Universitat de Girona

AVENÇOS EN ELS FONAMENTS MATEMÀTICS DE L'ANÀLISI COMPOSICIONAL DE DADES: CONVEXITAT I NORMES L^p. APLICACIÓ A LA REGRESSIÓ LINEAL LASSO AMB COVARIABLE COMPOSICIONAL

Jordi Saperas i Riera

http://creativecommons.org/licenses/by/4.0/deed.ca

Aquesta obra està subjecta a una llicència Creative Commons Reconeixement

Esta obra está bajo una licencia Creative Commons Reconocimiento

This work is licensed under a Creative Commons Attribution licence

Universitat de Girona

TESI DOCTORAL

Avenços en els fonaments matemàtics de l'anàlisi composicional de dades: convexitat i normes L^p. Aplicació a la regressió lineal LASSO amb covariable composicional

Jordi Saperas i Riera

2024

Universitat de Girona

TESI DOCTORAL

Avenços en els fonaments matemàtics de l'anàlisi composicional de dades: convexitat i normes L^p. Aplicació a la regressió lineal LASSO amb covariable composicional

Jordi Saperas i Riera

2024

Programa de Doctorat en Tecnologia

Directors

Dra. Glòria Mateu Figueras

i

Dr. Josep A. Martín Fernández

Memòria presentada per optar al títol de doctor per la Universitat de Girona

Publicacions

Aquesta tesi es presenta com a compendi dels següents articles:

- Saperas-Riera, J., Martín-Fernández, J.A., Mateu-Figueras, G. (2023), **Fundamentals of convex optimization for compositional data.** Statistics and Operations Research Transactions, 47 (2), pp. 323-344. Factor d'impacte: 1.6, segon quartil (Q2) del *Journal Citation Report* (JCR) de l'*Institute of Scientific Information.*

- Saperas-Riera, J., Martín-Fernández, J.A., Mateu-Figueras, G. (2023), **Lasso regression method for a compositional covariate regularised by the norm L^1 pairwise logratio**. Journal of Geochemical Exploration, 255. Factor d'impacte: 3.9, primer quartil (Q1) del *Journal Citation Report* (JCR) de l'*Institute of Scientific Information.*

- Saperas-Riera, J., Mateu-Figueras, G., Martín-Fernández, J.A. (2024), **L^p-norm for compositional data: exploring the CoDa L^1-norm in penalised regression**. Mathematics, 12(9): 1388. Factor d'impacte: 2.4, primer quartil (Q1) del *Journal Citation Report* (JCR) de l'*Institute of Scientific Information.*

Del treball de la tesi se n'han derivat les següents aportacions a congressos:

- 50th meeting of the Italian Statistical Society (SIS 2021). **What is a convex set on compositional data analysis?** Jordi Saperas-Riera, Josep Antoni Martín-Fernández. On-line, Italy.

- 5th International Conference on Econometrics and Statistics (EcoSta 2022). **Contributions of the compositional data methodology to constrained optimization in economics.** Jordi Saperas-Riera, Josep Antoni Martín-Fernández. On-line, Japan.

- XXXIX Congreso Nacional de Estadística e Investigación Operativa (SEIO 2022). **El análisis composicional del perfil de glucosa.** Jordi Saperas-Riera, Josep Antoni Martín-Fernández, Josep Vehí-Casellas. Granada, Spain.

- 9th International Workshop on Compositional Data Analysis (CODAWORK 2022). **Some thoughts on constrained convex optimization in the simplex.** Jordi Saperas-Riera, Josep Antoni Martín-Fernández. Toulouse, France.

- XL Congreso Nacional de Estadística e Investigación Operativa (SEIO 2023). **Balance selection: new insights on penalised regression models with compositional predictors.** Jordi Saperas-Riera, Glòria Mateu-Figueras, Josep Antoni Martín-Fernández. Elx, Spain.

- 10th International Workshop on Compositional Data Analysis (CODAWORK 2024). **Exploring Penalty Term Metrics in LASSO Regression for Compositional Data: A Comprehensive Data-Driven Analysis.** Jordi Saperas-Riera, Josep Antoni Martín-Fernández. Girona, Spain.

- 52nd meeting of the Italian Statistical Society (SIS 2024). **Balance selection for low dimension: the time-use case.** Jordi Saperas-Riera, Glòria Mateu-Figueras, Josep Antoni Martín-Fernández. Bari, Italy.

Llista d'abreviatures

plr	Pairwise logratio
alr	Additive logratio
clr	Centered logratio
CoDa	Compositional Data
LASSO	Least Absolute Shrinkage and Selection Operator
Med	mediana
olr	Orthonormal logratio
SBP	Sequential Binary Partition

"Tornar sempre és la millor part de l'aventura..."
Els Amics de les Arts

A tots els que hi han estat.
*A la Marta. A en Biel. ***.*

Agraïments

Vull expressar el meu agraïment més profund als meus directors de tesi, en Martin i la Glòria, pel seu coneixement excepcional, el seu suport constant i l'excel·lent orientació que m'han ofert al llarg de tot aquest procés. Vull destacar l'acurat equilibri amb el qual m'han acompanyat durant aquests anys: m'han ofert orientació sense ser autoritaris, m'han donat consells i indicacions permetent-me alhora que resolgués els reptes per mi mateix. La seva influència ha estat decisiva no només en aquesta tesi, sinó també en el meu desenvolupament com a investigador.

També vull expressar agraïment a tot el Grup de Recerca en Estadística i Anàlisi de Dades Composicionals, així com als companys de departament, per la seva col·laboració i suport. Aquest agraïment no és només una formalitat. La recerca és cada cop més col·laborativa i interdisciplinària. La complexitat dels problemes actuals sovint requereix un esforç col·lectiu, combinant recursos, coneixements i habilitats per aconseguir resultats complets i impactants. Les contribucions individuals continuen sent valuoses, però són més efectives quan s'integren dins d'un marc més ampli com és la recerca col·laborativa. Avui dia, m'és difícil entendre la recerca des de la individualitat i l'acompanyament del grup esdevé un element clau en la construcció de nou coneixement.

Aquesta tesi ha estat finançada mitjançant un contracte predoctoral (FPI) del Ministerio de Ciéncia e Innovación (Ref: PRE2019-090976). També a través dels projectes de recerca: Anàlisi de Dades Composicionals i Espacials. Compositional and Spatial Data Analysis (COSDA) de l' Agència de Gestió d'Ajuts Universitaris i de Recerca (Ref: 2021SGR01197), MÉTodos del análisis COmposicional de DAtos del Ministerio de Ciencia e Innovación (Ref: RTI2018-095518-B-C21) i Generation and Transfer of Compositional Data Analysis Knowledge del Ministerio de Ciencia e Innovación (Ref: PID2021-123833OB-I00).

Índex

Índex de figures

Resum

Les dades composicionals fan referència a dades multivariants en què les variables representen parts d'un tot. Aquest tipus de dades són habituals en àmbits com ara la geologia, la biologia molecular, l'economia i la química, per citar-ne només alguns. Normalment, estan limitades a sumar una constant, com ara 1 o 100 %, tot i que, de manera més general, també poden aparèixer sense aquesta restricció. Un aspecte clau a tenir en compte en l'anàlisi de dades composicionals és que la informació rellevant no es troba en els valors absoluts de les parts, sinó en les relacions relatives entre elles. Els mètodes estadístics tradicionals, que pressuposen variables sense restriccions i que consideren el valor absolut de les parts, poden conduïr a resultats enganyosos o incoherents. Això inclou, especialment, correlacions espúries. Aquestes es produeixen quan les relacions entre les parts semblen significatives però, en realitat, són el resultat de les restriccions inherents a la naturalesa composicional de les dades, i no de cap relació real entre les variables. Per evitar aquests problemes, és essencial utilitzar mètodes que tinguin en compte la naturalesa relativa de les dades, és a dir, que utilitzin la geometria específica de les dades composicionals, la coneguda com geometria d'Aitchison.

La geometria d'Aitchison proporciona un marc rigorós i coherent per a l'anàlisi composicional de les dades. Aquesta geometria introdueix operacions específiques com ara la pertorbació (equivalent a la suma), la potenciació (equivalent al producte per escalars) i un producte escalar, dissenyades per respectar l'estructura particular de l'espai mostral, el símplex. Un concepte central en aquest camp és l'isomorfisme logarítmic. Aquesta transformació permet l'aplicació de logquocients i balanços, fent possible treballar amb dades logtransformades en l'espai euclidià, on es poden aplicar les tècniques estadístiques convencionals de manera coherent. Aquest enfocament facilita la interpretació correcta de les distàncies, angles i altres propietats geomètriques, la qual cosa fa que l'anàlisi composicional de les dades sigui més robust i fiable.

Aquesta tesi contribueix al desenvolupament dels fonaments matemàtics de l'anàlisi composicional de dades. En particular, adapta les definicions de convexitat i de normes L^p al símplex. L'optimització convexa té un paper crucial en nombroses tècniques estadístiques, especialment en la resolució de problemes de minimització per trobar solucions òptimes. En el context de les dades composicionals, és essencial redefinir els conjunts i les funcions convexes dins del símplex per tal de respectar la geometria d'Aitchison. Aquesta tesi aborda aquesta necessitat adaptant l'optimització convexa a l'anàlisi composicional de dades. Es presenten definicions rigoroses de conjunts i funcions convexes en el símplex i s'ofereixen exemples que permeten una aplicació coherent a conjunts de dades reals. Exemples d'optimització convexa, com ara la regressió penalitzada, l'anàlisi de components principals i molts altres, inclouen mètriques. Per aquest motiu s'han adaptat les normes L^p al símplex i s'han explorat les seves propietats principals en el context composicional.

Finalment, aquesta tesi aplica aquests avenços en els fonaments matemàtics a la metodologia LASSO. La tècnica de regularització *Least Absolute Shrinkage and Selection Operator* (LASSO) és àmpliament reconeguda per la seva eficàcia a l'hora d'ajustar models lineals mentre realitza la selecció de variables. No obstant això, l'aplicació de LASSO a les dades composicionals presenta nous reptes, ja que el terme de penalització ha de respectar la geomètrica d'Aitchison. En resposta a aquest repte, s'ha proposat un enfocament que defineix una nova norma en l'espai composicional anomenada L^1-*plr*, coherent amb l'estructura del símplex. El model LASSO resultant redueix eficaçment la dimensionalitat, seleccionant logquocients significatius entre parts, cosa que representa un avenç important en la seva aplicació a les dades composicionals. A més, s'ha realitzat una comparació entre els models de regressió LASSO obtinguts amb diferents normes en el terme de penalització, analitzant com el procés de regularització afecta l'estructura subcomposicional del model lineal.

En resum, les principals contribucions d'aquesta tesi doctoral en el camp de l'anàlisi composicional de dades són: establir un marc coherent per a l'optimització convexa dins de la geometria d'Aitchison, i desenvolupar normes composicionals consistents per a la regressió LASSO en aquest marc. A més, la introducció de la norma L^1-*plr* facilita la selecció de balanços en el model de regressió lineal amb covariables composicionals. En última instància, aquests avenços formals amplien el conjunt d'eines metodològiques disponibles per als investigadors que treballen amb dades composicionals en una àmplia gamma de disciplines científiques.

Resumen

Las datos composicionales hacen referencia a datos multivariantes en los que las variables representan partes de un todo. Este tipo de datos son comunes en campos como la geología, la bi, la ecoología molecular, la economía y la química, por citar solo algunos. Normalmente, están limitados a sumar una constante, como 1 o 100 %, aunque, de manera más general, también pueden aparecer sin esta restricción. Un aspecto clave a tener en cuenta en el análisis composicional de datos es que la información relevante no se encuentra en los valores absolutos de las partes, sino en las relaciones relativas entre ellas. Los métodos estadísticos tradicionales, que suponen variables sin restricciones y que consideran el valor absoluto de las partes, pueden conducir a resultados engañosos o incoherentes. Esto incluye, especialmente, correlaciones espurias. Estas se producen cuando las relaciones entre las partes parecen significativas, pero en realidad son el resultado de las restricciones inherentes a la naturaleza composicional de los datos, y no de ninguna relación real entre las variables. Para evitar estos problemas, es esencial utilizar métodos que tengan en cuenta la geometría específica de los datos composicionales, la conocida como geometría de Aitchison.

La geometría de Aitchison proporciona un marco riguroso y coherente para el análisis composicional de los datos. Esta geometría introduce operaciones específicas como la perturbación (equivalente a la suma), la potenciación (equivalente al producto por escalares) y un producto escalar, que están diseñadas para respetar la estructura particular del espacio muestral, el símplex. Un concepto central en este campo es el isomorfismo logarítmico. Esta transformación permite la aplicación de logcocientes y balances, haciendo posible trabajar con datos logtransformados en el espacio euclidiano, donde se pueden aplicar técnicas estadísticas convencionales de manera coherente. Este enfoque facilita la interpretación correcta de las distancias, ángulos y otras propiedades geométricas, lo que hace que el análisis composicional de los datos sea más robusto y fiable.

Esta tesis contribuye al desarrollo de los fundamentos matemáticos del

análisis composicional de datos. En particular, adapta las definiciones de convexidad y de norma L^p al símplex. La optimización convexa desempeña un papel crucial en numerosas técnicas estadísticas, especialmente en la resolución de problemas de minimización para encontrar soluciones óptimas. En el contexto de los datos composicionales, es esencial redefinir los conjuntos y las funciones convexas dentro del símplex para respetar la geometría de Aitchison. Esta tesis aborda esta necesidad adaptando la optimización convexa específicamente para el análisis composicional de datos. Se presentan definiciones rigurosas de conjuntos y funciones convexas en el símplex y se ofrecen ejemplos que permiten una aplicación coherente a conjuntos de datos reales. Algunos ejemplos de optimización convexa, como la regresión penalizada, el análisis de componentes principales y muchos otros, incluyen métricas. Por este motivo, se han adaptado las normas L^p al símplex y se han explorado las principales propiedades en el contexto composicional.

Finalmente, esta tesis aplica estos avances en los fundamentos matemáticos a la metodología LASSO. La técnica de regularización *Least Absolute Shrinkage and Selection Operator* (LASSO) es ampliamente reconocida por su eficacia a la hora de ajustar modelos lineales mientras realiza la selección de variables. Sin embargo, la aplicación de LASSO a los datos composicionales presenta nuevos retos, ya que el término de penalización debe respetar la geométrica de Aitchison. En respuesta a este desafío, se ha propuesto un enfoque que define una nueva norma en el espacio composicional llamada L^1-*plr*, coherente con la estructura del símplex. El modelo LASSO resultante permite reducir efectivamente la dimensionalidad, seleccionando logcocientes entre partes, lo que representa un avance importante en su aplicación a los datos composicionales. Además, se ha realizado una comparación entre los modelos de regresión LASSO obtenidos con diferentes normas en el término de penalización, analizando cómo el proceso de regularización afecta a la estructura subcomposicional del modelo lineal.

En resumen, las principales contribuciones de esta tesis doctoral en el campo del análisis composicional de datos son: establecer un marco coherente para la optimización convexa dentro de la geometría de Aitchison y desarrollar normas composicionales consistentes para la regresión LASSO en este marco. Además, la introducción de la norma L^1-*plr* facilita la selección de balances en el modelo de regresión lineal con covariables composicionales. En última instancia, estos avances formales amplían el conjunto de herramientas metodológicas disponibles para los investigadores que trabajan con datos composicionales en una amplia gama de disciplinas científicas.

Abstract

Compositional data refers to multivariate data where the variables refer to parts of a whole. These data are common in fields such as geology, molecular biology, economics, and chemistry, to name just a few. Typically, they are constrained to sum up to a constant, such as 1 or 100 %; although more generally, they can be found as relative data without such constant-sum constraint. A key aspect to consider for their statistical analysis is that the relevant information lies not in the absolute values of the parts but in the relative relationships between them. Using ordinary statistical methods, which assume unconstrained variables, may lead to misleading or inconsistent results. This notably includes spurious correlations, which occur when the relationships between parts appear to be significant but are, in reality, an artefact of the inherent compositional nature of the data. To address these challenges, it is essential to use methods that account for the specific geometry of compositional data, the so-called Aitchison's geometry.

Aitchison's geometry provides a rigorous and coherent framework for the compositional analysis of data. This geometry introduces specific operations such as perturbation (equivalent to addition), powering (equivalent to ordinary scalar product), and an inner product, which are designed to respect the particular structure of a simplex as a sample space for compositional data. A pivotal concept in this field is the logarithmic isomorphism. This transformation allows the application of logratios and balances, making it possible to work with logtransformed data in the Euclidean real space where conventional statistical techniques can be coherently applied. This approach facilitates the correct interpretation of distances, angles, and other geometric properties, making the compositional approach more robust and reliable.

This doctoral thesis contributes to the development of the mathematical foundations of compositional data analysis. In particular, it adapts the definitions of convexity and L^p norms to the simplex. Convex optimization plays a crucial role in numerous statistical techniques, especially in solving minimization problems to find optimal solutions. In the context of compositional

data, it is essential to redefine convex sets and functions within the simplex to fulfil the structure of Aitchison's geometry. The present work addresses this by adapting convex optimization to the compositional case. It presents rigorous definitions of convex sets and functions in the simplex, providing examples that allow for a coherent application to real-world compositional data sets. Examples of convex optimization, such as penalized regression, principal component analysis, and others, contain metrics. Thus, L^p norms are redefined for the simplex, and their main properties are explored in the compositional context.

Finally, this thesis applies these advancements in mathematical foundations to the LASSO regression methodology. The Least Absolute Shrinkage and Selection Operator (LASSO) regularisation method is widely recognized for its effectiveness in fitting linear models while performing variable selection. However, applying the LASSO to compositional data entails new challenges, since the penalty term involved must respect Aitchison's geometry. In response to this, an approach is proposed defining a novel compositional norm named L^1-*plr*, which is consistent with the structure of the simplex. The resulting LASSO model effectively reduces dimensions by selecting meaningful logratios between parts, representing a significant advancement in its application to compositional data. Furthermore, a comparison is made between LASSO regression models obtained by using different norms in the penalty term, specifically investigating how the regularization process affects the subcompositional structure of the fitted linear model.

In summary, the main contributions of the current doctoral thesis to the field of compositional data analysis are: establishing a coherent framework for convex optimization within Aitchison's geometry and developing consistent compositional norms for LASSO regression within such framework. Moreover, the introduction of the L^1-*plr* norm facilitates the selection of logratio balances in regression modelling with compositional covariates. Ultimately, these formal advances expand the methodological toolkit available for researchers working with compositional data across a varied range of scientific disciplines.

Capítol 1

Introducció

1.1 Motivació

El treball de recerca desenvolupat en aquesta tesi doctoral neix de la necessitat d'adaptar a les dades composicionals (CoDa) els conceptes essencials per al plantejament i resolució de problemes d'optimització convexa, un àmbit crucial en les tècniques estadístiques. És àmpliament conegut que l'espai mostral de les dades composicionals, conegut com a símplex, presenta una estructura algebraicogeomètrica particular que difereix de l'estructura estàndard de l'espai euclidià real. Aquesta singularitat implica la necessitat d'adaptar els conceptes i les tècniques estadístiques tradicionals per tal de poder analitzar correctament les dades composicionals. Sovint, en la resolució d'un problema d'optimització es combinen tècniques i metodologies tradicionals amb desenvolupaments específics per a les dades composicionals, una barreja que pot conduir a resultats erronis o incoherents. Per tal d'abordar de manera precisa i eficient aquests reptes, s'ha desenvolupat una línia de recerca que ha cobert des dels fonaments teòrics fins a la creació d'algoritmes estadístics avançats. Aquesta línia de recerca s'ha vist reflectida en el projecte coordinat CoDaMET: "MÉTodos del análisis COmposicional de DAtos" (Ref: RTI2018-095518-B-C21; Ministerio de Ciencia, Innovación y Universidades), més concretament dins el subprojecte CODA-OPT, que té com un dels objectius principals el desenvolupament dels "Fonaments d'optimització restringida en dades composicionals". Aquesta línia de recerca ha tingut continuïtat en el projecte CoDaGenera: "Generation and Transfer of Compositional Data Analysis Knowledge" (Ref: PID2021-123833OB-I00; Ministerio de Ciencia, Innovación y Universidades), dins l'objectiu "Optimització en dades composicionals".

1.2 Situació dins la recerca

L'optimització convexa és una branca fonamental de les matemàtiques aplicades que se centra en la resolució de problemes en els quals es vol minimitzar o maximitzar una funció convexa subjecta a un conjunt de restriccions que defineixen un conjunt convex. En el camp de l'estadística, l'optimització convexa és essencial per diverses raons. Molts mètodes estadístics, com ara la regressió lineal, l'anàlisi discriminant, l'anàlisi de components principals o els models de màxima versemblança, es poden formular com a problemes d'optimització convexa. La convexitat garanteix que qualsevol solució local sigui també una solució global, i això facilita la resolució eficient d'aquests problemes. Això permet als estadístics desenvolupar mètodes més robustos i eficients per a l'anàlisi de dades i assegurar resultats fiables i interpretables.

Un exemple destacat d'optimització convexa en estadística és el problema de tipus LASSO (Least Absolute Shrinkage and Selection Operator). Aquest es pot formular com un problema d'optimització restringida, en què l'objectiu és minimitzar l'error de predicció subjecte a una restricció. La restricció o penalització es formula en termes de la norma L^1 dels coeficients i imposa esparsitat en els coeficients del model. Això significa que alguns coeficients es redueixen a zero i s'eliminen així les variables corresponents del model. Aquesta propietat permet una selecció automàtica de variables, fa el model més simple i interpretatiu i a més prevé el sobreajustament.

