming", *Proc. of the 32nd conference on decision and control*, San Antonio, Texas, December, 1993.

(Gustafsson and Mäkilä, 1994) T. Gustafsson and P.M. Mäkilä, *ℓ_1-identification toolbox for Matlab*, Åbo Akademi, Finland, 1994.

(Gustafsson and Mäkilä, 1996), Gustafsson, T.K., Mäkilä, P.M., "Modelling of uncertain systems via linear programming", *Automatica*, **32**: 219-335, 1996.

(Gustafsson and Mäkila, 2001) Gustafsson, T.K. and P.M. Mäkilä, "Modelling of uncertain systems with application to robust process control", *Journal of Process Control*, **11**: 251-264, 2001.

(Hakvoort and Van den Hof, 1997) R. G. Hakvoort and P. M. J. Van den Hof, "Identification of probabilistic system uncertainty regions by explicit evaluation of bias and variance errors", *IEEE Transactions on Automatic Control*, **42**(11), 1997.

(Hastings, 1970) W. Hastings, "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, **57**, pp.97-109.

(Helmicki, Jacobson, and Nett, 1991) A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control-oriented system identification: A worst-case/deterministic approach in $\mathcal{H}_\infty$", *IEEE Transactions on Automatic Control*, **36**(10): 1163-1176, 1991.

(Herrero, 2006) J. M. Herrero Durá, *Identificación Robusta de Sistemas no Lineales mediante Algoritmos Evolutivos*, Ph. D. Thesis, Spvs.: M. Martínez Iranzo and X. Blasco Ferragud, Universidad Politécnica de Valencia, 2006.

(Heuberger, Van den Hof, and Bosgra, 1995) P.S.C. Heuberger, P.M.J. Van den Hof, and O.H. Bosgra, "A generalised orthonormal basis for linear dynamical systems", *IEEE Transactions on Automatic Control*, **40**(3):451-465, March 1995.

(Hildebrand and Gevers, 2003) Hildebrand, R. and M. Gevers, "Identification for control: optimal input design with respect to a worst-case ν-gap cost function", *SIAM Journal Control Optim*, **41**(5), 1586-1608, 2003.

(Hjalmarsson and Gustafsson, 1995) H. Hjalmarsson and F. Gustafsson, "Composite Modelling of Transfer Functions", *IEEE Transactions on Automatic Control*, vol. **40**, no.5, May 1995.

(Hjalmarsson, 2005) H. Hjalmarsson, "From experiment design to closed-loop control", *Automatica*, **41**, March 2005.

(Hoeting *et al*., 1999) J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky, "Bayesian Model Averaging: A tutorial", *Statistical Science*, vol.**14**, no. 4, 382-417, 1999.

(Hoeting *et al*., 1999) J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky, "Bayesian Model Averaging: A tutorial", *Statistical Science*, vol.**14**, no. 4, 382-417, 1999.

(Horwich, 1982) Horwich, P., *Probability and Evidence*, Cambridge University Press, 1982.

(Houpis, Rasmussen, and García-Sanz, 2006) C. H. Houpis, S. J. Rasmussen, M. García-Sanz, Quantitative Feedback Theory. Fundamentals and Applications, 2nd ed., CRC Press, 2006

(Howson and Urbach, 1989) Howson C. and P. Urbach, *Scientific Reasoning. The Bayesian Approach*. La Salle, Ill.: Open Court, 1989.

(Igusa *et al*, 2002) Igusa. T., S.G.Buonoparte, and B.R., Ellingwood, "Bayesian analysis of uncertainty for structural engineering applications", *Structural Safety*, **24**, 165-186, 2002.

(Igusa, Buonoparte, and Ellingwood, 2002) Igusa. T., S.G.Buonoparte, and B.R., Ellingwood, "Bayesian analysis of uncertainty for structural engineering applications", *Structural Safety*, **24**, 165-186, 2002.