Aquesta tesi s'ha centrat en l'adaptació dels fonaments de convexitat, les mètriques L^p i la metodologia LASSO a l'espai composicional. El resultat final és un marc teòric precís i coherent, acompanyat d'un enriquiment significatiu de les eines metodològiques per a la regularització de models lineals, millorant així la capacitat predictiva i interpretativa dels models estadístics. Més concretament:

- A la bibliografia és fàcil trobar problemes d'optimització, no necessàriament composicionals, enquè el domini és un subconjunt del símplex. Podem trobar exemples en problemes d'ús del temps (Xie *et al.*, 2022) i en disseny d'experiments (Fang i Chan, 2006) entre d'altres. Tot i l'ús notable que es fa de l'optimització en el símplex no hi ha cap treball que descrigui els fonaments de convexitat en el símplex amb la seva particular geometria, anomenada geometria d'Aitchison. Per aquest motiu, amb la intenció d'omplir aquest buit en la bibliografia, aquesta tesi aborda els fonaments teòrics de la convexitat en l'espai composicional, proporcionant una base sòlida per a la correcta classificació dels problemes d'optimització convexa i el desenvolupament

d'eines metodològiques eficients i coherents.

- La mètrica d'Aitchison és una eina fonamental per a l'anàlisi composicional de dades (Aitchison, 1986), actuant com l'equivalent de la mètrica euclidiana en aquest espai particular. A la bibliografia s'han introduït altres mètriques al símplex, com la restricció de la mètrica L^1 aplicada al subespai de les parts logtransformades i centrades (clr). Aquesta mètrica ha estat àmpliament utilitzada en la selecció de variables (Lin *et al.*, 2014). Un altre exemple és la definició de la mètrica L^1 sobre una sistema fixat de coordenades ortonormal (Wang *et al.*, 2021). En tots aquests treballs, l'ús de la mètrica és purament instrumental. Malgrat la importància de les mètriques en el desenvolupament d'eines i metodologies estadístiques, no s'ha presentat fins ara una anàlisi exhaustiva de les normes L^p en l'espai composicional, ni tampoc una comparativa dels efectes que té l'ús d'una mètrica o una altra sobre les eines estadístiques. Aquesta tesi també ha volgut abordar aquesta llacuna, definint de manera rigorosa les normes L^p. A més, hem examinat els efectes que té el canvi de mètrica en el terme de penalització en la regularització LASSO, per proporcionar una comprensió més profunda de com aquestes mètriques impacten en els models estadístics.

- LASSO és una tècnica de regressió que combina la selecció de variables i la regularització per millorar tant la interpretabilitat com la capacitat predictiva del model. Aquesta característica és especialment útil en contextos on hi ha un gran nombre de variables predictores, perquè facilita la identificació de les variables més rellevants. Quan s'aplica LASSO a dades composicionals, es fa necessari adaptar la formulació per respectar les propietats logquocient d'aquestes dades (Lin *et al.*, 2014; Bates i Tibshirani, 2018). El seu impacte s'ha vist potenciat pel desenvolupament d'algoritmes específics en l'anàlisi estadística de la microbiota, en què l'alta dimensionalitat justifica l'ús de tècniques que simplifiquin els models lineals (Calle *et al.*, 2023). No obstant això, la selecció de variables no és l'única simplificació rellevant en el camp de les dades composicionals. La naturalesa relativa d'aquestes dades fa que també es pugui considerar simplificat un model lineal que es pugui expressar en funció d'uns pocs balanços, és a dir, logquocients entre grups de parts. Aquesta tesi aprofundeix en la regularització de balanços a través de la regressió penalitzada LASSO, oferint una alternativa a la selecció de variables que permeti a l'investigador descobrir nous patrons i millorar la interpretació del model lineal.

Aquesta tesi presenta tres articles que aporten avenços significatius en els fonaments matemàtics de l'anàlisi composicional de dades: convexitat i normes L^p. D'aquesta manera, contribueix a formalitzar un marc teòric més robust. Aquests avenços teòrics permeten la introducció d'eines metodològiques per a la regularització i simplificació de models en l'anàlisi composicional. Els tres treballs proporcionen una base sòlida per a futurs estudis i aplicacions destacant la importància de les mètriques adequades i l'adaptació de tècniques convencionals a contextos composicionals.

1.3 Presentació dels articles

La recerca desenvolupada en el marc d'aquesta tesi doctoral s'ha concretat en aquests articles:

- El primer, titulat "Fundamentals of convex optimization for compositional data", ha estat publicat a la revista *Statistical & Operational Research Transactions* (SORT). Se'n dona una transcripció de l'article a partir de la pàgina 51. En aquest article abordem el buit en la bibliografia especialitzada introduint i definint de manera rigorosa nous conceptes d'optimització convexa per a dades composicionals segons la geometria d'Aitchison. Entre aquestes noves definicions s'inclouen els conceptes de conjunts convexos i funcions convexes dins del símplex. L'article conclou amb un exemple d'optimització convexa resolt utilitzant tant la geometria euclidiana com la geometria d'Aitchison, i s'il·lustra com la primera pot portar a resultats incongruents.

- El segon article, titulat "Lasso regression method for a compositional covariate regularised by the norm L^1 pairwise logratio", ha estat publicat a la revista *Journal of Geochemical Exploration*. Se'n dona una transcripció a partir de la pàgina 75. En aquest article es presenta, per primer cop, un procés de regularització LASSO que fa selecció de balanços mantenint totes les parts en el resultat final. El punt clau d'aquest procés de regularització és la definició d'una nova norma, anomenada L^1 *pairwise logratio* , L^1-*plr*, que produeix esparsitat en els balanços de tipus *pairwise*. Aquest esquema de regressió LASSO generalitzada s'il·lustra amb l'anàlisi d'un conjunt de dades geoquímiques.

- El tercer article, titulat "L^p-Norm for Compositional Data: Exploring the CoDa L^1-Norm in Penalised Regression", ha estat publicat a la

revista *Mathematics*. Se'n dona una transcripció a partir de la pàgina 87. Aquest article presenta de manera rigorosa la família de normes L^p en l'espai composicional, parant especial atenció a les normes L^1-CoDa, L^2-CoDa i L^∞-CoDa. Seguidament, es condueix el procés de regularització LASSO amb tres normes diferents: L^1-*plr*, L^1-CoDa i L^1-clr. Les regularitzacions s'apliquen a un conjunt de dades de microbioma, que permete fer una comparació detallada entre aquests processos per avaluar-ne l'eficàcia i els impactes respectius sobre els models lineals.

1.4 Estructura de la tesi

Els capítols de la tesi s'estructuren de la següent manera. En el Capítol 2 es descriuen els objectius generals i específics de la tesi. En el Capítol 3 es presenten de manera breu i sintètica els aspectes metodològics bàsics de l'anàlisi composicional relacionats amb la recerca realitzada. També s'inclou una revisió de les normes L^p en espais quocients euclidians i del procés de regularització LASSO estàndard. En el Capítol 4, que constitueix el nucli central de la tesi, s'adjunta una còpia dels articles publicats. En el Capítol 5 es discuteixen els principals resultats obtinguts i es presenten les conclusions, així com les futures línies de recerca.

Capítol 2

Objectius

En el moment en què es va iniciar aquesta tesi, no hi havia cap estudi rigorós en la bibliografia sobre optimització convexa aplicada a l'espai mostral de les dades composicionals amb la geometria d'Aitchison. Tot i que hi havia treballs que involucraven elements d'optimització convexa en el seu procés, aquests sovint no complien de manera adequada les propietats de la geometria d'Aitchison, i això posa en qüestió la seva validesa en aquest context específic (Wang *et al.*, 2021). La tesi que es presenta sota el títol *Avenços en els fonaments matemàtics de l'anàlisi composicional de dades: convexitat i normes L^p. Aplicació a la regressió lineal LASSO amb covariable composicional* té com a objectiu principal omplir aquest buit i definir amb rigor els elements essencials de l'optimització convexa en el marc de les dades composicionals. El propòsit és establir un marc teòric i pràctic per a l'aplicació de tècniques d'optimització convexa que respectin la naturalesa intrínseca de les dades composicionals, per garantir així la coherència i la validesa dels resultats obtinguts. D'altra banda, dins de l'optimització convexa, les funcions que inclouen elements mètrics tenen un paper crucial i són fonamentals en el moment de desenvolupar metodologies estadístiques com ara els models de regressió lineal o l'anàlisi de clústers. Aquesta tesi també inclou un treball teòric original que adapta a l'espai mostral composicional les tradicionals mètriques L^p i defineix una nova norma, la norma L^1-*plr*. Aquest treball teòric s'ha aplicat per ampliar i desenvolupar amb major profunditat la regressió lineal penalitzada amb covariable composicional, i obrir noves possibilitats per a l'anàlisi i la interpretació de dades composicionals en el context dels models lineals. En resum, aquesta tesi ofereix contribucions en dues àrees clarament diferenciades: d'una banda, un treball de fonament matemàtic rigorós en l'espai mostral composicional, acompanyat d'exemples

il·lustratius, que proporciona una base sòlida per a futurs desenvolupaments en aquest camp i de l'altra, un desenvolupament metodològic en el camp de la regressió lineal penalitzada que amplia les eines disponibles per a l'anàlisi composicional de dades en situacions pràctiques.

2.1 Objectius de la tesi

Els objectius d'aquesta tesi es detallen a continuació:

O1. Adaptar la teoria sobre convexitat a l'espai mostral composicional.

O2. Adaptar la definició de norma L^p a l'espai mostral composicional. Definir una nova norma: L^1-plr.

O3. Definir el marc teòric on dur a terme la reducció de dimensió en models lineals mitjançant regressió penalitzada (LASSO).

O4. Explorar els algoritmes de resolució numèrica existents en llenguatge R per tal de resoldre els problemes d'optimització convexa de manera eficient.

Aquests objectius s'han desenvolupat en tres publicacions diferents que han estat revisades per revisors externs. En aquestes publicacions es mostren els conceptes i la metodologia proposats a través d'exemples amb aplicacions pràctiques de casos reals. A continuació es resumeix la relació entre cadascuna de les publicacions i els objectius descrits:

	Objectiu			
Article	**O1**	**O2**	**O3**	**O4**
Fundamentals of convex optimization for compositional data (SORT, IF=0.7 (Q3))	✓			
Lasso regression method for a compositional covariate regularized by the norm L1 pairwise logratio (J. Geochem. Explor., IF=3.4 (Q1))		✓	✓	✓
Lp-norm for compositional data: exploring the CoDa L1-norm in penalised regression (Mathematics, IF=2.3 (Q1))		✓	✓	✓

Capítol 3

Metodologia

En aquest capítol es pretén oferir una introducció als conceptes i definicions bàsiques necessàries per entendre el treball desenvolupat en aquesta tesi. En primer lloc, es detallen els fonaments de l'anàlisi composicional de dades i l'estructura del seu espai mostral. Seguidament s'exposen els conceptes de conjunt convex i funció convexa adaptats a l'espai mostral composicional. A continuació, es presenten les mètriques L^p en espais quocients euclidians. Finalment, es proporciona una introducció a la metodologia de la regressió LASSO estàndard. Aquesta base teòrica proporciona el context necessari per comprendre les aportacions específiques i els resultats presentats en els capítols posteriors.

3.1 Anàlisi composicional de dades

Històricament, les CoDa s'han definit com dades que representen proporcions o parts d'un tot i són àmpliament utilitzades en camps com ara la geologia, la química o les ciències socials. Aquestes dades tenen una naturalesa intrínsecament relativa, ja que la informació significativa es troba en les relacions entre les parts, més que en els seus valors absoluts. Tradicionalment, les dades composicionals es representen en el símplex, lloc geomètric dels vectors de components positives amb suma constant (habitualment igual a 1). Una aproximació més moderna de l'anàlisi composicional de dades assumeix que les dades són classes d'equivalència d'un espai quocient, on dos vectors són equivalents si un és múltiple de l'altre. A aquestes dades les anomenem *dades composicionals* o *composicions*. L'espai quocient permet tractar les dades composicionals adequadament, ja que les operacions matemàtiques tradicionals no són directament aplicables. Per exemple,

no es poden sumar o restar dues composicions sense perdre la seva naturalesa relativa. La solució és utilitzar la funció logarítmica per establir un isomorfisme amb un espai quocient euclidià en l'espai real multidimensional on es poden aplicar les operacions habituals i les tècniques estadístiques tradicionals. Aquest enfocament assegura que les operacions realitzades siguin coherents amb la naturalesa relativa de les dades.

3.1.1 Conceptes bàsics

Els conceptes i la metodologia explorats en aquesta tesi estan estretament vinculats amb la noció que les composicions es poden conceptualitzar com a classes d'equivalència. Un article clau per aprofundir en la comprensió de l'espai mostral de les dades composicionals com a espai quocient és Barceló-Vidal i Martín-Fernández (2016). A continuació, procedirem a analitzar detalladament l'estructura d'espai quocient de les composicions.

Assumim que les nostres dades són realitzacions de composicions aleatòries que es materialitzen en vectors $\mathbf{x} = (x_1, \ldots, x_D)$ amb D components estrictament positives, conegudes com a parts. En una anàlisi composicional assumim que l'interès se centra en l'estudi de la informació relativa entre les parts, més que en la magnitud absoluta de cada part individualment. En aquest context, un vector i un múltiple d'aquest es consideren vectors equivalents, ja que contenen la mateixa informació relativa. En aquesta tesi, si no es diu el contrari, els vectors $\mathbf{x}$ són vectors columna i $\mathbf{x}'$ denota el vector transposat.

Definició 3.1. Dos vectors de D-parts, $\mathbf{x},\ \mathbf{y} \in \mathbb{R}^D_+$, són composicionalment equivalents, i ho denotarem $\mathbf{x} \sim \mathbf{y}$, si un és múltiple de l'altre, és a dir, si existeix un nombre real positiu k tal que $\mathbf{x} = k\mathbf{y}$.

La relació d'equivalència 3.1 parteix l'ortant positiu, $\mathbb{R}^D_+$, en classes d'equivalència que anomenarem *composicions* (Fig. 3.1 (a)). Podem visualitzar geomètricament les composicions com a rajos de l'ortant positiu que passen per l'origen (Aitchison, 1986). Per simplicitat, denotarem amb $\mathbf{x}$ tant el vector $\mathbf{x}$ com la seva classe d'equivalència.

Definició 3.2. Definim l'*espai composicional* com el conjunt de totes les composicions, és a dir, l'espai quocient $\mathbb{R}^D_+/\sim$. L'espai composicional es denota per $\mathcal{S}^D$.

De tots els representants de l'espai composicional que podem trobar, l'espai mostral natural de les composicions és el *símplex unitari* (Fig. 3.1

(a)). Utilitzarem indistintament $\mathcal{S}^D$ per denotar l'espai composicional i el símplex unitari:

$$\mathcal{S}^D = \{(x_1, x_2, \ldots, x_D) \in \mathbb{R}_+^D \mid x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = 1\}.$$

Donada una composició $\mathbf{x}$, per tal d'obtenir el representant de la classe d'equivalència situat en el símplex unitari s'ha de dividir cada part de $\mathbf{x}$ per la suma de totes elles. Aquesta operació s'anomena *clausura* i la denotem per C:

$$C(\mathbf{x}) = \left(\frac{x_1}{\sum_{i=1}^{D} x_i}, \frac{x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i} \right).$$

La clausura és un reescalat de les parts d'una composició per tal que la seva suma sigui 1. Cal fer notar que l'operació clausura proporciona un vector composicionalment equivalent, i per tant no modifica la informació relativa entre les parts de la composició.

3.1.2 L'espai composicional com a espai vectorial

El marc geomètric formal per a l'anàlisi composicional de dades, conegut com a *geometria d'Aitchison*, va ser introduït per primera vegada per Pawlowsky-Glahn i Egozcue (2001) i Billheimer *et al.* (2001). Aquesta geometria es fonamenta en dues operacions bàsiques definides en l'espai composicional: la pertorbació i la potència. La pertorbació es refereix a l'operació de combinar dues composicions per la multiplicació component a component. La potència, en canvi, implica elevar cada component d'una composició a una potència donada. Aquestes operacions permeten tractar les composicions de manera coherent amb la seva naturalesa relativa, i proporcionen una base matemàtica sòlida per a l'anàlisi estadística en aquest context.

Definició 3.3. Siguin $\mathbf{x}, \mathbf{y}$ dues composicions amb D parts. Es defineix l'operació pertorbació com:

$$\mathbf{x} \oplus \mathbf{y} = (x_1 y_1,\ x_2 y_2,\ \ldots,\ x_D y_D)\,.$$

Definició 3.4. Sigui $\mathbf{x}$ una composició amb D parts i sigui α un escalar de $\mathbb{R}$. Es defineix l'operació potència com:

$$\alpha \odot \mathbf{x} = (x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha).$$

Cal notar que les operacions pertorbació i potència, denotades per $\oplus$ i $\odot$ respectivament, són coherents amb la relació d'equivalència $\sim$, és a dir, el resultat d'aquestes operacions no depèn del representant escollit (vegeu la propietat 3.1). Les operacions de pertorbació i potència doten l'espai composicional amb estructura d'espai vectorial sobre el cos $\mathbb{R}$.

Propietat 3.1. Siguin $\mathbf{x}, \mathbf{y}$ dues composicions amb D parts. Sigui α un nombre real i λ un nombre real positiu. Les operacions pertorbació i potència estan ben definides, és a dir,

- $\lambda\mathbf{x} \oplus \mathbf{y} = \lambda\,(\mathbf{x} \oplus \mathbf{y})$
- $\mathbf{x} \oplus \lambda\mathbf{y} = \lambda\,(\mathbf{x} \oplus \mathbf{y})$
- $\alpha \odot (\lambda\mathbf{x}) = \lambda^{\alpha}\,(\alpha \odot \mathbf{x})$

Per facilitar l'aplicació efectiva d'eines estadístiques estàndard, és essencial capturar la informació relativa proporcionada pels quocients entre les components i integrar-la en l'espai euclidià convencional. L'ús de la funció logarítmica com a isomorfisme entre l'ortant positiu, $\mathbb{R}^D_+$, i l'espai real, $\mathbb{R}^D$, és la clau per a la comprensió i la interpretació precises de les composicions (Barceló-Vidal i Martín-Fernández, 2016).

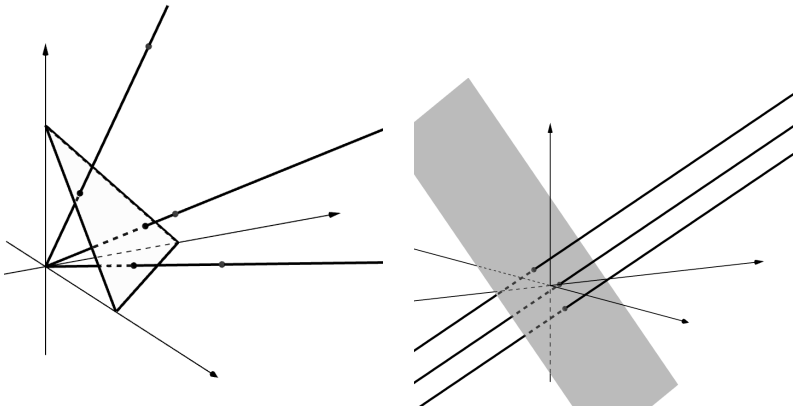
La funció logarítmica indueix un isomorfisme entre l'espai composicional, $\mathcal{S}^D = \mathbb{R}^D_+/\sim$, i l'espai quocient $\mathcal{L}^D = \mathbb{R}^D/\equiv$ (Fig. 3.1):

$$\begin{aligned} \ln : \mathcal{S}^D &\to \mathcal{L}^D \\ \mathbf{x} &\mapsto \mathbf{z} = \ln \mathbf{x}, \end{aligned} \tag{3.1}$$

on $\mathbf{z} \equiv \mathbf{z}^*$ si i només si $\mathbf{z}^* = \mathbf{z} + \lambda\mathbf{1}_D$, $\lambda \in \mathbb{R}$ i $\mathbf{1}_D = (1, \ldots, 1)$. Per simplicitat, denotarem amb $\mathbf{z}$ tant el vector $\mathbf{z}$ com la seva classe d'equivalència. De tots els representats de l'espai quocient $\mathcal{L}^D$, el més habitual és el subpespai vectorial $\{\mathbf{z} \in \mathbb{R}^D |\ \sum_{i=1}^D z_i = 0\}$ anomenat *clr* (de l'anglès *centered logratio*) (Fig. 3.1 (b)). Donat $\mathbf{z} \in \mathbb{R}^D$, i $\mathbf{1}_D = (1, \ldots, 1)$, el representant de la classe d'equivalència de $\mathbf{z}$ en el subespai clr és el vector centrat $\mathbf{z} - \overline{\mathbf{z}}\mathbf{1}_D = H_D\mathbf{z}$, on $\overline{\mathbf{z}} = \frac{\sum_{j=1}^D z_j}{D}$, $H_D = I_D - \frac{1}{D}\mathbf{1}_D\mathbf{1}'_D$ és la matriu de centrat i I_D és la matriu identitat d'ordre $D \times D$.

3.2 Convexitat en l'espai composicional

La convexitat té un paper crucial en moltes àrees de l'estadística, des de la teoria fins a les aplicacions pràctiques. Molts problemes d'estimació en

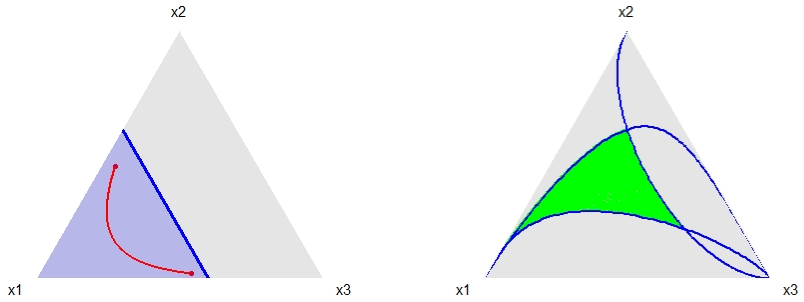


(a) (b)

Figura 3.1: (a) L'espai composicional de 3 parts, $\mathcal{S}^3$, on el triangle és el símplex unitari. Els rajos que passen per l'origen són composicions ($\mathbf{x}$). (b) L'espai quocient $\mathcal{L}^3$ on el pla gris és el subespai vectorial clr. Les rectes perpendiculars al subespai vectorial clr són les classes d'equivalència ($\mathbf{z} = \ln \mathbf{x}$)

estadística es poden formular com a problemes d'optimització convexa. Per exemple, en la regressió lineal l'objectiu és minimitzar la suma dels quadrats dels residus, que és una funció convexa de les estimacions dels paràmetres. Això assegura que qualsevol solució òptima local és també la solució òptima global, cosa que facilita el càlcul de les estimacions dels paràmetres. En l'anàlisi de regressió, els problemes de regressió com ara LASSO i Ridge són exemples de problemes convexos amb restricció: LASSO afegeix una restricció a l'espai de coeficients basada en la norma L^1, mentre que Ridge afegeix una restricció basada en la norma L^2. Ambdós problemes són problemes d'optimització convexa, la qual cosa permet utilitzar mètodes d'optimització eficients per trobar les solucions òptimes. En l'anàlisi multivariant, les funcions convexes es poden utilitzar per a la reducció de dimensionalitat, com en l'anàlisi de components principals (PCA). La convexitat de les funcions cost en PCA garanteix que les solucions obtingudes són òptimes i úniques, i això facilita la interpretació dels resultats i l'extracció de les components rellevants.

A Saperas-Riera *et al.* (2023) s'adapten de manera precisa les definicions de conjunt convex i funció convexa a la geometria d'Aitchison, il·lustrant-les amb exemples detallats. Un conjunt convex és un conjunt de punts en un espai vectorial tal que, per a qualsevol parell de punts dins del conjunt,



(a) Conjunt $\mathcal{A}$-convex (b) Triangle $\mathcal{A}$-convex

Figura 3.2: Conjunts $\mathcal{A}$-convexos en $\mathcal{S}^3$: (a) El conjunt $\Sigma^+ = \{\mathbf{x} \in \mathcal{S}^3 | -6 + 4\frac{x_2}{x_1} - 4\frac{x_3}{x_1} - 1 \leq 0\}$ (regió blava). La corba vermella representa un segment que uneix dues composicions del conjunt; (b) Un triangle (àrea verda) determinat per la intersecció de tres semiplans (línies blaves).

qualsevol combinació convexa d'aquests punts també pertany al conjunt. Per tal de ser coherents amb la geometria d'Aitchison, cal que utilitzem les operacions de pertorbació i potència. Denotarem amb el prefix $\mathcal{A}$ els conceptes compatibles amb la geometria d'Aitchison.

Definició 3.5. Un conjunt $\mathcal{B}$ de l'espai vectorial $\mathcal{S}^D$ s'anomena $\mathcal{A}$-convex si, per a qualsevol parell de composicions $\mathbf{x},\ \mathbf{y} \in \mathcal{B}$ i qualsevol $\lambda \in [0, 1]$, es compleix que:

$$(1 - \lambda) \odot \mathbf{x} \oplus \lambda \odot \mathbf{y} \in \mathcal{B}$$

Això significa que el segment que uneix qualsevol parell de composicions dins del conjunt es troba completament dins del conjunt (Figura 3.2a). Exemples comuns de conjunts $\mathcal{A}$-convexos inclouen segments de línia, plans, poliedres i discs amb la geometria d'Aitchison. El triangle de la Figura 3.2b és un conjunt $\mathcal{A}$-convex ja que és la intersecció de tres semiplans amb la geometria d'Aitchison.

Una funció convexa és una funció definida sobre un conjunt convex que té la propietat que la línia recta que uneix qualsevol parell de punts en el seu gràfic es troba per sobre o al mateix nivell que la gràfica de la funció. Abans, cal notar que una funció definida en un domini $\mathcal{B} \subset \mathcal{S}^D$ està ben definida si és invariant per a escalars, $f(\lambda\mathbf{x}) = f(\mathbf{x})$ (Barceló-Vidal *et al.*, 2011).

Definició 3.6. Sigui $\mathcal{B} \subset \mathcal{S}^D$ un conjunt $\mathcal{A}$-convex. La funció $f : \mathcal{B} \to \mathbb{R}$ és una funció $\mathcal{A}$-convexa si per tot parell de composicions $\mathbf{x},\ \mathbf{y} \in \mathcal{B}$ i $\lambda \in [0, 1]$:

$$f((1-\lambda) \odot \mathbf{x} \oplus \lambda \odot \mathbf{y}) \leq (1-\lambda) f(\mathbf{x}) + \lambda f(\mathbf{y})$$

Definició 3.7. Sigui $\mathcal{B} \subset \mathcal{S}^D$ un conjunt $\mathcal{A}$-convex. La funció $f : \mathcal{B} \to \mathbb{R}$ és una funció $\mathcal{A}$-còncava si la funció $-f$ és $\mathcal{A}$-convexa.

En l'anàlisi composicional de dades, la combinació lineal dels logaritmes de les parts amb coeficients que sumen zero, coneguda com a *logcontrast*, exerceix un paper anàleg al de l'aplicació lineal en la geometria euclidiana. Els logcontrasts, que seran importants a l'hora de definir els problemes de programació lineal en l'anàlisi composicional de dades (Saperas-Riera *et al.*, 2023), són un exemple de funció $\mathcal{A}$-convexa i $\mathcal{A}$-còncava.

A Saperas-Riera *et al.* (2023), la norma d'Aitchison es presenta com a exemple de funció $\mathcal{A}$-convexa. En la Figura 3.3, podem veure les corbes de nivell de la funció distància d'Aitchison a una composició donada (Def. 3.11). En general, qualsevol mètrica compatible amb la geometria d'Aitchison és una funció $\mathcal{A}$-convexa. En canvi, la mètrica euclidiana no és una funció $\mathcal{A}$-convexa. En la Figura 3.4, podem veure les corbes de nivell de la funció distància euclidiana a una composició donada, $d_{E,\mathbf{c}}(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}\|_E$. Des d'una perspectiva geomètrica, podem veure que la funció distància euclidiana no és $\mathcal{A}$-convexa comprovant que els subnivells de la funció (conjunt de punts on el valor de la funció és menor o igual a un cert valor determinat) no són conjunts $\mathcal{A}$-convexos (Saperas-Riera *et al.*, 2023). Això es fa evident en la Figura 3.4(b), on les corves de nivell de la funció distància en coordenades mostren clarament que els subnivells de la funció distància euclidiana no són convexos.

3.3 Espai euclidià i mètriques L^p

L'isomorfisme logarítmic és de gran utilitat ja que permet que l'estructura euclidiana real definida a $\mathcal{L}^D$ es transfereixi a $\mathcal{S}^D$ (Barceló-Vidal i Martín-Fernández, 2016). La idea és definir primer el producte escalar a $\mathcal{L}^D$ i després transferir aquest producte escalar a $\mathcal{S}^D$, per construir d'aquesta manera la geometria d'Aitchison.

Definició 3.8. Donades $\mathbf{z},\ \mathbf{z}^* \in \mathcal{L}^D$, es defineix el producte escalar, $\langle\ ,\ \rangle_{\mathcal{L}}$, com

$$\langle \mathbf{z}, \mathbf{z}^* \rangle_{\mathcal{L}} = \mathbf{z}' H_D \mathbf{z}^*.$$

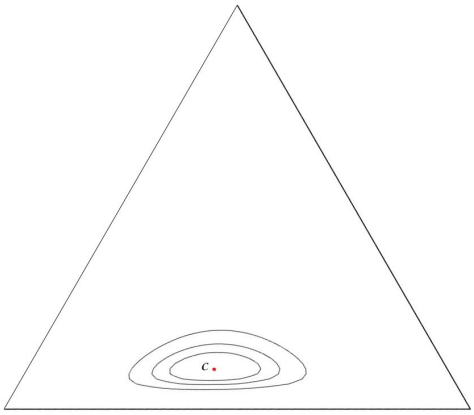
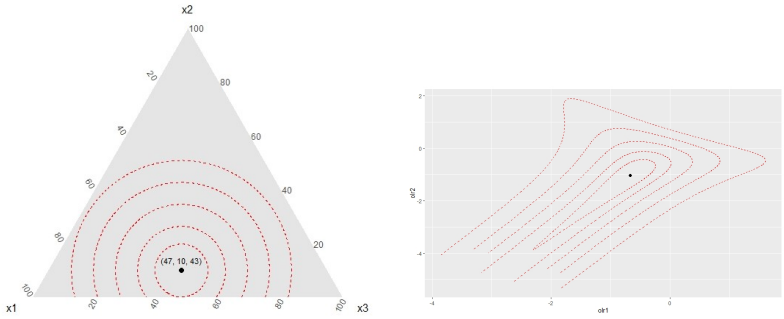


Figura 3.3: Corbes de nivell de la funció distància d'Aitchison a la composició $\mathbf{c} = (0.47,\ 0.1,\ 0.43) : d_{\mathcal{A},c}(\mathbf{x}) = \|\mathbf{x} \ominus \mathbf{c}\|_{\mathcal{A}}$



(a) (b)

Figura 3.4: Corbes de nivell de la funció distància euclidiana a una composició $\mathbf{c} = (0.47,\ 0.1,\ 0.43) : d_c(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}\|_E$. (a) Representació sobre el símplex de 3 parts. (b) Representació en coordenades olr. Els subnivells, conjunt de punts on el valor de la funció és menor o igual a un cert valor determinat, no són conjunts convexos.

És habitual escriure el producte escalar $\langle , \rangle_{\mathcal{L}}$ a partir dels representants del subespai clr de les classes d'equivalència:

$$\langle \mathbf{z}, \mathbf{z}^* \rangle_{\mathcal{L}} = \langle \mathbf{z} - \overline{\mathbf{z}}\mathbf{1}_D, \mathbf{z}^* - \overline{\mathbf{z}^*}\mathbf{1}_D \rangle_E = \sum_{j=1}^{D} (z_j - \overline{\mathbf{z}})(z^*{}_j - \overline{\mathbf{z}^*}),$$

on $\langle \ , \ \rangle_E$ és el producte escalar euclidià habitual.

Definició 3.9. Siguin $\mathbf{x},\ \mathbf{y} \in \mathcal{S}^D$ dues composicions amb D parts. Es defineix el producte escalar d'Aitchison com:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{A}} = \langle \ln \mathbf{x}, \ln \mathbf{y} \rangle_{\mathcal{L}} = \sum_{j=1}^{D} \ln \frac{x_j}{g(\mathbf{x})} \ln \frac{y_j}{g(\mathbf{y})},$$

on $\mathcal{A}$ denota la geometria d'Aitchison i $g(\mathbf{x})$ és la mitjana geomètrica de les parts de la composició $\mathbf{x}$.

A partir del producte escalar, $\langle \ , \ \rangle_{\mathcal{A}}$, es pot definir la norma i la distància d'Aitchison a l'espai composicional.

Definició 3.10. Sigui $\mathbf{x} \in \mathcal{S}^D$ una composició de D parts. Es defineix la norma d'Aitchison (Aitchison, 1986) com:

$$\|\mathbf{x}\|^2_{\mathcal{A}} = \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{A}} = \sum_{j=1}^{D} \left(\ln \frac{x_j}{g(\mathbf{x})} \right)^2.$$

Definició 3.11. Siguin $\mathbf{x},\ \mathbf{y} \in \mathcal{S}^D$ dues composicions de D parts. Es defineix la distància d'Aitchison (Aitchison, 1986) com:

$$d_{\mathcal{A}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{\mathcal{A}},$$

on $\ominus$ és la pertorbació diferència: $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$

En l'anàlisi composicional, sovint hi ha un interès específic en un subconjunt de parts que resulten especialment rellevants per a l'estudi. Això pot ocórrer, per exemple, en l'anàlisi de components microbians, en què només un grup reduït d'espècies pot ser el focus d'atenció, o en estudis geològics, en què només alguns minerals són determinants per comprendre les característiques d'un territori. En aquests casos és crucial que les conclusions obtingudes del subconjunt de dades siguin coherents amb les que es derivarien de l'anàlisi de la composició completa.

El principi de coherència subcomposicional estableix que qualsevol anàlisi realitzada sobre aquest subconjunt de parts, que anomenarem *subcomposició*, ha de ser consistent amb els resultats que s'obtindrien si s'analitzés la composició global. Això vol dir que les relacions i les proporcions entre les parts seleccionades han de mantenir-se coherents, tant si es considera tota la composició com si es focalitza només en un subconjunt. Si aquest principi no es compleix, es poden obtenir conclusions contradictòries o distorsionades, la qual cosa compromet la validesa dels resultats i pot portar a interpretacions incorrectes.

Quan les tècniques estadístiques impliquen mètriques, la coherència subcomposicional pot ser formalitzada en termes mètrics, fet que dona lloc al concepte de dominància subcomposicional.

Definició 3.12. Sigui $\mathbf{x} \in \mathcal{S}^D$ una composició de D parts. Considerem una subcomposició qualsevol de $\mathbf{x}$, $sub(\mathbf{x}) = (x_{i_1}, \ldots, x_{i_d})$, amb $d < D$, on $i_1, \ldots, i_d$ és un subconjunt de les parts de $\mathbf{x}$. Sigui $\|\cdot\|_*$ una mètrica definida en l'espai composicional. *La dominància composicional* requereix que la norma de la subcomposició sigui menor o igual a la norma de la composició completa, $\|sub(\mathbf{x})\|_* \leq \|\mathbf{x}\|_*$.

3.3.1 Coordenades en l'espai composicional

L'estructura d'espai euclidià permet representar el símplex a través d'un sistema de coordenades. De totes les bases amb les quals podem expressar les composicions, la bibliografia ha consolidat l'ús de les coordenades clr (centered logratio) , alr (additive logratio) i olr (orthonormal logratio) .

Coordenades clr

Seguint l'estructura presentada en el capítol 3 de Pawlowsky-Glahn i Buccianti (2011), el sistema generador habitual, $\mathbf{W}$, és el conjunt format per les composicions $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$, on cada $\mathbf{w}_i$ és de la forma $(1, \ldots, e, \ldots, 1)$, amb e ocupant la i-èsima posició. Observem que, amb la geometria d'Aitchison, l'espai composicional és un espai multiplicatiu, en què l'element neutre és l'1 i la unitat és e. Per tant, podem dir que el sistema $\mathbf{W}$ és el sistema generador canònic. Per tant, tota composició $\mathbf{x} = (x_1, \ldots, x_D)$ es pot expressar com a combinació lineal dels $\mathbf{w}_i$ utilitzant les operacions de pertorbació i potència:

$$\mathbf{x} = (\ln x_1 \odot \mathbf{w}_1) \oplus (\ln x_2 \odot \mathbf{w}_2) \oplus \ldots \oplus (\ln x_D \odot \mathbf{w}_D)$$

Aquesta representació no és única. Cada composició representa una classe d'equivalència i per tant, les coordenades respecte a $\mathbf{W}$ no estan unívocament determinades:

$$\mathbf{x} = (\ln \lambda x_1 \odot \mathbf{w}_1) \oplus (\ln \lambda x_2 \odot \mathbf{w}_2) \oplus \ldots \oplus (\ln \lambda x_D \odot \mathbf{w}_D)$$

De tots els representants, és habitual prendre $\lambda = \frac{1}{g(\mathbf{x})}$. A aquestes coordenades se les anomena *coordenades* clr:

$$\mathbf{x} = (\ln \frac{x_1}{g(\mathbf{x})} \odot \mathbf{w}_1) \oplus (\ln \frac{x_2}{g(\mathbf{x})} \odot \mathbf{w}_2) \oplus \ldots \oplus (\ln \frac{x_D}{g(\mathbf{x})} \odot \mathbf{w}_D)$$

Recordem que les coordenades clr compleixen que la seva suma és zero, $\sum_{j=1}^{D} \ln \frac{x_j}{g(\mathbf{x})} = 0$. Per comprendre millor el significat geomètric d'aquesta elecció, cal estudiar les coordenades clr en l'espai $\mathcal{L}^D$ (Fig. 3.1b). La projecció ortogonal de $\mathbf{z} = \ln \mathbf{x}$ en el subespai clr s'interpreta com el centratment de la variable $\mathbf{z}$, i queda definit per la matriu H_D.

L'aplicació que assigna a cada composició un vector del subespai clr és la *transformació logquocient centrada* (Aitchison, 1986).

Definició 3.13. Sigui $\mathbf{x}$ una composició amb D parts. Les coordenades clr es defineixen com:

$$\operatorname{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \ \ldots, \ln \frac{x_D}{g(\mathbf{x})} \right).$$

Coordenades alr

Per tal de tenir una base, n'hi ha prou de prendre $D-1$ vectors del sistema generador $\mathbf{W}$. Per exemple, podem prendre com a base el conjunt $\{\mathbf{w}_1, \ldots, \mathbf{w}_{D-1}\}$ (Pawlowsky-Glahn i Buccianti, 2011). Qualsevol composició $\mathbf{x}$ pot ser escrita com a combinació lineal de $\mathbf{w}_i$, $i = 1, \ldots, D-1$, de manera única:

$$\mathbf{x} = (\ln \frac{x_1}{x_D} \odot \mathbf{w}_1) \oplus (\ln \frac{x_2}{x_D} \odot \mathbf{w}_2) \oplus \ldots \oplus (\ln \frac{x_{D-1}}{x_D} \odot \mathbf{w}_{D-1})$$

Aitchison (1986) defineix les coordenades alr d'una composició $\mathbf{x}$.

Definició 3.14. Sigui $\mathbf{x}$ una composició amb D parts. Les coordenades alr es defineixen com:

$$\text{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right)$$

Cal observar que la base $\{\mathbf{w}_1, \ldots, \mathbf{w}_{D-1}\}$ no és ortonormal:

- $\|\mathbf{w}_i\|^2_{\mathcal{A}} = \frac{D-1}{D}$
- $\langle \mathbf{w}_i, \mathbf{w}_j \rangle_{\mathcal{A}} = -\frac{1}{D}$

Coordenades olr

Les bases ortonormals en l'espai composicional van ser introduïdes per Egozcue *et al.* (2003). Donada una base qualsevol, podem obtenir una base ortonormal aplicant el procediment d'ortonormalització de Gram-Schmidt.

Prenem l'espai vectorial composicional, $\mathcal{S}^D$, amb el producte escalar d'Aitchison, $\langle\ ,\ \rangle_{\mathcal{A}}$. Donat un conjunt de composicions linealment independents $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{D-1}\}$, el procés de Gram-Schmidt genera una base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ de la manera següent:

1. Inicialitzem el primer vector:

$$\mathbf{u}_1 = \mathbf{v}_1$$

2. Per a $k = 2, 3, \ldots, D-1$:

$$\mathbf{u}_k = \mathbf{v}_k \ominus \bigoplus_{j=1}^{k-1} \frac{\langle \mathbf{v}_k, \mathbf{u}_j \rangle_{\mathcal{A}}}{\langle \mathbf{u}_j, \mathbf{u}_j \rangle_{\mathcal{A}}} \odot \mathbf{u}_j$$

3. Normalitzem els vectors $\mathbf{u}_k$:

$$\mathbf{e}_k = \frac{1}{\|\mathbf{u}_k\|} \odot \mathbf{u}_k$$

Les coordenades de la composició $\mathbf{x}$ en la base ortonormal $\{\mathbf{e}_k\}$ s'anomenen coordenades olr (Martín-Fernández, 2019):

$$\mathbf{x} = \bigoplus_{j=1}^{D-1} \text{olr}(\mathbf{x})_j \odot \mathbf{e}_j$$

Entre totes les bases ortonormals que es poden construir, les bases obtingudes mitjançant la metodologia de partició binària seqüencial (SBP, Egozcue i Pawlowsky-Glahn (2003)) tenen una importància especial en la bibliografia. La seva principal virtut rau en la simplicitat de la seva interpretació:

1. Dividim la composició de D-parts, $(x_1, \ldots, x_D)$, en dues subcomposicions no buides. A continuació, dividim cadascuna de les subcomposicions en dues subcomposicions no buides, iterant aquest procés fins que cada grup contingui només un element. Aquest procés consisteix en un total de $D-1$ particions.