(Jaulin, 2010) L. Jaulin, "Probabilistic set-membership approach for robust regression", *Journal of Statistical Theory and Practice*, **4**(1), 2010.

(Jeffrey, 2004) R. Jeffrey, *Subjective Probability: The Real Thing*, Cambridge University Press, 2004.

(Johansson, 2000) K.H. Johansson, "The Quadruple-Tank Process: A Multivariable Laboratory Process with an Adjustable Zero", *IEEE Transactions on Control Systems Technology*, vol. **8**, no. 3, May 2000.

(Kass and Wasserman, 1996) Kass. R. and Wasseman, L., "The selection of prior distributions by formal rules", *Journal of the American Statistical Association*, **91**:1343-1370, 1996.

(Katafygiotis and Beck, 1998) Katafygiotis, L.S., and J.L. Beck, "Updating models and their uncertainties II: Model Identifiability", *J. Eng. Mech., ASCE*, **124** (4), 463-467, 1998.

(kk-electronic, 2012) http://www.kk-electronic.com/Default.aspx?ID=9589.

(Landau, Karimi, and Hjalmarsson, 2003) I. D. Landau, A. Karimi, and H. Hjalmarsson, "Design and optimisation of restricted complexity controllers", *European Journal of Control Special Issue*, vol.9, no.1, 2003. Benchmark website: http://lawww.epfl.ch/page11534.html

(Lee, 2008) Y. K. Lee, *A Fault Diagnosis Technique for Complex Systems using Bayesian Data Analysis*, Ph. D. Thesis, Georgia Institute of Technology, 2008.

(Lehman and Casella, 1998) E.L. Lehmann and G. Casella, *Theory of Point Estimation*. *Springer Texts in Statistics*, Springer, 2nd ed., 1998.

(Letellier *et al*., 2011) C. Letellier, G. Hoblos, and H. Chafouk, "Robust Fault Detection based on Multimodel and Interval Approach. Application to a Throttle Valve", *19th Mediterranean Conference on Control and Automation*, Corfu, Greece, June 2011.

(Ljung, 1995) Ljung, L., *System Identification Toolbox. User's guide*, 4th ed., The MathWorks, 1995.

(Ljung, 1997) Ljung, L., "Identification, model validation and control", *In 36th Conference on Decision and Control. Plenary lecture*, 1997.

(Ljung, 1999a) Ljung, L., *System Identification - Theory for the User*, 2nd ed., Prentice Hall, 1999.

(Ljung, 1999b) L. Ljung, "Model validation and model error modeling", Control and Communications Group Technical Report LiTH-ISY-R-2125, Linkoping University, 1999. *In Proc. of the Aström Symposium on Control*, pages 15-42, Lund, Sweden, 1999.

(Malan *et al.*, 2001) S. Malan, M. Milanese, D. Regruto, and M. Taragna, "Robust control from data via uncertainty model sets identification", *Proc of the 40th IEEE Conference on Decision and Control,* Orlando, Florida, USA, December 2001.

(Maraoui and Messaoud, 2001) S. Maraoui and H. Messaoud, "Design and comparative study of limited complexity bounding error identification algorithms", *IFAC Symposium on System Structure and Control*, Prague, 29-31, 2001.

(Mazzaro, Parrilo, and Sánchez Peña, 2004) M.C. Mazzaro, P.A. Parrilo, and R.S. Sánchez Peña, "Robust Identification Toolbox", *Latin American Applied Research*, 34:91-100, 2004.

(Mbarek, Messaoud, and Favier, 2003) A. Mbarek, H. Messaoud, and G. Favier, "New algorithm for parallelotope updating in output error bounding", *Proc of the ICECS-2003*, pp.758-761, 2003.

(McVinish, Braslavsky, and Mengersen., 2006) R. McVinish, J.H. Braslavsky, and K. Mengersen, "A Bayesian-decision theoretic approach to model error modelling", 14th *IFAC SYSID*, 2006.