2. Per a la partició k-èsima, creem un vector assignant pesos positius d'igual magnitud a les parts d'un grup, $\{x_{n1_k}, \ldots, x_{nr_k}\}$, i pesos negatius d'igual magnitud a les parts de l'altre grup, $\{x_{d1_k}, \ldots, x_{ds_k}\}$. Les parts que no participen en aquesta partició reben un pes zero. Triem els pesos positius i negatius de manera que la seva suma sigui zero. Aquesta elecció dels pesos garanteix l'ortogonalitat de la base. Un cop tenim els $D-1$ vectors de la base, els normalitzem per assegurar-nos que tenen longitud unitària. Els $D-1$ vectors ortonormals resultants són:

$$\mathbf{e}_j = \bigoplus_{j=n1_k}^{nr_k} \sqrt{\frac{s_k}{r_k(r_k+s_k)}} \odot \mathbf{w}_j \oplus \bigoplus_{j=d1_k}^{ds_k} -\sqrt{\frac{r_k}{s_k(r_k+s_k)}} \odot \mathbf{w}_j,\ j = 1, \ldots, D-1$$

El fet d'assignar pesos d'igual magnitud a cada grup en cada partició és el que facilita la interpretació de les coordenades en aquesta base. La coordenada k-èsima d'una composició $\mathbf{x}$ en una base ortonormal $\{\mathbf{e}_i\}$ creada a partir d'una SBP té l'expressió següent:

$$\operatorname{olr}(\mathbf{x})_k = \sqrt{\frac{r_k \cdot s_k}{r_k + s_k}} \ln \frac{(x_{n1_k} \ldots x_{nr_k})^{\frac{1}{nr_k}}}{(x_{d1_k} \ldots x_{ds_k})^{\frac{1}{ds_k}}}$$

Egozcue i Pawlowsky-Glahn (2003) denominen balanços les coordenades $\operatorname{olr}(\mathbf{x})_k$, a causa de l'expressió dels logquocients que s'hi obtenen.

3.3.2 Mètriques L^p

La naturalitat amb què el subpespai clr fa de representant de l'espai quocient $\mathcal{L}^D$ ha condicionat profundament el desenvolupament de l'espai composicional com a espai mètric. Les dues mètriques més utilitzades en l'anàlisi

composicional es defineixen habitualment prenent com a referència el representant del subespai clr:

1. Norma d'Aitchison:

$$\|\mathbf{x}\|_{\mathcal{A}}^2 = \|\mathrm{clr}(\mathbf{x})\|_2^2 = \sum_{j=1}^{D} \left(\ln \frac{x_j}{g(\mathbf{x})} \right)^2$$

2. Norma L^1 restringida al subespai clr:

$$\|\mathbf{x}\|_{1-\mathrm{clr}} = \|\mathrm{clr}(\mathbf{x})\|_1 = \sum_{j=1}^{D} \left| \ln \frac{x_j}{g(\mathbf{x})} \right|$$

Aquestes definicions fixen el representant del subespai clr a l'hora de definir la norma d'una composició. És important destacar que, en general, la mètrica obtinguda en restringir una mètrica euclidiana L^p al subespai clr no respecta els principis composicionals de la geometria d'Aitchison. En concret, no es verifica la dominància subcomposicional. Aquesta característica és fonamental, ja que la seva absència pot generar resultats incoherents, especialment en metodologies com ara els algoritmes de clúster, quan definim distàncies entre les observacions.

Propietat 3.2. La restricció de la norma L^p euclidiana en el subespai clr, $\|\mathbf{x}\|_{p-\mathrm{clr}} = \|\mathrm{clr}(\mathbf{x})\|_p = \left(\sum_{j=1}^{D} \left| \ln \frac{x_j}{g(\mathbf{x})} \right|^p \right)^{\frac{1}{p}}$, no és subcomposicionalment dominant per a $p \in (1,\ 2) \cup (2,\ \infty)$.

Demostració. Sigui $\mathbf{x}$ una composició de D parts, amb $D \geq 4$, amb coordenades $\mathrm{clr}(\mathbf{x}) = (0, 20, -10, -10, 0, \ldots, 0)$. És fàcil trobar composicions $\mathbf{x}^*$ tals que:

- $\mathrm{clr}(\mathbf{x}^*) = \mathrm{clr}(\mathbf{x}) + t(1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, 0, \ldots, 0)$.
- $\|\mathbf{x}^*\|_{p-\mathrm{clr}}^p < \|sub(\mathbf{x}^*)\|_{p-\mathrm{clr}}^p$, on $sub(\mathbf{x}) = (x_2, \ldots, x_D)$.

Per exemple, per a $p = 1.5$, podem prendre $\mathbf{x}^*$ tal que $\mathrm{clr}(\mathbf{x}^*) = (-0.6, 20.2, -9.8, -9.8)$. Es verifica que $\|\mathbf{x}^*\|_{p-\mathrm{clr}}^p = 152.61 < 152.688 = \|sub(\mathbf{x}^*)\|_{p-\mathrm{clr}}^p$. Un altre exemple: per a $p = 3$, podem prendre $\mathbf{x}^*$ tal que $\mathrm{clr}(\mathbf{x}^*) = (6, 18, -12, -12, 0, \ldots, 0)$. Es verifica que $\|\mathbf{x}^*\|_{p-\mathrm{clr}}^p = 9504 < 10000 = \|sub(\mathbf{x}^*)\|_{p-\mathrm{clr}}^p$. □

Propietat 3.3. La restricció de la norma L^∞ euclidiana en el subespai clr, $\|\mathbf{x}\|_{\infty-\text{clr}} = \|\text{clr}(\mathbf{x})\|_\infty = \max_j \left\{\left|\ln \frac{x_j}{g(\mathbf{x})}\right|\right\}$, no és subcomposicionalment dominant.

Demostració. Com en la propietat anterior, sigui $\mathbf{x}$ una composició de D parts, amb $D \geq 4$, amb coordenades $\text{clr}(\mathbf{x}) = (0, 20, -10, -10, 0, \ldots, 0)$. És fàcil trobar composicions $\mathbf{x}^*$ tals que:

- $\text{clr}(\mathbf{x}^*) = \text{clr}(\mathbf{x}) + t(1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, 0, \ldots, 0)$.
- $\|\mathbf{x}^*\|_{\infty-\text{clr}} < \|sub(\mathbf{x}^*)\|_{\infty-\text{clr}}$.

Per exemple, podem prendre $\mathbf{x}^*$ tal que $\text{clr}(\mathbf{x}^*) = (15, 15, -15, -15, 0, \ldots, 0)$. Es verifica que $\|\mathbf{x}^*\|_{\infty-\text{clr}} = 15 < 20 = \|sub(\mathbf{x}^*)\|_{\infty-\text{clr}}$ □

Per descriure el conjunt de mètriques L^p en l'espai composicional, cal abordar aquest problema sense comprometre'ns amb cap representant específic, induint l'habitual mètrica L^p directament sobre l'espai composicional. El procediment és similar al que hem utilitzat per definir el producte escalar en l'espai composicional en la secció 3.3:

- Definim la mètrica L^p en l'espai quocient $\mathcal{L}^D$ (Brezis, 2011):
  - Prenem la mètrica L^p en $\mathbb{R}^D$, $\|\mathbf{z}\|_p,\ \mathbf{z} \in \mathbb{R}^D$.
  - Induïm la mètrica L^p en $\mathcal{L}^D$ tot assignant a cada classe d'equivalència el mínim de la mètrica anterior,

$$\|\mathbf{z}\|_{p,\mathcal{L}^D} = \min_{\lambda \in \mathbf{R}} \|\mathbf{z} + \lambda \mathbf{1}_D\|_p.$$

- Transferim la mètrica L^p induïda en $\mathcal{L}^D$ a $\mathcal{S}^D$ amb l'isomorfisme logarítmic. La mètrica resultant l'anomenem L^p-CoDa.

A Saperas-Riera *et al.* (2024) es troben els detalls d'aquest procediment. A continuació trobem les definicions de les normes L^p induïdes en l'espai composicional en els casos $p = 1, 2, \infty$.

Definició 3.15. La norma L^1-CoDa es defineix com:

$$\|\mathbf{x}\|_{1,\mathcal{S}^D} = \sum_{j=1}^{D} \left|\ln \frac{x_j}{Med(\mathbf{x})}\right|,$$

on $Med(\mathbf{x})$ denota la mediana del conjunt $\{x_1, \ldots, x_D\}$. Denotem per $\{x_{(1)}, \ldots, x_{(D)}\}$ els valors del conjunt $\{x_1, \ldots, x_D\}$ ordenats en ordre creixent, és a dir, $x_{(1)} \leq \ldots \leq x_{(D)}$. Si el conjunt $\{x_1, \ldots, x_D\}$ conté un nombre senar de valors, aleshores el valor de la mediana està perfectament definit, $Med(\mathbf{x}) = x_{((D+1)/2)}$. Si el conjunt $\{x_1, \ldots, x_D\}$ conté un nombre parell de valors, aleshores, donat que estem en un espai multiplicatiu, prenem com a representant de la mediana la mitjana geomètrica dels valors centrals: $Med(\mathbf{x}) = (x_{(D/2)}x_{(D/2+1)})^{\frac{1}{2}}$.

Definició 3.16. La norma L^2-CoDa és la norma d'Aitchison:

$$\|\mathbf{x}\|_{2,\mathcal{S}^D} = \left(\sum_{j=1}^{D} \left(\ln \frac{x_j}{g(\mathbf{x})} \right)^2 \right)^{1/2}$$

Definició 3.17. La norma L^∞-CoDa es defineix com:

$$\|\mathbf{x}\|_{\infty,\mathcal{S}^D} = \frac{1}{2} \max_{i,j} \ln \frac{x_i}{x_j}$$

3.3.3 Altres mètriques i propietats

Per concloure aquests apartats sobre mètriques en l'espai composicional, volem mostrar alguns resultats sobre desigualtats de normes. Abans però, volem definir la norma L^1-*pairwise logratio* (Saperas-Riera *et al.*, 2023). La norma d'Aitchison pot expressar-se en funció dels quadrats dels logquocients entre dues parts:

$$\|\mathbf{x}\|^2_{2,\mathcal{S}^D} = \frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right)^2$$

Des dels primers treballs en l'anàlisi composicional de dades, els logquocients entre dues parts, $\ln \frac{x_i}{x_j}$, han captat un interès notable dins la comunitat científica. Això es deu a la seva simplicitat matemàtica, que en facilita el seu càlcul i comprensió, i els converteix en eines accessibles tant per a l'anàlisi teòrica com per a l'aplicació pràctica. Durant el desenvolupament d'aquesta tesi ens vam preguntar si podia existir una norma de tipus L^1 (basada en la suma de valors absoluts) que s'expressés en funció de logquocients entre parts i que complís els principis composicionals. La proposta que fem es basa en la suma dels valors absoluts dels logquocients entre parts.

Definició 3.18. Sigui $\mathbf{x}$ una composició. La norma L^1-*pairwise logratio* es defineix com:

$$\|\mathbf{x}\|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| \ln \frac{x_i}{x_j} \right|$$

A Saperas-Riera *et al.* (2023) s'estudia amb detall la norma L^1-*plr*, i es demostra que compleix les propietats d'una norma i que verifica les propietats composicionals.

La primera desigualtat que estudiem afecta les normes que es defineixen a partir de la suma de valors absoluts: L^1-CoDa, L^1-clr i L^1-*plr*. El resultat que es presenta a continuació millora el que es presenta en l'annex de l'article Saperas-Riera *et al.* (2023).

Proposició 3.1. Per a tota composició $\mathbf{x} \in \mathcal{S}^D$, es verifica que:

$$\|\mathbf{x}\|_{1-plr} \leq \|\mathbf{x}\|_{1,\mathcal{S}^D} \leq \|\mathbf{x}\|_{1-clr}$$

Els punts de la forma $\mathrm{clr}(\mathbf{x}) = (0,\ldots,0,\underbrace{\pm\frac{a}{2}}_{i},0,\ldots,0,\underbrace{\mp\frac{a}{2}}_{j},0,\ldots,0)$ compleixen la igualtat $\|\mathbf{x}\|_{1-plr} = \|\mathbf{x}\|_{1,\mathcal{S}^D} = \|\mathbf{x}\|_{1-clr} = a$ (vegeu la Fig. 3.5).

Demostració. Provem la primera desigualtat, $\|\mathbf{x}\|_{1-plr} \leq \|\mathbf{x}\|_{1,\mathcal{S}^D}$:

$$\begin{aligned}
\|\mathbf{x}\|_{1-plr} &= \frac{1}{D-1}\sum_{i<j}\left|\ln\frac{x_i}{x_j}\right| = \frac{1}{D-1}\sum_{i<j}\left|\ln\frac{x_i/Med(\mathbf{x})}{x_j/Med(\mathbf{x})}\right| = \\
&= \frac{1}{D-1}\sum_{i<j}\left|\ln\frac{x_i}{Med(\mathbf{x})} - \ln\frac{x_j}{Med(\mathbf{x})}\right| \leq \\
&\leq \frac{1}{D-1}\sum_{i<j}\left(\left|\ln\frac{x_i}{Med(\mathbf{x})}\right| + \left|\ln\frac{x_j}{Med(\mathbf{x})}\right|\right)
\end{aligned}$$

Per a $i = 1,\ldots,D$, el terme $\left|\ln\frac{x_i}{Med(\mathbf{x})}\right|$ apareix $D-1$ vegades. Per tant,

$$\|\mathbf{x}\|_{1-plr} \leq \sum_{i=1}^{D}\left|\ln\frac{x_i}{Med(\mathbf{x})}\right| = \|\mathbf{x}\|_{1,\mathcal{S}^D}$$

La segona desigualtat és immediata per definició. La mediana, $Med(\mathbf{z})$, és el valor que minimitza la funció desviació absoluta, $\sum_{j=1}^{D}|z_i - \lambda|$:

$$\begin{aligned}
\|\mathbf{x}\|_{1,\mathcal{S}^D} = \sum_{i=1}^{D}\left|\ln\frac{x_i}{Med(\mathbf{x})}\right| &= \min_{\lambda>0}\sum_{i=1}^{D}|\ln x_i - \ln\lambda| \leq \\
&\leq \sum_{i=1}^{D}|\ln x_i - \ln g(\mathbf{x})| = \|\mathbf{x}\|_{1-\mathrm{clr}}
\end{aligned}$$

□

Per visualitzar la desigualtat de la proposició 3.1, ens situem en l'espai composicional de 4 parts, $\mathcal{S}^4$, i representarem les boles unitat corresponents a les diferents mètriques (L^1-*plr*, L^1-CoDa i L^1-clr) en coordenades clr. Denotem per $B(1)_*$ la bola unitat de la mètrica $*$. Observem que es compleix que $B(1)_{1-plr} \supseteq B(1)_{1,\mathcal{S}^D} \supseteq B(1)_{1-\mathrm{clr}}$. La bola unitat amb la mètrica L^1-*plr* correspon al poliedre cub tetrakis: format per 24 triangles isòsceles és un dels 13 poliedres de Catalan (Figura 3.5a). El cub tetrakis s'obté a partir del cub afegint un nou vèrtex cap enfora al centre de cada cara (operació *kis*). La bola unitat amb la mètrica L^1-CoDa correspon al cub: format per 6 quadrats, és un dels cinc sòlids platònics (Figura 3.5b). Finalment, la bola unitat amb la mètrica $L^1 - \mathrm{clr}$ correspon al poliedre cubooctaedre. El cubooctaedre és un dels 13 sòlids arquimedians (Figura 3.5c) i s'obté de rectificar el cub. Això vol dir truncar els vèrtexs del cub amb plans que passen pels punts mitjos de les seves arestes. Els punts mitjans de les arestes del cub corresponen a les composicions amb coordenades del tipus $\mathrm{clr}(\mathbf{x}) = (\pm\frac{1}{2}, \mp\frac{1}{2}, 0, 0)$ i totes les seves permutacions.

La segona desigualtat que presentem és ben coneguda i afecta les normes L^p-composicionals.

Proposició 3.2. Sigui $\mathbf{z} \in \mathbf{R}^D$. Sigui $1 \leq p \leq q \leq \infty$. Aleshores,

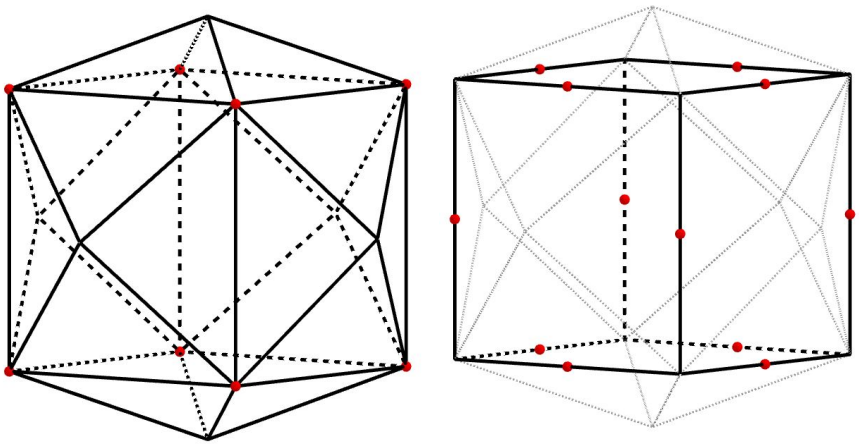
$$\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\|_q \leq \|\mathbf{z}\|_p \leq \|\mathbf{z}\|_1$$

Per construcció, aquest resultat també es pot escriure amb les normes L^p-composicionals.

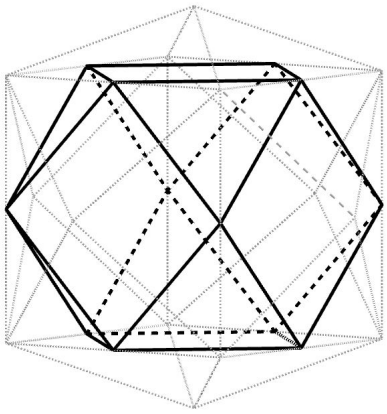
Proposició 3.3. Sigui $\mathbf{x} \in \mathcal{S}^D$. Sigui $1 \leq p \leq q \leq \infty$. Aleshores,

$$\|\mathbf{x}\|_{\infty,\mathcal{S}^D} \leq \|\mathbf{x}\|_{q,\mathcal{S}^D} \leq \|\mathbf{x}\|_{p,\mathcal{S}^D} \leq \|\mathbf{x}\|_{1,\mathcal{S}^D}$$

En la Figura 3.6, podem veure com les boles unitàries corresponents a les normes L^1-CoDa, L^2-CoDa i L^∞-CoDa estan contingudes una dins de l'altra seguint la desigualtat de la proposició 3.3. Per facilitar-ne la comprensió, es presenta el gràfic per al cas $D = 3$. A diferència de l'espai real $\mathbb{R}^D$, on les normes L^p prenen el mateix valor sobre els eixos de coordenades, $\mathbf{z} = (a, 0, \ldots, 0)$, $\|\mathbf{z}\|_p = a$, $1 \leq p \leq \infty$, les normes composicionals no coincideixen en cap composició. Això es deu al fet que l'espai composicional és un espai quocient en la direcció del vector $\mathbf{1}_D = (1, \ldots, 1)$, i no hi ha cap eix perpendicular a aquest vector.

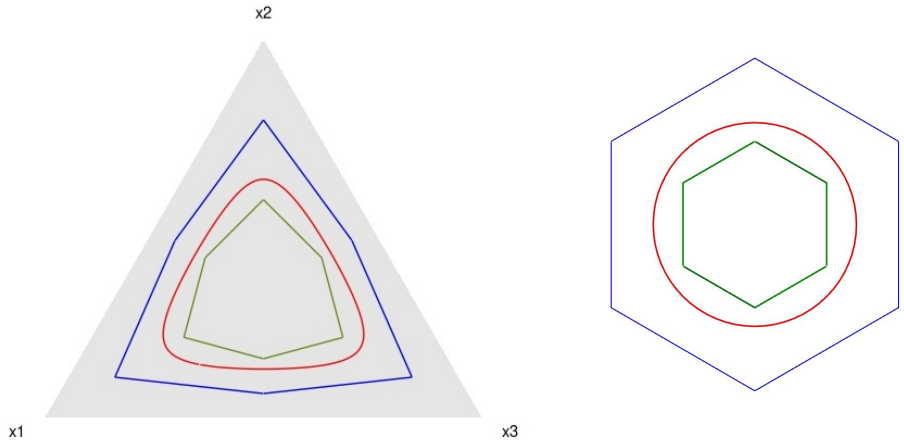


(a) $B(1)_{1-plr}$ (b) $B(1)_{1-CoDa}$



(c) $B(1)_{1-clr}$

Figura 3.5: Representació per a l'espai composicional $\mathcal{S}^4$ en coordenades olr, utilitzant la base ortonormal $\{(\frac{1}{2},\frac{1}{2},-\frac{1}{2},-\frac{1}{2}),\ (\frac{1}{2},-\frac{1}{2},\frac{1}{2},-\frac{1}{2}),\ (\frac{1}{2},-\frac{1}{2},-\frac{1}{2},\frac{1}{2})\}$: (a) Bola unitat amb la mètrica L^1-*plr*. En vermell, els vèrtexs que formen el cub corresponent a la bola unitat amb la mètrica L^1-CoDa. (b) Bola unitat amb la mètrica L^1-CoDa. En vermell, els vèrtexs que formen el cubooctaedre corresponent a la bola unitat amb la mètrica L^1-clr. (c) Bola unitat amb la mètrica L^1-clr.



(a) (b)

Figura 3.6: Boles unitàries corresponents a les normes L^1-CoDa (verd), L^2-CoDa (vermell) i L^∞-CoDa (blau) representades: (a) en el símplex de tres parts, (b) en coordenades olr.

3.4 El model lineal amb covariable composicional

Una funció important definida sobre $\mathcal{S}^D$ que és invariant per canvi d'escala és el *logcontrast*. Un logcontrast es defineix com qualsevol combinació lineal dels logaritmes de les parts d'una composició, en què la suma dels coeficients és zero: $\sum_{j=1}^{D} a_i \ln x_i$, $\sum_{j=1}^{D} a_i = 0$, $a_i \in \mathbb{R}$. El logcontrast exerceix el paper típic de la combinació lineal de variables. Un exemple de logcontrast és la diferència logarítmica entre dues parts composicionals: $\ln \frac{x_i}{x_j} = \ln x_i - \ln x_j$. En el context estadístic, els logcontrastos es poden utilitzar en models de regressió (Aitchison, 1984; Hron *et al.*, 2012), anàlisi de components principals (Aitchison, 1983) i altres tècniques multivariants adaptades per treballar amb dades composicionals. La importància dels logcontrastos en l'anàlisi composicional es reflecteix en la seva capacitat per simplificar i fer més comprensibles les relacions entre les parts d'una composició, així com per millorar l'eficiència i la interpretabilitat dels models estadístics.

Definició 3.19. Donada una variable dependent y i una composició explicativa de D-parts $\mathbf{x}$, un *model de regressió lineal* en termes de logcontrast

es pot expressar com:

$$y = a_0 + \sum_{j=1}^{D} a_j \ln x_j + \epsilon, \ \sum_{j=1}^{D} a_j = 0, \tag{3.2}$$

on a_0 és el terme independent, $\boldsymbol{a} = (a_1, \ldots, a_D)$ és el vector de coeficients del model i ϵ és el terme d'error.

La restricció $\sum_{j=1}^{D} a_j = 0$ és necessària per garantir la naturalesa composicional del model. Altrament, podem escriure la definició del model de regressió lineal en termes mètrics (Boogaart i Tolosana, 2013):

$$y = \beta_0 + \langle \boldsymbol{\beta}, \ \mathbf{x} \rangle_{\mathcal{A}} + \epsilon, \tag{3.3}$$

on β_0 és el terme independent, $\boldsymbol{\beta}$ és el coeficient de regressió i ϵ és el terme d'error. Notem que els models de regressió lineal anteriors (3.2, 3.3) són equivalents ($a_0 = \beta_0, \boldsymbol{a} = \operatorname{clr}(\boldsymbol{\beta})$).

3.5 Regressió LASSO

La regressió LASSO és una tècnica de regressió que combina la selecció de variables i la regularització per millorar la predicció i la interpretabilitat dels models estadístics. Desenvolupada per Tibshirani (1996), LASSO s'ha convertit en una eina popular en l'anàlisi de dades per la seva capacitat per gestionar models amb un gran nombre de covariables predictores, especialment quan aquestes covariables predictores estan altament correlacionades.

LASSO aconsegueix la selecció de variables i la regularització mitjançant l'addició d'un terme de penalització a la funció de cost de la regressió lineal tradicional. Aquest terme de penalització és proporcional a la suma dels valors absoluts dels coeficients dels predictors, la qual cosa força alguns coeficients a ser exactament zero. Així, LASSO no només redueix la complexitat del model eliminant predictors irrellevants, sinó que també ajuda a prevenir el sobreajustament.

Definició 3.20. Sigui y_i la i-èsima observació de la variable resposta i $\mathbf{Z}$ la matriu de dimensió $n \times D$, on cada fila $\mathbf{z}_i = (z_{i1}, \ldots, z_{iD})$ conté les observacions de la covariable explicativa. El *model LASSO* es pot expressar com la solució del següent problema d'optimització restringida:

$$\min_{a_0, a_1, \ldots, a_D} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - a_0 - \sum_{j=1}^{D} z_{ij} a_j \right)^2 \right\},$$

subjecte a

$$\sum_{j=1}^{D} |a_j| \leq t$$

per a $t > 0$. Seguint amb la mateixa notació, a_0 és el terme independent i $\boldsymbol{a} = (a_1, \ldots, a_D)$ és el vector de coeficients del model.

Aquest és un problema de programació quadràtica en què la primera part de l'expressió, $\frac{1}{2}\sum_{i=1}^{n}\left(y_i - a_0 - \sum_{j=1}^{D} z_{ij}a_j\right)^2$, és el terme de pèrdua quadràtica que mesura l'ajust del model lineal a les dades. La segona part, la restricció $\sum_{j=1}^{D} |a_j| \leq t$, promou l'encongiment dels coeficients a_j, reduint-ne alguns exactament a zero. En la bibliografia, és comú presentar el problema d'optimització en la forma de regressió penalitzada:

$$\min_{a_0, a_1, \ldots, a_D} \left\{ \frac{1}{2}\sum_{i=1}^{n}\left(y_i - a_0 - \sum_{j=1}^{D} z_{ij}a_j\right)^2 + \lambda \sum_{j=1}^{D} |a_j| \right\},$$

on λ és el paràmetre que controla la quantitat de penalització del model.

A més, la flexibilitat de LASSO ha portat al desenvolupament de diverses extensions i variants, com ara l'*elastic net* (Wang i Hastie, 2005), que combina les penalitzacions de LASSO i Ridge (James *et al.*, 2021), i el *LASSO generalitzat* (Tibshirani i Taylor, 2011), que permet penalitzacions més complexes. Aquestes extensions permeten adaptar la tècnica de LASSO a diferents tipus de problemes i dades, fet que n'augmenta encara més la utilitat en diverses àrees de recerca i aplicacions pràctiques.

Definició 3.21. Sigui y_i la i-èsima observació de la variable resposta i $\mathbf{Z}$ la matriu de dimensió $n \times D$, on cada fila $\mathbf{z}_i = (z_{i1}, \ldots, z_{iD})$ conté les observacions de la covariable explicativa. El *model LASSO generalitzat* es pot expressar com la solució del següent problema de regressió penalitzada:

$$\min_{a_0, \boldsymbol{a}} \left\{ \frac{1}{2}\sum_{i=1}^{n}\left(y_i - a_0 - \langle \boldsymbol{a}, \mathbf{z}_i \rangle_E\right)^2 + \lambda \|\mathbf{F}\boldsymbol{a}\|_1 \right\},$$

on a_0 és el terme independent, $\boldsymbol{a} = (a_1, \ldots, a_D)$ és el vector de coeficients del model, $\mathbf{F}$ és una matriu $m \times D$, $\langle\ ,\ \rangle_E$ és el producte escalar euclidià, $\|\cdot\|_1$ és la norma L^1 habitual i λ és el paràmetre que controla la quantitat de penalització del model.

3.5.1 Regressió LASSO amb covariable composicional: selecció de variables

Fins ara, l'aplicació de la regressió LASSO a les dades composicionals s'ha centrat a adaptar el procés de regularització LASSO estàndard als casos en què la covariable explicativa és una composició. Aquest procés de regularització s'explica com un procés que permet discriminar entre les parts d'una composició que influeixen en la variable resposta i les parts que poden ser eliminades del model. Lin *et al.* (2014) proposen un procés de regularització basat en la norma L^1 per al model lineal del logcontrast que respecta la naturalesa composicional del problema. El problema es formula com un problema d'optimització convexa restringida.

Definició 3.22. Sigui y_i la i-èsima observació de la variable resposta i $\mathbf{X}$ la matriu de dimensió $n \times D$, on cada fila $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ conté les observacions de la covariable composicional explicativa. El *model LASSO composicional restringit* es pot expressar com la solució del següent problema de regressió penalitzada:

$$\min_{a_0, \boldsymbol{a}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - a_0 - \langle \boldsymbol{a}, \ln \mathbf{x}_i \rangle_E\right)^2 + \lambda \|\boldsymbol{a}\|_1 \right\}, \ \sum_{j=1}^{D} a_i = 0,$$

on a_0 és el terme independent, $\boldsymbol{a} = (a_1, \ldots, a_D)$ és el vector de coeficients del model, $\langle\ ,\ \rangle_E$ és el producte escalar euclidià, $\|\cdot\|_1$ és la norma L^1 habitual i λ és el paràmetre que controla la quantitat de penalització del model.

Bates i Tibshirani (2018) presenten un problema d'optimització equivalent basat en un model lineal que conté totes les parelles del tipus $\ln \frac{x_i}{x_j}$:

Definició 3.23. Sigui y_i la i-èsima observació de la variable resposta i $\mathbf{X}$ la matriu de dimensió $n \times D$, on cada fila $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ conté les observacions de la covariable composicional explicativa. El *model LASSO logratio* es pot expressar com la solució del següent problema de regressió penalitzada:

$$\min_{\theta_0, \boldsymbol{\theta}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{1 \leq k < j \leq D} \theta_{kj} \ln \frac{x_{ik}}{x_{ij}} \right)^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\},$$

on θ_0 és el terme independent, $\boldsymbol{\theta} = (\theta_{12}, \theta_{13}, \ldots, \theta_{(D-1)D})$ és el vector de coeficients del model i està definit per θ_{kj} amb $k < j$, $\|\cdot\|_1$ és la norma L^1 habitual i λ és el paràmetre que controla la quantitat de penalització del model.

La selecció de variables en el camp de les dades composicionals ha tingut un gran impacte i ha trobat aplicacions destacades en diversos camps, especialment en l'estudi del microbioma, en què l'alta dimensionalitat del problema representa un repte important (Calle *et al.*, 2023). Aquesta alta dimensionalitat fa que la identificació dels tàxons més rellevants sigui essencial per entendre millor el model. La capacitat de seleccionar eficientment aquests tàxons, tot respectant la naturalesa composicional de les dades, és fonamental per obtenir resultats fiables i significatius en l'estudi del microbioma (Gloor *et al.*, 2017).

El *LASSO path plot* és un gràfic que mostra com evolucionen els coeficients d'un model de regressió LASSO a mesura que augmenta el terme de penalització. En aquest tipus de gràfica, l'eix vertical representa els valors dels coeficients dels predictors, mentre que l'eix horitzontal representa el $\log(\lambda)$. En la figura 3.7, veiem un exemple de procés de regularització LASSO amb covariable composicional on els coeficient clr tendeixen a zero, eliminant les variables predictores menys rellevants per tal d'aconseguir un model més senzill.

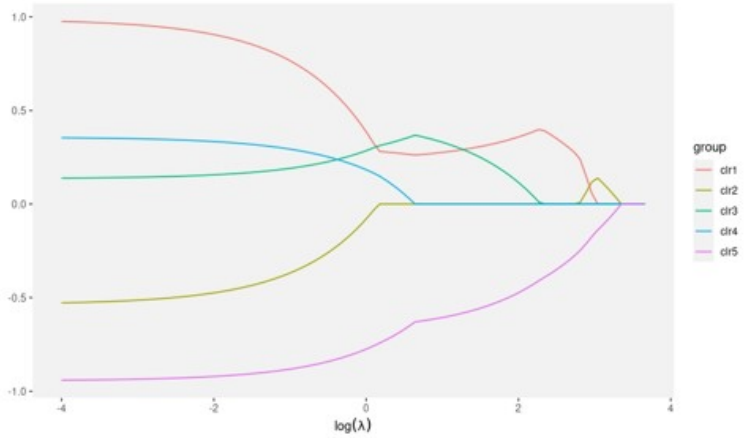


Figura 3.7: LASSO path plot: en diferents colors els camins que segueixen els coeficients clr al llarg del procés de regularització LASSO

3.5.2 Regressió LASSO amb covariable composicional: selecció de balanços

Una de les contribucions d'aquesta tesi és reinterpretar la regressió LASSO composicional restringida utilitzant els elements mètrics propis de l'espai composicional (Saperas-Riera *et al.*, 2024).

Definició 3.24. Sigui y_i la i-èsima observació de la variable resposta i $\mathbf{X}$ la matriu de dimensió $n \times D$, on cada fila $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ conté les observacions de la covariable composicional explicativa. El model L^1-clr *LASSO* s'expressa com la solució del següent problema d'optimització:

$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_{\mathcal{A}}\right)^2 + \lambda \|\boldsymbol{\beta}\|_{1-\mathrm{clr}} \right\},$$

on β_0 és el terme independent, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)$ és el vector de coeficients del model, $\langle\ ,\ \rangle_{\mathcal{A}}$ és el producte escalar d'Aitchison, $\|\boldsymbol{\beta}\|_{1-clr} = \sum_{j=1}^{D} \left| \ln \frac{\beta_j}{g(\boldsymbol{\beta})} \right|$ i λ és el paràmetre que controla la quantitat de penalització del model.

La importància d'aquesta definició, més enllà de descriure el problema de regularització en termes estrictament composicionals, resideix en el fet que el terme de penalització no selecciona directament parts de la composició, sinó que selecciona balanços. Concretament, selecciona balanços de la forma $\ln \frac{\beta_i}{g(\boldsymbol{\beta})}$, o sigui, en les direccions del sistema generador canònic $\mathbf{W}$. La conseqüència d'aquest procés de selecció és que podem expressar el model lineal en funció d'una subcomposició i eliminar la resta de les parts de la composició del model lineal. Sense pèrdua de generalitat, suposem que el balanç $\ln \frac{\beta_1}{g(\boldsymbol{\beta})}$ és zero, indicant que la relació entre la part x_1 i la mitjana geomètrica de la resta de parts, $g(x_2, \ldots, x_D)$, no és rellevant. En conseqüència, el model lineal acumula tota la seva variació en el subespai ortogonal al vector $\ln \frac{x_1}{g(\mathbf{x})}$, o sigui, en el subespai generat per la subcomposició $(x_2, \ldots, x_D)$, i això elimina la part x_1 de la nostra anàlisi.

Proposició 3.4. El model L^1-clr LASSO és equivalent al model LASSO composicional restringit.

Demostració. Partim del model LASSO composicional restringit:

$$\min_{a_0, \boldsymbol{a}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - a_0 - \langle \boldsymbol{a}, \ln \mathbf{x}_i \rangle_E\right)^2 + \lambda \|\boldsymbol{a}\|_1 \right\},\ \sum_{j=1}^{D} a_i = 0.$$

Per tal d'eliminar la restricció $\mathbf{1}_D \boldsymbol{a} = \sum_{j=1}^{D} a_i = 0$, considerem un conjunt de vectors que generin el nucli de l'aplicació $\mathbf{1}_D : \mathbb{R}^D \to \mathbb{R}$. Concretament, prenem el conjunt de vectors de la forma $\mathbf{v}_i = (-\frac{1}{D}, \ldots, 1 - \frac{1}{D}, \ldots, -\frac{1}{D})$, amb $1 - \frac{1}{D}$ ocupant la i-èsima posició. Definim el canvi de variable,

$$\boldsymbol{a} = \mathbf{H} \ln \boldsymbol{\beta},$$

on la matriu $\mathbf{H}$ té per columnes els vectors $\mathbf{v}_i$. La restricció queda escrita de la següent manera, $\mathbf{1}_D\boldsymbol{a} = \mathbf{1}_D\mathbf{H}\ln\boldsymbol{\beta} = \mathbf{0}\ln\boldsymbol{\beta} = 0$. La restricció pot ser eliminada ja que es compleix per a tota composició $\boldsymbol{\beta}$. Per tant, el problema de minimització és:

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\beta_0-\langle\mathbf{H}\ln\boldsymbol{\beta},\ln\mathbf{x}_i\rangle_E\right)^2+\lambda\|\mathbf{H}\ln\boldsymbol{\beta}\|_1\right\},$$

on $\beta_0 = a_0$. Si escrivim el resultat utilitzant els elements mètrics composicionals, $\mathbf{H}\ln\boldsymbol{\beta} = \mathrm{clr}\boldsymbol{\beta}$ i $\langle\mathrm{clr}\boldsymbol{\beta},\ln\mathbf{x}_i\rangle_E = \langle\boldsymbol{\beta},\mathbf{x}_i\rangle_{\mathcal{A}}$, tenim

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\beta_0-\langle\boldsymbol{\beta},\mathbf{x}_i\rangle_{\mathcal{A}}\right)^2+\lambda\|\boldsymbol{\beta}\|_{1-\mathrm{clr}}\right\}.$$

□

Per tant, cal entendre el procés de regularització amb covariables composicionals, conegut com a selecció de variables, com un cas particular de selecció de balanços. Això significa que, en lloc de seleccionar directament parts individuals de la composició, el procés identifica quines relacions entre les parts (balanços) són més rellevants per a la variable de resposta. Aquesta perspectiva no només preserva la naturalesa composicional de les dades, sinó que també permet una interpretació més rica i precisa de les influències relatives entre les parts sobre la variable resposta. Així, la selecció de balanços esdevé una eina poderosa per a la simplificació i millora de models estadístics en l'anàlisi de dades composicionals. Amb l'objectiu d'enriquir i ampliar les eines metodològiques per a la regularització dels models lineals amb covariables composicionals mitjançant la selecció de balanços, aquesta tesi presenta alternatives a la selecció de variables (Saperas-Riera *et al.*, 2023, 2024).

En l'anàlisi composicional, la informació rellevant és la informació relativa entre les parts, en lloc de les magnituds absolutes de cada part. Aquesta idea es reflecteix en l'estudi dels models lineals, en què l'objectiu és analitzar la influència dels vectors logquocients sobre la variable resposta, així com els coeficients associats a aquests vectors. Quan expressem un model lineal en una base de logquocients, els coeficients representen el pendent o gradient del model lineal. En aquest context, si el coeficient associat a un logquocient del model lineal és zero, això indica que la variable resposta és independent d'aquest logquocient específic. No obstant això, aquest coeficient zero no implica necessàriament la independència de la variable resposta respecte a les parts individuals que componen el logquocient. Aquesta independència

es refereix només a la relació específica capturada pel logquocient, no de manera individual a les parts que el conformen. Per tant, en l'anàlisi de dades composicionals és crucial entendre que la selecció de logquocients significatius, a través de procediments com la regressió LASSO, permet identificar relacions rellevants entre les parts de la composició que influeixen en la variable resposta. Aquesta perspectiva subratlla la importància de treballar amb logquocients en lloc de parts individuals. En aquest context, Boogaart *et al.* (2013) introdueixen les nocions de subcomposició internament independent i subcomposició externament independent respecte a la variable explicada y.

Definició 3.25. En un model lineal amb covariable composicional, una subcomposició és *internament independent* si la variable explicada no varia quan canvien les relacions entre les parts que componen aquesta subcomposició, suposant que la resta de relacions es conserven.

Quan una subcomposició de k parts és internament independent, podem visualitzar l'impacte sobre el model lineal a través de les coordenades log-transformades del vector gradient. Per a fer-ho, considerem una base olr de la subcomposició internament independent, $\{\mathbf{e}_j,\ j = 1, \ldots, k-1\}$, i la completem fins a obtenir una base ortonormal per a tot l'espai composicional, $\{\mathbf{e}_j,\ j = k, \ldots, D-1\}$. En aquesta base, les $k-1$ primeres coordenades del vector gradient seran zero, $\mathrm{olr}(\boldsymbol{\beta})_j = 0,\ j = 1, \ldots, k-1$, fet que indica que les variacions dins de la subcomposició no tenen efecte sobre la variable resposta. Si, en canvi, expressem el model lineal en coordenades clr, aquesta independència interna de la subcomposició es manifesta de manera diferent. Sense pèrdua de generalitat, suposem que les k parts de la subcomposició internament independent són les k primeres, $x_j,\ j = 1, \ldots, k$. Les coordenades clr del vector gradient associades a les parts de la subcomposició seran iguals, $\mathrm{clr}(\boldsymbol{\beta})_1 = \ldots = \mathrm{clr}(\boldsymbol{\beta})_k$. En aquest context, el model captura com a influent la relació que hi ha entre la mitjana geomètrica de les k parts de la subcomposició internament independent i la resta de parts.

La forma més elemental de subcomposició internament independent es presenta quan la relació relativa entre dues parts (*pairwise*) és no influent. En general, per identificar una subcomposició internament independent, és suficient determinar si les relacions dos a dos entre les parts de la subcomposició són no influents en la variable resposta. Per aquest motiu, quan volem regularitzar un model tot identificant subcomposicions internament independents, les mètriques utilitzades en el terme de penalització es basen en *pairwise* (L^1-CoDa i L^1-*plr*). A Saperas-Riera *et al.* (2023), s'estudia en

detall el procés de regularització amb la norma L^1-*plr* en el terme de penalització. El punt fort d'aquesta mètrica és que permet identificar múltiples subcomposicions internament independents.

Definició 3.26. Sigui y_i la i-èsima observació de la variable resposta i $\mathbf{X}$ la matriu de dimensió $n \times D$, on cada fila $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ conté les observacions de la covariable composicional explicativa. El model L^1-*plr LASSO* s'expressa com la solució del següent problema d'optimització:

$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_{\mathcal{A}} \right)^2 + \lambda \|\boldsymbol{\beta}\|_{1-plr} \right\},$$

on β_0 és el terme independent, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)$ és el vector de coeficients del model, $\langle\ ,\ \rangle_{\mathcal{A}}$ és el producte escalar d'Aitchison, $\|\boldsymbol{\beta}\|_{1-plr} = \sum_{i<j} \left| \ln \frac{\beta_i}{\beta_j} \right|$ i λ és el paràmetre que controla la quantitat de penalització del model.

En canvi, a Saperas-Riera *et al.* (2024) s'estudia el procés de regularització amb la norma L^1-CoDa en el terme de penalització i es compara el model lineal resultant amb els models lineals obtinguts mitjançant les regularitzacions amb les normes L^1-clr i L^1-*plr*.