(Messaoud and Favier, 1994) H. Messaoud and G. Favier, "Recursive determination of parameter uncertainty intervals for linear models with unknown but bounded errors", *10th IFAC Symposium on System Identification*, Copenhagen, pp.365-370, 1994.

(Metropolis et al., 1953) M. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, "Equations of state calculations by fast computing machines", *Journal of Chemical Physics*, **21**, pp. 1087-1091.

(Milanese and Taragna, 2002) M. Milanese and M. Taragna, "Optimality, approximation, and complexity in set-membership $\mathcal{H}_\infty$ identification", *IEEE Transactions on Automatic Control*, **47**(10):1682-1690, 2002.

(Milanese and Taragna, 2005) M. Milanese and M. Taragna, "$\mathcal{H}_\infty$ set-membership identification: a survey", *Automatica*, **41**(12):2019-2032, December 2005.

(Milanese and Vicino, 1991) M. Milanese and A. Vicino. "Estimation theory for nonlinear models and set-membership uncertainty". *Automatica*, **27**:403-408, 1991.

(Milanese *et al*, 1996) Milanese, M., J.P. Norton, H. Piet-Lahanier, and E. Walter, editors, *Bounding approaches to System Identification*, Plenum Press, New York, USA, 1996.

(Milanese, 1998) M. Milanese, "Learning models from data: the set-membership approach", *Proc. American Control Conf.*, (**1**):178-182, Philadelphia, USA, 1998.

(Mo and Norton, 1990) S. H. Mo and J.P. Norton, "Fast and robust algorithm to compute exact polytope parameter bounds", *Mathematics and Computers in Simulation*, vol.**32**, pp.481-493, 1990.

(Ninness and Goodwin, 1995) Ninness, B., Goodwin, G.C., "Estimation of model quality", *Automatica*, vol.**31**, no.12, pp.1771-1797, 1995.

(Ninness and Gustafsson, 1997) Ninness, B., and F. Gustafsson, "A unifying construction of orthonormal bases for system identification", *IEEE Transactions on Automatic Control*, **42**: 515:521, 1997.

(Ninness and Henriksen, 2010) B. Ninness and S. Henriksen, "Bayesian system identification via Markov chain Monte Carlo techniques", *Automatica*, **46**:40-51, 2010.

(Ninness and Hjalmarsson, 2003a) Ninness, B., and H. Hjalmarsson, "On the frequency domain accuracy of closed loop estimates", *Proceedings of the 42nd IEEE Conference on Decision and Control*, Maui, Hawaïi, pp.5997-6002, December 2003.

(Ninness and Hjalmarsson, 2003b) Ninness, B., and H. Hjalmarsson, "Variance error quantifications that are exact for finite model order", *Proceedings of the 42nd IEEE Conference on Decision and Control*, Maui, Hawaïi, pp.6003-6008, December 2003.

(Ninness, 2009) B. Ninness, "Some System Identification Challenges and Approaches", Plenary Lecture, *15th IFAC Symposium on System Identification* (SYSID), Saint-Malo, France, July 6-8, 2009.

(Ninness, Hjalmarsson, and Gustafsson, 1999) Ninness, B., H. Hjalmarsson, and F. Gustafsson, "The fundamental role of general orthonormal bases in system identification", *IEEE Transactions on Automatic Control*, **44**(7), July 1999.

(Odgaard and Johnson, 2012) P.F. Odgaard and K. E. Johnson, "Wind turbine fault detection and fault tolerant control – a second challenge", available at http://www.kk-electronic.com/Default.aspx?ID=9589.

(Odgaard, Stoustrup, and Kinnaert, 2009) P.F. Odgaard, J. Stoustrup, and M. Kinnaert, "Fault tolerant control of wind turbines - a benchmark model", In *Proceedings of 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes*, Barcelona, Spain, 2009.