Definició 3.27. Sigui y_i la i-èsima observació de la variable resposta i $\mathbf{X}$ la matriu de dimensió $n \times D$, on cada fila $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ conté les observacions de la covariable composicional explicativa. El model L^1-*CoDa LASSO* s'expressa com la solució del següent problema d'optimització:

$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_{\mathcal{A}} \right)^2 + \lambda \|\boldsymbol{\beta}\|_{1-CoDa} \right\},$$

on β_0 és el terme independent, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)$ és el vector de coeficients del model, $\langle\ ,\ \rangle_{\mathcal{A}}$ és el producte escalar d'Aitchison, $\|\boldsymbol{\beta}\|_{1-CoDa} = \sum_{j=1}^{D} \left| \ln \frac{\beta_j}{Med(\boldsymbol{\beta})} \right|$ i λ és el paràmetre que controla la quantitat de penalització del model.

El *LASSO path plot* també és una eina valuosa per comprendre el procés de regularització quan seleccionem logquocients entre dues parts. En la figura 3.8 presentem un nou gràfic que mostra com les diferències en els coeficients clr es redueixen progressivament cap a zero a mesura que augmenta el paràmetre de regularització λ. La novetat d'aquest gràfic rau en el fet que, en lloc de centrar l'atenció en el punt en què el camí del coeficient clr es fa zero (com es fa tradicionalment en la selecció de variables), l'atenció es posa

en el moment en què dos camins clr es fusionen. Això permet identificar quins logquocients entre dues parts s'eliminen a mesura que la regularització s'intensifica, oferint informació valuosa sobre la selecció de logquocients entre dues parts i l'esparsitat del model en termes de balanços.

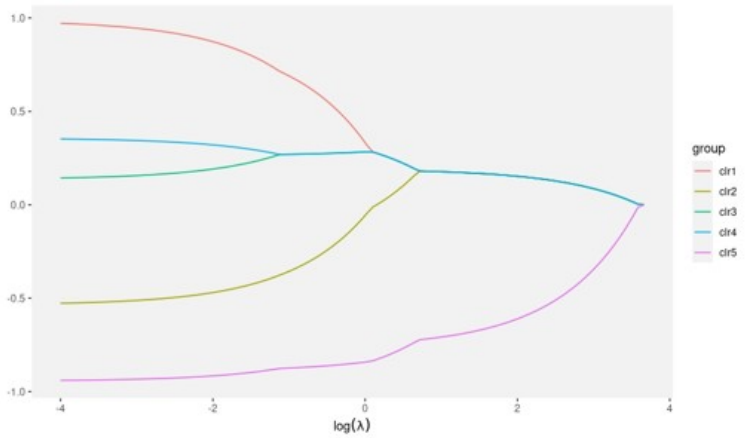


Figura 3.8: LASSO path plot: en diferents colors els camins que segueixen els coeficients clr al llarg del procés de regularització LASSO

Per tant, aquesta tesi aporta la teoria i la metodologia suficient per identificar, a través d'un procediment no supervisat, la subcomposició o subcomposicions internament independents.

Definició 3.28. En un model lineal amb covariable composicional, una subcomposició és *externament independent* si la variable explicada no varia quan varia el balanç de les parts que componen aquesta subcomposició i la resta de parts de la composició, suposant que la resta de relacions es conserven.

La primera observació que cal fer és que si una subcomposició de k parts és externament independent, la subcomposició que conté la resta de $D - k$ parts de la composició també és externament independent.

Com en el cas anterior, quan una subcomposició de k parts és externament independent, podem visualitzar l'impacte sobre el model lineal a través de les coordenades logquocient del vector gradient. Per tant, construïm una base olr adequada: primer, prenem una base olr de la subcomposició de k parts externament independent, $\mathbf{e}_j$, $j = 1, \ldots, k-1$. Seguidament, la completem fins a tenir una base ortonormal de tot l'espai afegint el vector que balanceja les parts de la subcomposició amb la resta de parts de

la composició, $\mathbf{e}_k$, i una base ortonormal per a la resta de $D-k$ parts, $\mathbf{e}_j,\ j=k+1,\ldots,D-1$. En aquesta base, el vector gradient té la k-èsima coordenada igual a zero, $\text{olr}(\boldsymbol{\beta})_k = 0$. Cal notar que el model lineal queda expressat en funció de balanços: els primers $k-1$ balanços depenen únicament de les k parts que formen la primera subcomposició externament independent, i els següents $D-k-1$ balanços depenen únicament de les $D-k$ parts que formen la segona subcomposició externament independent. Si ara s'expressa el model lineal en coordenades clr, observarem que els k coeficients corresponents a les parts de la primera subcomposició sumen zero i els $D-k$ coeficients corresponents a les parts de la segona subcomposició també sumen zero. Per tant, la composició inicial queda partida en dues subcomposicions que per si mateixes es poden interpretar com a composicions. Per tant, el model lineal pot ser interpretat com a suma de dos models lineals amb covariable composicional.

El cas més senzill de subcomposició externament independent és el cas degenerat, en què la primera subcomposició està formada per una part i la segona subcomposició està formada per $D-1$ parts. La mètrica L^1-clr el que fa és explorar els D balanços de la forma una part contra les altres $D-1$ parts i en selecciona quins són no influents sobre la variable resposta. Si un balanç de la forma $\ln \frac{x_j}{g(\mathbf{x})}$ és no influent, la composició queda dividida en dues subcomposicions: una primera subcomposició degenerada, $\{x_j\}$, que és eliminada del model lineal i una segona subcomposició, $\{x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_D\}$, que passa a ser la composició que fa de covariable composicional en el model lineal. Una futura línia de recerca és buscar una mètrica que generalitzi la mètrica L^1-clr i que permeti explorar de manera exhaustiva totes les possibles subcomposicions externament independents.

Una propietat destacable és que quan una subcomposició de k parts és doblement independent, és a dir, internament i externament independent, aleshores la màxima variació de la variable explicada es deu exclusivament a les relacions entre les $D-k$ parts que no formen part de la subcomposició, i per tant les parts d'aquesta subcomposició doblement independent es poden eliminar del model. La manera fàcil d'entendre aquesta propietat és construint una base olr: primer, prenem una base olr de la subcomposició de k parts, $\mathbf{e}_j,\ j=1,\ldots,k-1$. Seguidament, la completem fins a tenir una base ortonormal de tot l'espai amb el vector que balanceja les k parts de la subcomposició amb la resta de $D-k$ parts de la composició, $\mathbf{e}_k$, i una base ortonormal per a la resta de $D-k$ parts que no formen part de la subcomposició, $\mathbf{e}_j,\ j=k+1,\ldots,D-1$. En aquesta base, el vector gradient té les primeres k coordenades igual a zero, $\text{olr}(\boldsymbol{\beta})_j = 0,\ j=1,\ldots,k$. A

Saperas-Riera *et al.* (2023) es proposa que una vegada finalitzat el procés de regularització L^1-*plr* LASSO, es testegi si alguna de les subcomposicions internament independents és també externament independent. D'aquesta manera no s'imposa el procés de selecció de variable com un prerequisit, sinó que sorgeix com a conseqüència d'estudiar les propietats d'independència interna i externa sobre una mateixa subcomposició. Actualment ja existeixen en la bibliografia diferents eines que permeten fer inferència sobre els coeficients de models de regressió, ja sigui LASSO (Chatterjee i Lahiri, 2011; Lee *et al.*, 2016) o LASSO generalitzat (Hyun *et al.*, 2018).

Capítol 4

Articles

4.1 Statistics and Operations Research Transactions

Aquest primer article cobreix el primer objectiu que ens vam marcar a l'inici de la tesi: O1. Adaptar la teoria sobre convexitat a l'espai composicional (2.1). En aquest article s'adapten les definicions de conjunt convex, funció convexa i funció quasi-convexa a l'espai composicional. A continuació, s'estudia la convexitat dels conjunts que apareixen habitualment en la bibliografia. També s'adapta la definició d'envolupant convexa. Finalment, es compara el resultat de resoldre un problema d'optimització convexa utilitzant la geometria d'Aitchison amb el resultat obtingut amb la geometria euclidiana (O5 2.1). Es confirma que usant una geometria no adequada es poden obtenir resultats incoherents.

L'article ha estat publicat a la revista Statistics and Operations Research Transactions (SORT).
Volum: 47, número: 2, pàgines: 323-344
Enviat: novembre 2022, acceptat: juny 2023
DOI: 10.57645/20.8080.02.11
Factor d'impacte: 0.7 (Q3).

SORT 47 (2) July-December 2023, 323-344 DOI: 10.57645/20.8080.02.11

Fundamentals of convex optimization for compositional data

Jordi Saperas Riera[1], Josep Antoni Martín Fernández[3]
and Glòria Mateu Figueras[2]

Abstract

Many of the most popular statistical techniques incorporate optimisation problems in their inner workings. A convex optimisation problem is defined as the problem of minimising a convex function over a convex set. When traditional methods are applied to compositional data, misleading and incoherent results could be obtained. In this paper, we fill a gap in the specialised literature by introducing and rigorously defining novel concepts of convex optimisation for compositional data according to the Aitchison geometry. Convex sets and convex functions on the simplex are defined and illustrated.

MSC: *62F40, 62H15, 62H99, 62J10, 62J15, 62P25.*

Keywords: *Compositional data, logratio, simplex, proportion, function, convexity, optimisation.*

1. Introduction

Convex optimisation plays an important role in a wide range of scientific fields where statistical methods are applied. Thus, it has become particularly relevant in the design of experiments (Coetzer and Haines, 2017), variable selection (Susin et al., 2020), robust statistics (Boogaart et al., 2021), cluster analysis (Wang et al., 2020), and principal components analysis (Campbell and Wong, 2022).

In a broad sense, a convex optimisation problem in $\mathbb{R}^D$ is defined as (Boyd and Vandenberghe, 2004) the problem of minimising a convex function (objective function) over a convex set (feasible region). The feasible region is the set of all possible values of variables that verify the constraints of the problem. Commonly, it is assumed

[1,2,3] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi, Edifici P-4, 17003 Girona, Spain.

[1] jordi.saperas@udg.edu

[2] gloria.mateu@udg.edu

[3] josepantoni.martin@udg.edu

Received: November 2022
Accepted: June 2023

that the convexity of the objective function and the feasible region corresponds to the ordinary convexity concept in real space, $\mathbb{R}^D$. However, in statistics, one has to take into account the particular geometry of the sample space of the variables. Compositional data (CoDa) (Aitchison, 1986) convey relative information because the variables describe relative contributions to a given total. These variables are called *parts* of a whole and are, usually, expressed in proportions, percentages or ppm. Historically (Aitchison, 1986), the sample space of CoDa is designed as the D-part unit simplex $\mathcal{S}^D = \{\mathbf{x} \in \mathbb{R}^D : x_j > 0; \sum x_j = 1; j = 1, \ldots, D\}$. The formal geometric framework for the analysis of CoDa was first introduced in Pawlowsky-Glahn and Egozcue (2001) and Billheimer, Guttorp and Fagan (2001). It was called the Aitchison geometry, and it was later formally established in Barceló-Vidal and Martín-Fernández (2016). Such geometry allows compositions to be expressed as coordinates on an orthonormal basis, formed by logratios and called olr-coordinates (Egozcue et al., 2003; Martín-Fernández, 2019). In short, the analysis of CoDa involves the use of standard techniques in olr-coordinates; what has been known as the principle of working on coordinates (Mateu-Figueras, Pawlowsky-Glahn and Egozcue, 2011).

Applications of the log-ratio methodology are increasingly found across a wide range of scientific fields such as geosciences (Martín-Fernández et al., 2018), chemistry and physics (Halim et al., 2021), and health (Bates and Tibshirani, 2019; Dumuid et al., 2020), among others. CoDa methods have become particularly relevant for the analysis of time-use data, especially in relation to public health studies (Chastin et al., 2015; Dumuid et al., 2021; Kitano et al., 2020; Fairclough et al., 2018; Gupta et al., 2020). In time-use data, the parts are the time spent on different activities such as sleep, work, and a range of physical activities. Some optimisation problems of practical relevance can be formulated in this context. For example, one can be interested in deciding what composition of daily activities agreeing with some medical recommendations (constraints) optimises a particular health biomarker (objective function). Alternatively, given a time-use composition of a patient not following some medical recommendation (feasible region), the question of interest is: what modification to the daily activity composition would enable a swift transition (objective function) into a healthy behaviour region?

The aim of this work is to adapt the concepts related to convex optimisation for problems involving CoDa and considering the Aitchison geometry. These definitions are essential to correctly identify and classify convex optimisation problems on the simplex. The paper is organised as follows. Section 2 summarises the basic concepts of CoDa. In Section 3, the concept of convex set on the simplex with the Aitchison geometry is defined. The basic sets on the simplex are analysed and their compositional convexity is studied. Section 4 develops the concept of compositional convex function and introduces the conditions to classify a problem as a compositional convex optimisation problem. These are illustrated in Section 5 through a case study based on time-use data. Finally, Section 6 concludes with some important remarks.

The analyses discussed in this article were carried out in R (R-Core-Team, 2022) and using the package *compositions* (van den Boogaart and Tolosana-Delgado, 2008).

2. Basic elements of Aitchison geometry

When analysing CoDa one assumes the property of *scale invariance*. That is, it is assumed that each D-part composition $\mathbf{w} \in \mathbb{R}^D_+$ is a member of an equivalence class (Barceló-Vidal and Martín-Fernández, 2016). In other words, the information contained in $\mathbf{w}$ is the same as in any other composition $k \cdot \mathcal{C}[\mathbf{w}]$ for any real scalar $k > 0$, where $\mathcal{C}[\mathbf{w}]$ is the closure operation defined by $\mathcal{C}[\mathbf{w}] = [w_1/\sum w_j, w_2/\sum w_j, \ldots, w_D/\sum w_j] = \mathbf{x} \in \mathcal{S}^D$.

The *perturbation operation*, $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \ldots, x_D y_D]$, defined on $\mathcal{S}^D \times \mathcal{S}^D$, and the *power transformation*, $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha]$, defined on $\mathbb{R} \times \mathcal{S}^D$, induce a vector space structure on the simplex $\mathcal{S}^D$ (Pawlowsky-Glahn and Egozcue, 2001). Another important element is the *logcontrast*, a log-linear combination

$$\sum_{i=1}^{D} \beta_i \ln x_i, \quad \text{with} \quad \sum_{i=1}^{D} \beta_i = 0, \quad \beta_i \in \mathbb{R} \tag{1}$$

which plays the typical role of the linear combination of variables (Aitchison, 1986).

Once we have a vector space structure, a metric structure is easily defined using the clr scores of a composition $\mathbf{x}$ (Aitchison, 1986):

$$\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})} \right),$$

where $g(\cdot)$ means the geometric mean. Note that the $clr(\mathbf{x})_k$ score, being a logcontrast that involves all the parts of a composition, provides information about the relative importance of part x_k in the composition. The basic metric elements of the Aitchison geometry as inner product ($<\cdot,\cdot>_{\mathcal{A}}$), norm ($||\cdot||_{\mathcal{A}}$), and distance ($d_{\mathcal{A}}(\cdot,\cdot)$) can be defined as:

$$<\mathbf{x},\mathbf{y}>_{\mathcal{A}} = <\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y})>_E , \quad ||\mathbf{x}||^2_{\mathcal{A}} = <\mathbf{x},\mathbf{x}>_{\mathcal{A}} , \quad d_{\mathcal{A}}(\mathbf{x},\mathbf{y}) = ||\mathbf{x} \ominus \mathbf{y}||_{\mathcal{A}},$$

where "$\mathcal{A}$" means the Aitchison geometry, "E" means the typical Euclidean geometry, and "$\ominus$" is the perturbation difference $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$.

The metric elements are used to construct orthonormal basis and to calculate the corresponding log-ratio coordinates of a composition ($\mathrm{olr}(\mathbf{x})$) (Egozcue et al., 2003; Martín-Fernández, 2019). The expression of these olr-coordinates depends on the selected basis. For example, following Egozcue and Pawlowsky-Glahn (2005) one can define particular olr-coordinates created through a sequential binary partition (SBP) of a complete composition $\mathbf{x} = (x_1, \ldots, x_D)$. In the first step of an SBP, when the first olr-coordinate is created, the complete composition $\mathbf{x} = (x_1, \ldots, x_D)$ is split into two groups of parts: one for the numerator and the other for the denominator. In the following steps, to create the following olr-coordinates, each group is in turn split into two groups. That is, in step k when the $\mathrm{olr}(\mathbf{x})_k$-coordinate is created, the r_k parts $(x_{n1_k}, \ldots, x_{nr_k})$ in the first group are placed in the numerator; the s parts $(x_{d1_k}, \ldots, x_{ds_k})$ in the second group will appear in the

denominator; and the rest of $D-(r_k+s_k)$ parts are not involved in the logratio. As a result, the $\text{olr}(\mathbf{x})_k$ is

$$\text{olr}(\mathbf{x})_k = \sqrt{\frac{r_k \cdot s_k}{r_k+s_k}} \ln \frac{(x_{n1_k} \cdots x_{nr_k})^{1/r_k}}{(x_{d1} \cdots x_{ds_k})^{1/s_k}},\ k = 1,\ldots,D-1. \tag{2}$$

where $\sqrt{\frac{r_k \cdot s_k}{r_k+s_k}}$ is the factor for normalising the coordinate. Note that $r_1+s_1 = D$ for $k=1$. The $\text{olr}(\mathbf{x})_k$ coordinate, being a logcontrast that involves two groups of parts of a composition, informs, on average, about the relative importance of one group of parts with regard to the other.

3. Convexity on the simplex

Following the basic definitions of convexity on $\mathbb{R}^D$ (Boyd and Vandenberghe, 2004), the counterpart definitions of convexity on the simplex $\mathcal{S}^D$ in a consistent manner with the Aitchison geometry (i.e. $\mathcal{A}$-convexity) are:

Definition 1. Let $\mathbf{x}_1,\ \mathbf{x}_2$ be two D-part compositions. The $\mathcal{A}$-segment $\overline{\mathbf{x}_1\mathbf{x}_2}$ is the set

$$\overline{\mathbf{x}_1\mathbf{x}_2} = \{\mathbf{y} \in \mathcal{S}^D | \mathbf{y} = \lambda \odot \mathbf{x}_1 \oplus (1-\lambda) \odot \mathbf{x}_2,\ \lambda \in [0,1]\}. \tag{3}$$

An $\mathcal{A}$-segment can be expressed in olr-coordinates as $\text{olr}(\overline{\mathbf{x}_1\mathbf{x}_2}) = \{\mathbf{z} \in \mathbb{R}^{(D-1)} | \mathbf{z} = \lambda \cdot \text{olr}(\mathbf{x}_1) + (1-\lambda) \cdot \text{olr}(\mathbf{x}_2),\ \lambda \in [0,1]\}$. That is the typical expression of a segment of a line on the real space. The definition of a compositional segment (Eq. (3)) can be used in the definition of a compositional convex set ($\mathcal{A}$-convex set).

Definition 2. A set $\mathcal{B} \subseteq \mathcal{S}^D$ is an $\mathcal{A}$-convex set if for all $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{B}$, the compositional segment $\overline{\mathbf{x}_1\mathbf{x}_2}$ is contained in $\mathcal{B}$. That is, for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{B}$ and any $\lambda \in [0,1]$, it holds that $\lambda \odot \mathbf{x}_1 \oplus (1-\lambda) \odot \mathbf{x}_2 \in \mathcal{B}$.

To illustrate the definition of an $\mathcal{A}$-convex set, 3-part compositions were selected in $\mathcal{S}^3$ for creating a compositional triangle using the Definition (1) of an $\mathcal{A}$-segment (Fig.1(a)). By construction, a compositional triangle is an $\mathcal{A}$-convex set. On the upper right, Fig.1(b) shows a typical strip (blue area). A compositional segment (red line) not entirely contained in the set shows the lack of the $\mathcal{A}$-convexity of the strip. Figure 1(b) shows how sets that look like convex sets in the simplex from a typical Euclidean point of view, are not compositional convex sets with the Aitchison geometry, and vice versa (Fig.1(a)). Figures 1(c) and (d), respectively, show the representation of Figures 1 (a) and (b) in olr-coordinates. In these figures, one recognises the typical form of a triangle and the shape of a non-convex set on the real space.