(Onatski and Williams, 2002) Onatski, A., N. Williams, "Modelling Model Uncertainty", Working Paper no.169, *European Central Bank Working Paper Series, International Seminar on Macroeconomics*, August 2002.

(Papadimitriou, Beck, and Katafygiotis, 2001) Papadimitriou, C., J.L. Beck, and L.S. Katafygiotis, "Updating robust reliability using structural test data", *Probabilistic Engineering Mechanics*, **16**, 203-113, 2001.

(Parrilo, Sánchez Peña, and Sznaier, 1999) P.A. Parrilo, R.S. Sánchez Peña, and M. Sznaier, "A parametric extension of mixed time/frequency robust identification", *IEEE Transactions on Automatic Control*, **44**(2):364-369, 1999.

(Pernestål, 2009) Pernestål, A., *Probabilistic Fault Diagnosis with Automotive Applications*, Ph. D. Thesis, Spv.: M. Nyberg, Linköping University, 2009

(Peterka, 1981) V. Peterka, "Bayesian System Identification", *Automatica,* **17**(1):41-53, 1981.

(Ploix and Adrot, 2006) S. Ploix and O. Adrot, "Parity relations for linear uncertain dynamic systems", *Automatica*, **42**(9): 1553-1562, 2006

(Puig *et al*., 2008) V. Puig, J. Quevedo, T. Escobet, F. Nejjari, and S. de las Heras, "Passive Robust Fault Detection of Dynamic Processes Using Interval Models", IEEE Trans. on Control Systems Technology, **16**(5): 1083-1089, 2008.

(Qian, Stow, and Borsuk, 2003) Song S. Qian, Craig A. Stow and Mark E. Borsuk, "On Monte Carlo methods for Bayesian inference", Ecological Modelling, Volume 159, Issues 2-3, 15 January 2003, Pages 269-277.

(Raafat *et al*., 2009) S. M. Raafat *et al*., "Robust identification of a single axis high precision positioning system", *2009 5th Int Colloquium on Signal Processing & Its Applications* (CSPA), 2009.

(Raïssi, Videau, and Zolghadri, 2010) T. Raïssi, G. Videau and A. Zolghadri, "Interval observer design for consistency checks of nonlinear continuous-time systems", *Automatica*, 2010.

(Reinelt, Garulli, and Ljung., 2002) Reinelt, W., A. Garulli, and L. Ljung, "Comparing different approaches to model error modelling in robust identification", *Automatica*, **38**, 2002.

(Reppa and Tzes, 2011) R. Reppa and A. Tzes, "Fault detection and diagnosis based on parameter set estimation", *IET Control Theory Appl*., **5**: 69-83, 2011.

(Robert and Casella, 1999) Robert, C.P., Casella, G., *Monte Carlo Statistical Methods*. Springer, New York, 1999.

(Robert, 2001) C.P. Robert. *The Bayesian Choice*. 2nd ed., Springer Texts in Statistics. Springer Verlag, 2001.

(Rosenkrantz, 1977) Rosenkrantz, R. D., *Inference, Method and Decision. Towards a Bayesian Philosophy of Science*. Dordrecht, Reidel., 1977.

(Sánchez Peña and Sznaier, 1998) Sánchez Peña, R.S. and M. Sznaier, *Robust Systems Theory and Applications*, John Wiley & Sons, Inc., 1998.

(Schön, Wills, and Ninness, 2011) T.B. Schön, A. Wills and B. Ninness, "System identification of nonlinear state-space models", *Automatica*, **47**:39-49, 2011.

(Schoukens and Pintelon, 1991) Schoukens, J., Pintelon, R., *Identification of linear systems: A practical guideline for accurate modeling*, London, Pergamon Press, 1991.

(Seron, Braslavsky, and Goodwin, 1997) M. M. Seron, J.H. Braslavsky, G.C. Goodwin, *Fundamental Limitations in Filtering and Control. Communications and Control Engineering Series*, Springer, 1997.