3.1. Convex hull on the simplex

The simplex endowed with the induced Euclidean geometry does not have the same structure and properties as the simplex with the Aitchison geometry. This difference

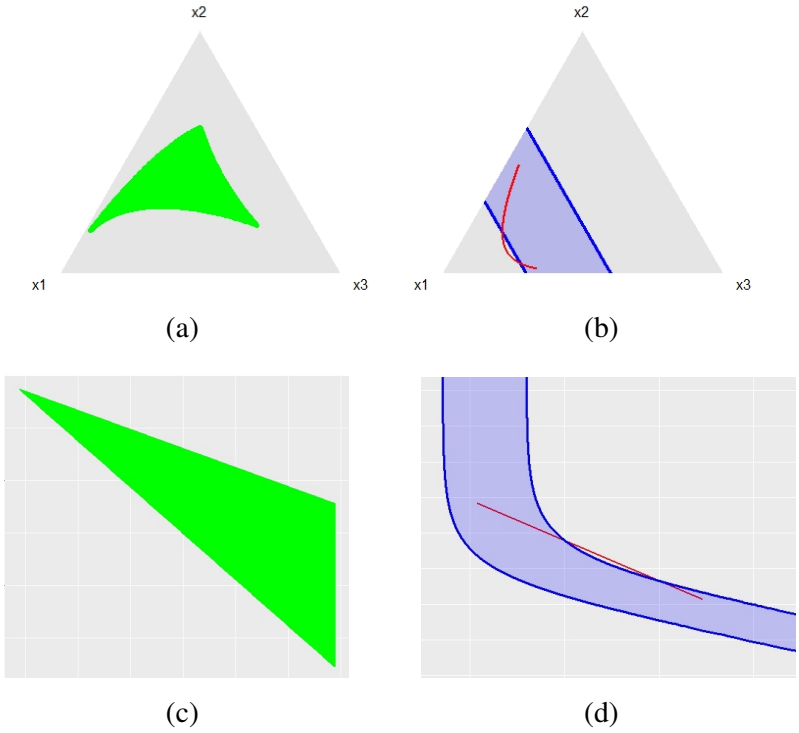


Figure 1. *Triangle and strip in $\mathcal{S}^3$: (a) $\mathcal{A}$-triangle (green area); (b) Strip (blue area) with an $\mathcal{A}$-segment (red line); (c) The $\mathcal{A}$-triangle (green area) in olr-coordinates ; and (d) The strip (blue area) with the $\mathcal{A}$-segment (red line) in olr-coordinates.*

between geometries has implications on the statistical techniques such as the peeling, which is a descriptive statistical technique based on the concept of convex hull (Caussinus, Ettinger and Tomassone, 2012; Small, 1990). Peeling can be described as an iterative algorithm that consists of removing layers of points. Each layer is formed by the points which form the border of the convex hull of the set of remaining points. The convex hull of a set of points $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ in $\mathbb{R}^D$ is the set

$$conv_E(\mathbf{X}) = \{\mathbf{z}\in\mathbb{R}^D | \mathbf{z} = \lambda_1\mathbf{x}_1 + \ldots + \lambda_n\mathbf{x}_n, \sum_{i=1}^{n}\lambda_i = 1,\ \lambda_i \geq 0, i = 1,\ldots,n\},$$

This definition is used by Tolosana-Delgado, von Eynatten and Karius (2011) for CoDa vectors. Among other applications, peeling allows us to graphically represent the centre of a set of points in the last *internal* layer and can be used for outlier detection (Harsh, Ball and Wei, 2016, chapter 4).

We propose the corresponding definition of the convex hull on the simplex in terms of Aitchison geometry:

Definition 3. The $\mathcal{A}$-convex hull of the set of compositions $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{S}^D$ is

$$conv_{\mathcal{A}}(\mathbf{X}) = \{\mathbf{y} \in \mathcal{S}^D | \mathbf{y} = \lambda_1 \odot \mathbf{x}_1 \oplus \ldots \oplus \lambda_n \odot \mathbf{x}_n, \sum_{i=1}^{n} \lambda_i = 1,\ \lambda_i \geq 0, i = 1, \ldots, n\}.$$

Note that the compositional triangle created in Fig. 1(a) is the most simple example of an $\mathcal{A}$-convex hull using the minimum number of compositions. The usefulness of this concept can be illustrated by means of a more complex example in $\mathcal{S}^3$. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{20}\} \in \mathcal{S}^3$ be a set of compositions randomly generated using a normal distribution on the simplex (Mateu-Figueras, Pawlowsky-Glahn and Egozcue, 2013). Figure 2 compares the $\mathcal{A}$-convex hull $conv_{\mathcal{A}}(\mathbf{X})$ (on the left) to the Euclidean convex hull $conv_E(\mathbf{X})$ (on the right). Figure 2(a) shows the successive layers created when peeling is applied to $\mathbf{X}$ using the $\mathcal{A}$-convex hull. The last internal layer (i.e., the smallest convex hull) is close to the compositional centre of $\mathbf{X}$, $g(\mathbf{X}) = \frac{1}{20} \odot (\oplus_{i=1}^{20} \mathbf{x}_i)$ (green dot), the column-wise geometric mean (e.g., Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015). On the other hand, the typical Euclidean centre of $\mathbf{X}$, the arithmetical mean $\overline{\mathbf{X}} = \frac{1}{20} \sum_{i=1}^{20} \mathbf{x}_i$ (red dot) is far from the last *external* layer of the peeling, suggesting a potential outlier. Fig. 2(b) shows the layers of the peeling using the E-convex hull. In this case, the Euclidean centre of $\mathbf{X}$ (red dot) is inside the last layer, whereas the compositional centre $g(\mathbf{X})$ (green dot) is far from it. In addition, when comparing the first external layers (i.e., the largest convex hull) in both geometries, one concludes that the $\mathcal{A}$-convex hull fits better with the common arch shape of a normally distributed CoDa set.

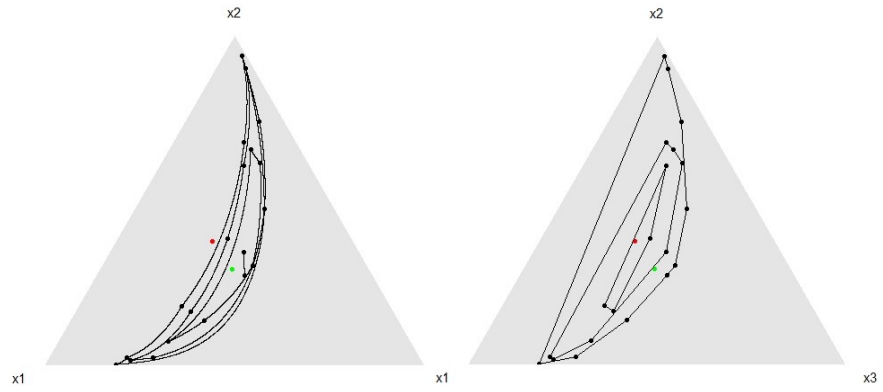


Figure 2. *Peeling applied to a CoDa set $\mathbf{X} \in \mathcal{S}^3$: (a) with $\mathcal{A}$-convex hull; (b) with E-convex hull. Geometric centre $g(\mathbf{X})$ (green dot) and arithmetic centre $\overline{\mathbf{X}}$ (red dot) of CoDa set are plotted.*

3.2. Some basic compositional sets

The most common sets in convex optimisation in areas such as, among others, design of experiments (DOE), optimisation with mixtures, or time-use data are defined by con-

straining some parts of the composition $\mathbf{x}$: $\{0 < l_i \leq x_i \leq u_i < 1;\ i = 1,\ldots,D\}$ (Chen et al., 2010) or by constraining the ratio of two parts: $\{0 < l_{ij} \leq \frac{x_i}{x_j} \leq u_{ij};\ i \neq j = 1,\ldots,D\}$(Lo Huang and Huang, 2009).

3.2.1. Constraining the ratios between parts: a logcontrast

The equation $\{\frac{x_i}{x_j} = k\ ,1 \leq i \neq j \leq D;\ k > 0\}$ is equivalent to the logcontrast $\{\ln x_i - \ln x_j = \ln k;\ 1 \leq i \neq j \leq D\}$. This equation describes an affine subspace of dimension $D-2$ on the simplex with the Aitchison geometry (Egozcue, Pawlowsky-Glahn and Gloor, 2018). Consequently, $l \leq \frac{x_i}{x_j}$ or $\frac{x_i}{x_j} \leq u$ define closed half-spaces of the simplex, and both sets, $\Pi^+ = \{\mathbf{x} \in \mathcal{S}^D | l \leq \frac{x_i}{x_j}\}$ and $\Pi^- = \{\mathbf{x} \in \mathcal{S}^D | \frac{x_i}{x_j} \leq u\}$, verify the condition of being an $\mathcal{A}$-convex set (Definition 2).

In general, affine subspaces such as $\mathcal{A}$-lines, $\mathcal{A}$-planes or $\mathcal{A}$-hyperplanes are defined by logcontrasts (Eq. (1)) (Egozcue et al., 2018). A logcontrast splits the simplex into two closed half-spaces, $\Pi^+ = \{\mathbf{x} \in \mathcal{S}^D | \sum_{j=1}^D \beta_j \ln x_j \geq k\}$ and $\Pi^- = \{\mathbf{x} \in \mathcal{S}^D | \sum_{j=1}^D \beta_j \ln x_j \leq k\}$. By construction, both half-spaces are $\mathcal{A}$-convex sets (Definition 2). Figure 3 shows four different logcontrasts (blue lines) whose intersection determines a quadrilateral (green area). Because the intersection of convex sets is a convex set (Boyd and Vandenberghe, 2004) it follows that the quadrilateral is an $\mathcal{A}$-convex set.

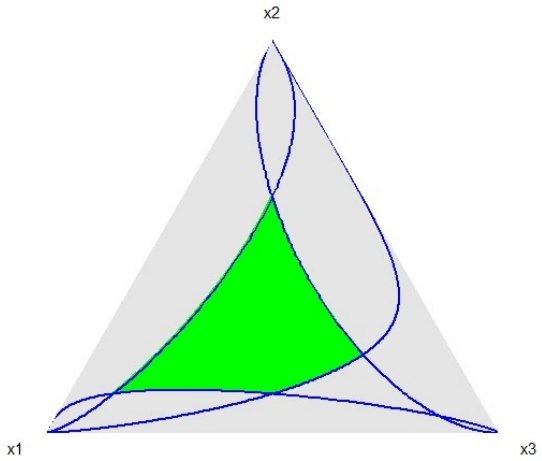


Figure 3. *An $\mathcal{A}$-convex quadrilateral in $\mathcal{S}^3$ (green area) determined by the intersection of four half-spaces defined by logcontrasts (blue lines).*

3.2.2. Constraining the parts of a composition

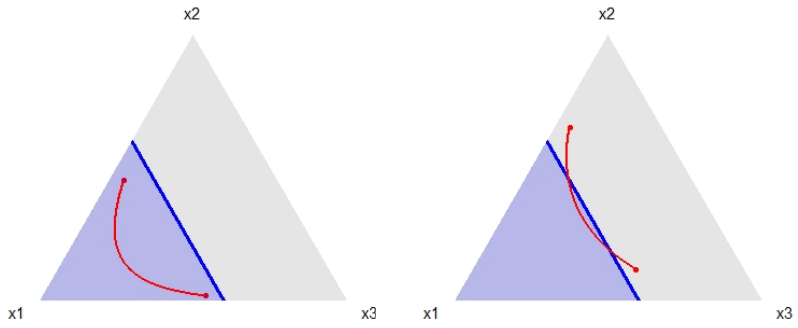
Despite the fact that an equation such as $\{x_i = k,\ k \in (0,1)\}$, for any $i = 1,\ldots,D$, cannot be expressed in terms of a logcontrast, this type of equation also splits the simplex into two sets: the upper set $\Sigma^+ = \{\mathbf{x} \in \mathcal{S}^D | x_i \geq k,\ k \in (0,1)\}$, and the lower set $\Sigma^- = \{\mathbf{x} \in$

$\mathcal{S}^D|x_i \leq k,\ k \in (0,1)\}$. However, in this case, the two sets are different with regard to their compositional convexity.

Proposition 1. *For any* $i = 1,\ldots,D$, *the set* $\Sigma^+ = \{\mathbf{x} \in \mathcal{S}^D|x_i \geq k,\ k \in (0,1)\}$ *is an* $\mathcal{A}$*-convex set.*

Proof. See Appendix. ■

In contrast, the lower set Σ^- is not an $\mathcal{A}$-convex set as is illustrated in Fig. 4(b). Figure 4 shows an example in S^3 of the type Σ^+ (blue area) and its complementary set, Σ^- (grey area), for part x_1 with $k = 0.4$. For each set, two compositions were selected and the corresponding $\mathcal{A}$-segment plotted (red line). Because the red $\mathcal{A}$-segment on Fig. 4(b) is not entirely contained in the set Σ^- (grey area), this set is not an $\mathcal{A}$-convex set.



(a) (b)

Figure 4. *A 3-part composition constrained for* $i = 1$ *and* $k = 0.4$*: (a) The set* Σ^+ *(blue area); (b) The set* Σ^- *(grey area). The red line represents the* $\mathcal{A}$*-segment for two compositions in each set.*

The sets Σ^+ and Σ^- can be generalised as follows:

$$\Sigma_i^+(\boldsymbol{\alpha}) = \{\mathbf{x} \in \mathcal{S}^D|\alpha_1 x_1 + \ldots + \alpha_D x_D \geq 0, \alpha_i > 0, \alpha_j \leq 0, 1 \leq j \neq i \leq D\}$$

$$\Sigma_i^-(\boldsymbol{\alpha}) = \{\mathbf{x} \in \mathcal{S}^D|\alpha_1 x_1 + \ldots + \alpha_D x_D \leq 0, \alpha_i > 0, \alpha_j \leq 0, 1 \leq j \neq i \leq D\}$$

Note that $\Sigma_i^+(\boldsymbol{\alpha})$ and $\Sigma_i^-(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha} = (-k,\ldots,-k,\underbrace{(1-k)}_{i},-k,\ldots,-k),\ k \in (0,1)$, become the sets Σ^+ and Σ^-, respectively. Importantly, an analogous proof to Proposition 1 states that a set $\Sigma_i^+(\boldsymbol{\alpha})$ is $\mathcal{A}$-convex. On the other hand, $\Sigma_i^-(\boldsymbol{\alpha})$ is not an $\mathcal{A}$-convex set. To illustrate these properties, Figure 5 shows the set $\Sigma_1^+(\boldsymbol{\alpha}) = \{\mathbf{x} \in \mathcal{S}^4|x_1 - \frac{2}{3}x_2 - \frac{2}{3}x_3 - \frac{1}{3}x_4 \geq 0\}$ (blue area) in $\mathcal{S}^4$, which generalises a set of type $\Sigma^+ = \{\mathbf{x} \in \mathcal{S}^D|x_1 \geq k,\ k \in (0,1)\}$.

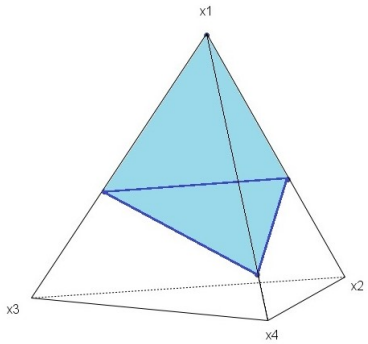


Figure 5. *The $\mathcal{A}$-convex set $\Sigma_1^+(\alpha) = \{\mathbf{x} \in \mathcal{S}^4 | x_1 - \frac{2}{3}x_2 - \frac{2}{3}x_3 - \frac{1}{3}x_4 \geq 0\}$ (blue area) in $\mathcal{S}^4$.*

Note that the intersection of sets of type Σ^+ or $\Sigma^+(\alpha)$ is an $\mathcal{A}$-convex set. However, when an optimisation problem includes a set of type $\Sigma^-(\alpha)$ the feasible region might be a non $\mathcal{A}$-convex set. In such a case, one should be cautious when applying convex optimisation techniques.

4. Convex functions on the simplex

The definition of a convex function in the Euclidean space can be adapted to the Aitchison geometry following the schema introduced in Luenberger and Ye (2008) and in Boyd and Vandenberghe (2004).

Definition 4. Let $W \subset \mathcal{S}^D$ be an $\mathcal{A}$-convex set. A function $f : W \to \mathbb{R}$ is an $\mathcal{A}$-convex function if for all $\mathbf{x}_1,\ \mathbf{x}_2 \in W$ and $\lambda \in [0,\ 1]$:

$$f((1-\lambda) \odot \mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2) \leq (1-\lambda) f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2)$$

Definition 5. Let $W \subset \mathcal{S}^D$ be an $\mathcal{A}$-convex set. A function $g : W \to \mathbb{R}$ is an $\mathcal{A}$-concave function if $f = -g$ is an $\mathcal{A}$-convex function.

Note that, as expected, a constant function $f(\mathbf{x}) = k$, with k a real number, is simultaneously an $\mathcal{A}$-convex and $\mathcal{A}$-concave function.

Importantly, the classification of convex function through the gradient or the Hessian matrix also applies for CoDa. That is, the common rule *A twice differentiable function of several variables is convex on a convex set if and only if its Hessian matrix of second partial derivatives is positive semidefinite on the interior of the convex set* can be used to classify $\mathcal{A}$-convex functions by means of the basic concepts of compositional differential calculus (Barceló-Vidal, Martín-Fernández and Mateu-Figueras, 2011) working in olr-coordinates. Whereas, using the $\mathcal{A}$-gradient, one has to check the usual expression

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla_A(f)(\mathbf{y})(\ln(\mathbf{x}) - \ln(\mathbf{y}))$$

for all $\mathbf{x},\ \mathbf{y} \in W$, where $\nabla_{\mathcal{A}}(f)$ is the $\mathcal{A}$-gradient vector in clr-coordinates Barceló-Vidal et al. (2011).

The following two propositions show that the sum of functions and product by a scalar are basic operations for creating more complex $\mathcal{A}$-convex functions.

Proposition 2. *Let f_1 and f_2 be two $\mathcal{A}$-convex functions on the $\mathcal{A}$-convex set $W \subset \mathcal{S}^D$. The function $f_1 + f_2$ is $\mathcal{A}$-convex on W.*

Proof. See Appendix. ■

Proposition 3. *Let f be an $\mathcal{A}$-convex function on the $\mathcal{A}$-convex set $W \subset \mathcal{S}^D$. For any $a \geq 0$, the function af is $\mathcal{A}$-convex on W*

Proof. This proof is immediate from the definition of convex function. ■

Using the two previous propositions, it follows that a positive linear combination of $\mathcal{A}$-convex functions is an $\mathcal{A}$-convex function. That is, given a set of $\mathcal{A}$-convex functions $f_1, ..., f_n$ on the $\mathcal{A}$-convex set $W \subset \mathcal{S}^D$, then the function $a_1 f_1 + \ldots + a_n f_n$ is an $\mathcal{A}$-convex function for any $a_j > 0,\ j = 1, \ldots, n$. This property is useful for creating more complex $\mathcal{A}$-convex functions using basic functions (see Section 4.1).

In addition, the following results state that the sublevel sets of an $\mathcal{A}$-convex function on the simplex verify the usual properties as regard to the convex sets.

Proposition 4. *Let f be an $\mathcal{A}$-convex function on the $\mathcal{A}$-convex set $W \subset \mathcal{S}^D$. The sublevel set $\mathcal{G}_\alpha^- = \{\mathbf{x} |\ \mathbf{x} \in W, f(\mathbf{x}) \leq \alpha\}$ is $\mathcal{A}$-convex for any real number α.*

Proof. See Appendix. ■

Even though for any $\mathcal{A}$-convex function, its sublevel sets $\mathcal{G}_\alpha^-$ are $\mathcal{A}$-convex sets, the converse is not true. This fact motivates the following definitions.

Definition 6. A function $f : \mathcal{S}^D \to \mathbb{R}$ is an $\mathcal{A}$-quasiconvex function if its domain and all its sublevel sets, $\mathcal{G}_\alpha^- = \{\mathbf{x} |\ \mathbf{x} \in dom(f), f(\mathbf{x}) \leq \alpha\}$ are $\mathcal{A}$-convex sets for any real number α.

Definition 7. A function $f : \mathcal{S}^D \to \mathbb{R}$ is an $\mathcal{A}$-quasiconcave function if its domain and all its superlevel sets, $\mathcal{G}_\alpha^+ = \{\mathbf{x} |\ \mathbf{x} \in dom(f), f(\mathbf{x}) \geq \alpha\}$ are $\mathcal{A}$-convex sets for any real number α.

As with convex optimisation problems in the Euclidean space when defining the feasible region (Boyd and Vandenberghe, 2004), it is recommended to represent the sublevels of an $\mathcal{A}$-quasiconvex function (or the superlevels of a $\mathcal{A}$-quasiconcave function) through inequalities of $\mathcal{A}$-convex functions. Therefore, an $\mathcal{A}$-quasiconvex function f, should be expressed by means of $\mathcal{A}$-convex functions, Φ_α such that,

$$f(\mathbf{x}) \leq \alpha \iff \Phi_\alpha(\mathbf{x}) \leq 0.$$

4.1. Some basic functions on the simplex

With the following examples, the $\mathcal{A}$-convexity of some popular functions on $\mathcal{S}^D$ is reviewed.

Example 1. The function $f(\mathbf{x}) = \frac{x_i}{x_j}$, $1 \leq i, j \leq D$ is an $\mathcal{A}$-convex function over its domain, $dom(f) = \mathcal{S}^D$.

Proof. See Appendix. ■

Example 2. For any $i = 1, \ldots, D$, the function $f(\mathbf{x}) = x_i$ is an $\mathcal{A}$-quasiconcave function over its domain, $dom(f) = \mathcal{S}^D$. Moreover, using the $\mathcal{A}$-convex function

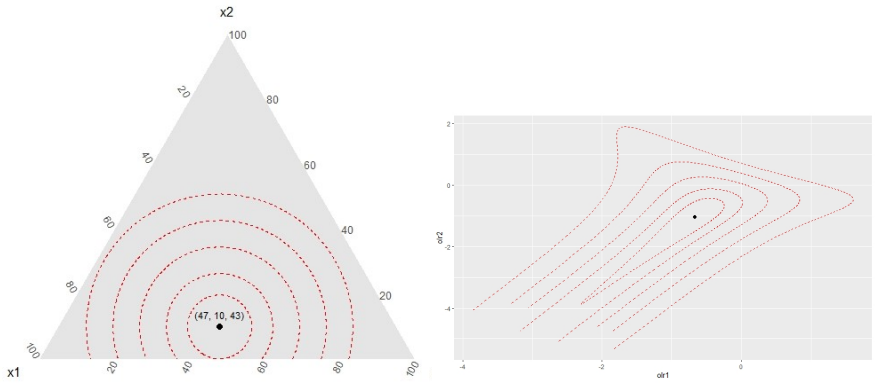
$$\Phi_\alpha(\mathbf{x}) = \alpha \sum_{j=1}^{D} \frac{x_j}{x_i} - 1\,,$$

a superlevel $\mathcal{G}_\alpha^+ = \{\mathbf{x} \in \mathcal{S}^D | \; x_i \geq \alpha\}$ for any $\alpha \in (0,\ 1)$ can be represented by means of $\Phi_\alpha(\mathbf{x}) \leq 0$.