(Sjöberg *et al*., 1995) J. Sjöberg, *et al*., "Nonlinear black-box modelling in system identification: a unified overview", *Automatica*, vol.**31**, no.12, 1995.

(Skogestad and Postlethwaite, 1996) S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control. Analysis and Design*, John Wiley, 1996.

(Söderström and Stoica, 1989), *System Identification*, Prentice Hall, Englewood Cliffs, NJ, 1989.

(Soudappan *et al*., 2004) Prabhu Soundappan, Efstratios Nikolaidis, Raphael T. Haftka, Ramana Grandhi and Robert Canfield, "Comparison of evidence theory and Bayesian theory for uncertainty modelling", Reliability Engineering & System Safety, Volume 85, Issues 1-3, July-September 2004, Pages 295-311.

(Tanner, 1996) Tanner, M.A., *Tools for Statistical Inference: Methods for the Exploration of Posterior Distribution and Likelihood Functions*. Springer, New York, 1996.

(Tharrault *et al*., 2009) Y. Tharrault, G. Mourot, J. Ragot, "WWTP Diagnosis based on Robust Principal Component Analysis", *Proc. of the 7th IFAC Symp. on Fault Detection, Supervision and Safety of Technical Processes*, Barcelona, Spain, 2009.

(Tierney, 1994) L. Tierney, "Markov chains for exploring posterior distributions", *The Annals of Statistics*, vol.**22**, no.4, 1701-1762, 1994.

(Tjärnström, 2002) Tjärnström, F., *Variance expressions and model reduction in system identification*, Ph D. Thesis, Linköping University, 2002.

(Traub, Wasilkowski, and Wozniakowski, 1988) J.F. Traub, H. Wasilkowsi, and H. Wozniakowsi, *Information-Based Complexity*, New York: Academic Press, 1988.

(Utkin, 2003) Utkin, L.V., "A second-order uncertainty model for calculation of the interval system reliability", *Realiability Engineering and System Safety*, **79**, 341-351, 2003.

(Van den Hof and Schrama, 1995) P.M.J. Van den Hof and R.J.P. Schrama, "Identification and control – closed loop issues", *Automatica*, **31**: 1751-1770, 1995.

(Verhaegen and Verdult, 2007) M. Verhaegen and V. Verdult, *Filtering and System Identification: A Least Squares Approach*, Cambridge University Press, 2007.

(Vicino and Zappa, 1996) A. Vicino and G. Zappa, "Sequential approximation of feasible parameter sets for identification with set membership uncertainty", *IEEE Transactions on Automatic Control*, 41:774–785, 1996.

(Wahlberg and Ljung, 1986) B. Wahlberg and L. Ljung, "Design variables for bias distribution in transfer function estimation", *IEEE Trans. Automat. Contr*., vol. AC-31, pp.134-144, 1986.

(Wahlberg, 1991) Wahlberg, B., "System identification using Laguerre models", *IEEE Transactions on Automatic Control*, **36**: 551-562, 1991.

(Wahlberg, 1994) Wahlberg, B., "System identification using Kautz models", *IEEE Transactions on Automatic Control*, **39**: 1276-1282, 1994.

(Walhberg and Ljung, 1992) B. Wahlberg and L. Ljung, "Hard frequency-domain model error bounds from least-squares like identification techniques", *IEEE Transactions on Automatic Control*, vol. **AC-37**, no.7, pp.900-912, 1992.

(Zanardelli, Strangas, and Aviyente, 2007) W. Zanardelli, E. Strangas, S. Aviyente, "Identification of Intermitent Electrical and Mechanical Faults in Permanent-Magnet AC Drives Based on Time-Frequency Analysis", *IEEE Trans. on Industry Applications*, **43**(4): 971-980, 2007.

(Zhou with Doyle, 1998) K. Zhou, with J.C. Doyle, *Essentials of Robust Control*, Prentice-Hall, 1998.

Index