Proof. See Appendix. ■

Example 3. For any $\mathbf{x}_0 \in \mathcal{S}^D$, the function squared Euclidean distance $f(\mathbf{x}) = d_E^2(\mathbf{x}, \mathbf{x}_0) = \sum_{j=1}^{D}(x_j - x_{0j})^2$ is not $\mathcal{A}$-convex.

Figure 6(a) shows the contour lines of the function $f(\mathbf{x}) = d_E^2(\mathbf{x}, \mathbf{x}_0)$ on $\mathcal{S}^3$ for $\mathbf{x}_0 = (47, 10, 43)$. The sublevel sets are not $\mathcal{A}$-convex sets, and therefore, the function $f(\mathbf{x})$ is not an $\mathcal{A}$-convex function (Proposition 4). Because it may be difficult to see the lack of convexity of a set on the ternary diagram, Figure 6(b) shows the sublevel sets in olr-coordinates.



(a) (b)

Figure 6. *Contour lines of the function $f(\mathbf{x}) = d_E^2(\mathbf{x}, \mathbf{x}_0)$ on $\mathcal{S}^3$ for $\mathbf{x}_0 = (47, 10, 43)$: (a) Ternary diagram; (b) olr-coordinates.*

Note that the function Euclidean distance and the function (squared) $d_E^2(\mathbf{x},\mathbf{x}_0)$ share the same contour lines. Consequently, the function (non-squared) $d_E(\mathbf{x},\mathbf{x}_0)$ is not $\mathcal{A}$-convex. The lack of convexity means that the Euclidean distance does not satisfy the triangular inequality, that is, it is not a distance function on the simplex endowed with the Aitchison geometry.

Example 4. For any $\mathbf{x}_0 \in \mathcal{S}^D$, the function squared Aitchison distance $f(\mathbf{x}) = d_{\mathcal{A}}^2(\mathbf{x},\mathbf{x}_0) = d_E^2(\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{x}_0))$ is $\mathcal{A}$-convex.

Proof. The proof is immediate because the function squared Euclidean distance is convex on the Real space. ■

4.2. Convex optimisation on the simplex

In a broad sense, a convex optimisation problem in $\mathbb{R}^D$ is defined as (Boyd and Vandenberghe, 2004):

$$\begin{aligned} &\textit{minimise} && f_0(\mathbf{x}) \\ &\textit{subject to} && f_j(\mathbf{x}) \leq 0 \;\; j = 1,\ldots,m \\ &&& g_k(\mathbf{x}) = b_k \; k = 1,\ldots,n \end{aligned} \tag{4}$$

where $\mathbf{x} \in \mathbb{R}^D$, $f_0,\ldots,f_m : \mathbb{R}^D \to \mathbb{R}$ are convex functions and $g_1,\ldots,g_n : \mathbb{R}^D \to \mathbb{R}$ are linear functions.

For CoDa, the above definition is adapted as:

Definition 8. An $\mathcal{A}$-convex optimisation problem in standard form is defined as

$$\begin{aligned} \text{minimise} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_j(\mathbf{x}) \leq 0 && j = 1,\ldots,m \\ & \boldsymbol{\beta}_k^{\mathsf{T}} \ln \mathbf{x} = b_k && k = 1,\ldots,n \end{aligned}$$

where $\mathbf{x} \in \mathcal{S}^D$, $f_0,\ f_1,\ldots,\ f_m$ are $\mathcal{A}$-convex functions and $\boldsymbol{\beta}_k^{\mathsf{T}} \ln \mathbf{x}$ are logcontrasts, that is, $\sum_{j=1}^D \beta_{k,j} = 0,\ k = 1,\ldots,n$.

An important case of convex optimisation problems is that of *linear programming*, that is, when the objective and all constraining functions are linear. For CoDa, an $\mathcal{A}$-linear programming problem is defined in terms of logcontrasts as follows:

$$\begin{aligned} \text{minimise} \quad & \boldsymbol{\beta}_0^{\mathsf{T}} \ln \mathbf{x} \\ \text{subject to} \quad & \boldsymbol{\beta}_k^{\mathsf{T}} \ln \mathbf{x} \leq b_k && k = 1,\ldots,m \\ & \boldsymbol{\beta}_k^{\mathsf{T}} \ln \mathbf{x} = b_k && k = m+1,\ldots,n \end{aligned}$$

where $\sum_{j=1}^D \beta_{k,j} = 0,\ k = 0,\ldots,n$.

5. Case Study

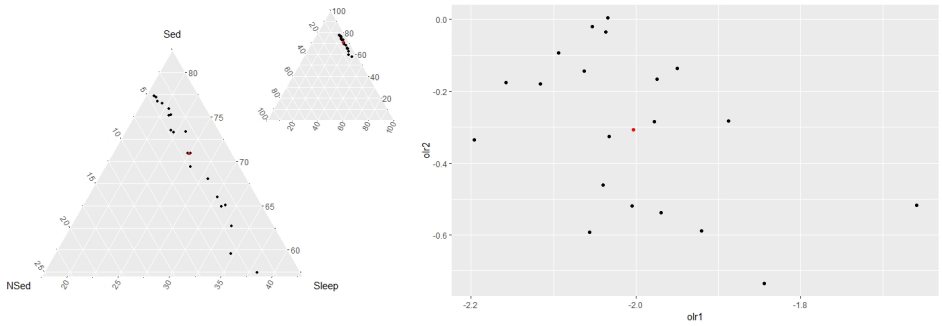
The example we present here is based on data from Aitchison (1986) and only serves for illustrative purposes. We consider a 3-part time-use composition of mutually exclusive and exhaustive parts: non-sedentary time (NSed), sedentary time (Sed), and sleeping time (Sleep). Table 1 shows the 3-part time-use compositions of a university associate professor with unhealthy physical activity habits. Figure 7 shows the data set in the ternary diagram (Fig. 7a) and in the olr-space (Fig. 7b), where the olr-basis used is $\mathrm{olr}_1(\mathbf{x}) = \frac{\sqrt{2}}{2} \ln \frac{NSed}{Sed}$ and $\mathrm{olr}_2(\mathbf{x}) = \sqrt{\frac{2}{3}} \ln \frac{\sqrt{NSed \cdot Sed}}{Sleep}$.

Table 1. *3-part time-use composition over 20 days.*

	Non-sedentary	Sedentary	Sleep
D1	0.04234	0.77218	0.18547
D2	0.03772	0.75235	0.20993
D3	0.04807	0.69388	0.25805
D4	0.05705	0.59596	0.34699
D5	0.04306	0.76733	0.18961
D6	0.03592	0.75916	0.20493
D7	0.03797	0.67973	0.28231
D8	0.03959	0.76519	0.19522
D9	0.04321	0.70868	0.24811
D10	0.04000	0.70886	0.25114
D11	0.04060	0.75101	0.20838
D12	0.04148	0.62683	0.33169
D13	0.04003	0.64864	0.31133
D14	0.04357	0.77365	0.18277
D15	0.04488	0.73273	0.22239
D16	0.04665	0.73483	0.21853
D17	0.03873	0.65937	0.30190
D18	0.03282	0.73313	0.23405
D19	0.03552	0.65058	0.31390
D20	0.04231	0.57445	0.38324

The centre or mean ($\mathbf{x}_0$) of a CoDa set is the vector of geometric means of its parts, scaled to sum 1 in order to obtain its representative on the unit simplex. Therefore, on dayly average, the associate professor engages in physical activity for one hour, exhibits sedentary behaviour for seventeen hours, and sleeps for six hours, $\mathbf{x}_0 = (1/24, 17/24, 6/24)$ (see the red point in Figure 7).

The location and spread of a compositional data set are summarised in the variation array (Table 2) through pairwise logratios of parts. The elements above the first diagonal are the pairwise log-ratio variances, whereas the elements below it are the arithmetic



(a) (b)

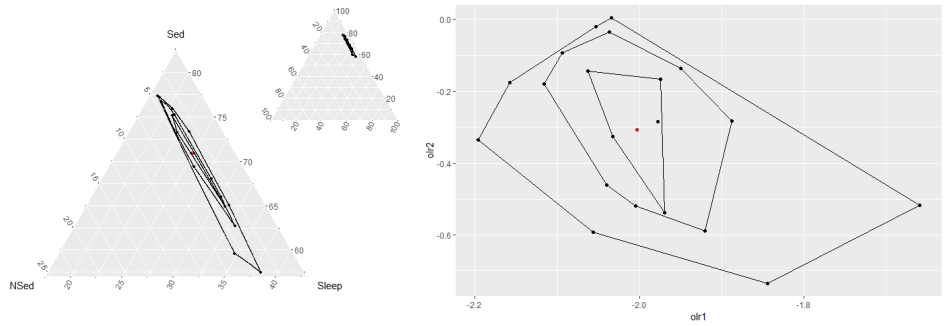
Figure 7. *(a) 3-part time-use compositions on the ternary diagram (small triangle). The large triangle zooms in on the data set region. In red, the centre of the data set* $\mathbf{x}_0 = (\frac{1}{24}, \frac{17}{24}, \frac{6}{24})$*; (b) the data set and its centre in the olr-space.*

means. As suggested by the ternary diagram (Fig. 7a), the largest log-ratio variance corresponds to $\{Sed, Sleep\}$, whereas the smallest value is $Var(\ln \frac{NSed}{Sed}) = 0.0265$. In this case, because the estimate of the log-ratio expectation is $E(\ln \frac{Sed}{NSed}) = 2.8332$, on average, sedentary time is approximately 17 times ($\approx \exp\{2.8332\}$) the non-sedentary time, as the centre $\mathbf{x}_0$ of the data set indicates.

Table 2. *Variation array of the 3-part time-use compositional data.*

	Pairwise log-ratio variance		
	NSed	Sed	Sleep
NSed		0.0265	0.0561
Sed	2.8332		0.0949
Sleep	1.7918	-1.0415	
	Pairwise log-ratio arithmetical mean		

The data set shown in the ternary diagram (Fig. 7a) suggests that the farthest point from the centre is the point located further down on the simplex: $D20 = (0.04231,\ 0.57445,\ 0.38324)$, with the smallest value in the part *NSed* (Table 1). At first glance, this point may be considered a potential outlier. Moreover, when representing the data set in olr-coordinates (Fig. 7b), the outermost point is the point located further to the right: $\text{olr}(D4) = (-1.659,\ -0.516)$. This fact is corroborated by the peeling using the $\mathcal{A}$-convex hull (3). Figure 8b suggests that each layer has an elliptical shape, the typical shape of a normal probability distribution. A multivariate Anderson-Darling test confirms that one can fail to reject normality ($A^2 = 0.7760$; $p - value = 0.2199$). Indeed, under this assumption, the χ^2 atipicality index indicates that the sample $D4$ is a potential outlier ($\chi^2 percentile = 98.75\%$), whereas the sample $D20$ is not ($\chi^2 percentile = 86.87\%$).



(a) (b)

Figure 8. *Peeling with the Aitchison geometry of the 3-part time-use compositions: (a) on the ternary diagram (small triangle). The large triangle zooms into the data set region. The centre of the data set is represented in red colour; (b) in the olr-space.*

The professor was recommended to increase non-sedentary time up to, on average, at least 13 hours per day. The question is how to distribute the time for the rest of the activities (i.e., sedentary and sleeping). One criterion may be to move from the centre of the data set ($\mathbf{x}_0$) to the closest point in the $\mathcal{A}$-convex set $\Sigma^+ = \{\mathbf{x} \in \mathcal{S}^D | x_1 \geq \frac{13}{24}\}$, whereas the spread is preserved. This is one of the simplest examples of convex optimisation problem: to find the minimum distance from a given point $\mathbf{x}_0 \in \mathcal{S}^D$ to a convex set Σ^+.

From a Euclidean approach, the E-convex optimisation problem is:

$$\begin{array}{ll} \text{minimise} & d_E^2(\mathbf{x}_0, \mathbf{x}) = \sum_{j=1}^3 (x_j - x_{0j})^2 \\ \text{subject to} & x_1 \geq 13/24 \\ & x_1 + x_2 + x_3 = 1 \\ & x_j \geq 0,\ j = 1, \ldots, 3 \end{array} \tag{5}$$

Figure 9(a) shows that the solution of the E-convex optimisation problem (Eq. 5) is $\mathbf{x} = (\frac{13}{24}, \frac{11}{24}, 0)$, where the proposed movement is $\mathbf{x} - \mathbf{x}_0 = (12/24, -6/24, -6/24)$. That is, in order to increase the fraction of non-sedentary time in $\frac{12}{24}$, the Euclidean approach subtracts from the rest of the parts (non-sedentary and sleeping time) the same amount of time, $\frac{6}{24}$ hours. Note that the sleeping time has to be zero, a solution that is not realistic in practice.

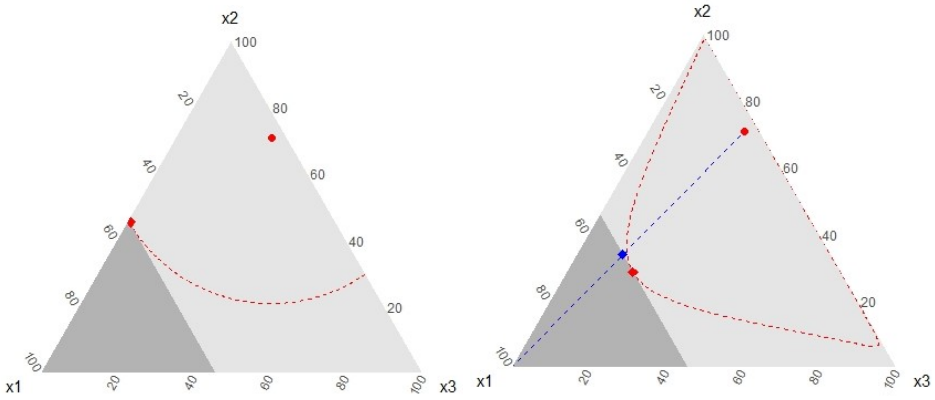
In this situation, one could consider applying an analogous procedure using the Aitchison geometry. That is, perturb by $\frac{13}{1}$ the non-sedentary time $\frac{1}{24}$ to verify that $x_1 = 13/24$, while perturbing the other two parts ($\{Sed, Sleep\}$) by the same factor to preserve its relative information. Following this idea, when the centre $\mathbf{x}_0 = (1/24, 17/24, 6/24)$ is perturbed by the vector $\mathbf{p} = (13/1, 11/23, 11/23)$ the composition obtained is $\mathbf{x} = (13/24, 8.13/24, 2.87/24)$, which verifies the constraint $x_1 \geq 13/24$ (see figure 9 (b)). When one calculates the (squared) Aitchison distance from $\mathbf{x}_0$ to the new centre $\mathbf{x}$ the result is 7.271. To confirm whether this distance is the minimum value one must formulate the convex optimisation problem using the Aitchison geometry:

$$\begin{array}{ll} \text{minimise} & d^2_{\mathcal{A}}(\mathbf{x}_0,\mathbf{x}) = \sum_{j=1}^{3}\left(\ln\frac{x_j}{g(\mathbf{x})} - \ln\frac{x_{0j}}{g(\mathbf{x_0})}\right)^2 \\ \text{subject to} & x_1 \geq \frac{13}{24} \end{array} \tag{6}$$

whose standard form is (Example 2)

$$\begin{array}{ll} \text{minimise} & d^2_{\mathcal{A}}(\mathbf{x}_0,\mathbf{x}) = \sum_{j=1}^{3}\left(\ln\frac{x_j}{g(\mathbf{x})} - \ln\frac{x_{0j}}{g(\mathbf{x_0})}\right)^2 \\ \text{subject to} & \frac{13}{24}\sum_{j=1}^{3}\frac{x_j}{x_1} - 1 \leq 0 \end{array}$$

Figure 9(b) shows that the solution to the $\mathcal{A}$-convex optimisation problem (Eq. 6) is $\mathbf{x} = \left(\frac{13}{24}, \frac{6.87}{24}, \frac{4.13}{24}\right)$. Note that, the Aitchison approach has a reasonable behaviour because the largest part of the initial composition (i.e., sedentary time) contributes more to increase the non-sedentary time. The way that the other parts contribute to increase the non-sedentary time is not proportional in any sense. In this case, the (squared) Aitchison distance from $\mathbf{x}_0$ to the new centre $\mathbf{x}$ is 6.989, smaller than the distance obtained when the new centre is created by perturbation. Importantly, the ratio $\frac{Sed}{Sleep}$ is not preserved when moving from $\mathbf{x}_0$ to the solution of the optimisation problem ($\mathbf{x}$). That is, using this approach, the parts $\{Sed,\ Sleep\}$ are not perturbed by the same factor.



(a) (b)

Figure 9. *Time-use optimisation for the composition* $\mathbf{x}_0 = (\frac{1}{24}, \frac{17}{24}, \frac{6}{24})$ *(red dot). The dark grey area is the set* $\Sigma^+ = \{\mathbf{x} \in \mathcal{S}^3 | x_1 \geq \frac{13}{24}\}$. *The red diamond* $\mathbf{x} \in \Sigma^+$ *is the closest point to* $\mathbf{x}_0$. *The dashed red line is the contour line for the corresponding distance. (a) Euclidean geometry:* $\mathbf{x} = (\frac{13}{24}, \frac{11}{24}, \frac{0}{24})$*; (b) Aitchison geometry:* $\mathbf{x} = \left(\frac{13}{24}, \frac{6.87}{24}, \frac{4.13}{24}\right)$. *Dashed blue line is the perturbation direction from* $\mathbf{x}_0$ *to* $\mathbf{x} = (\frac{13}{24}, \frac{8.13}{24}, \frac{2.87}{24}) \in \Sigma^+$ *(blue diamond)*

In the optimisation problem, the movement from $\mathbf{x}_0$ to $\mathbf{x}$ is explained by the perturbation difference $\mathbf{x} \ominus \mathbf{x}_0 = (13/1, 6.87/17, 4.13/6)$. Figure 10 shows how the original data set (grey) is moved to the data set perturbed (black) by $\mathbf{x} \ominus \mathbf{x}_0 = (13/1, 6.87/17, 4.13/6)$, preserving the data spread. The centre of the perturbed data set fulfils the condition $x_1 \geq \frac{13}{24}$. With the Aitchison geometry approach, the solution is more realistic.

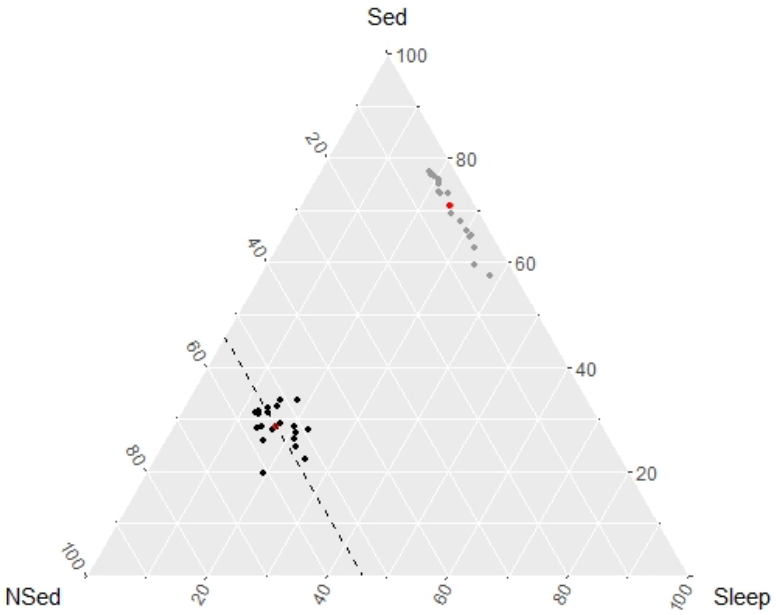


Figure 10. *Resolving the optimisation problem on the ternary diagram: the original data set (grey) is perturbed by* $\mathbf{x} \ominus \mathbf{x}_0 = (13/1, 6.87/17, 4.13/6)$ *for becoming the new data set (black). The dashed segment is the border of the set* $x_1 \geq \frac{13}{24}$*. The centre of each data set is shown in red.*

6. Conclusions and final remarks

The adequate definitions of convex set, convex hull and convex function in a compatible manner with the Aitchison geometry play an important role in the analysis of CoDa. The most popular statistical techniques include optimisation problems that are sensitive to the geometry of the sample space. We have compared the Euclidean geometry to the Aitchison geometry when solving a convex optimisation problem for CoDa. While the Euclidean approach found inconsistent solutions, the Aitchison geometry provided more satisfactory results. This concludes that, despite the fact that any method may give a reasonable solution in some scenarios, it is advisable to use methods that are consistent with the geometry of the sample space.

With basic concepts of constrained convex optimisation for CoDa on hand, a revision of some statistical techniques should be carried out. The pending challenge is to investigate the implications in popular techniques such as, among others, outlier detection, experimental design of mixtures, or Lasso regression.

Appendix: Proofs

Proposition 1 *For any $i = 1,\ldots,D$, the set $\Sigma^+ = \{\mathbf{x} \in \mathcal{S}^D | x_i \geq k,\ k \in (0,1)\}$ is an $\mathcal{A}$-convex set.*

Proof. Let $\mathbf{x}_1 = (x_{11},\ldots,x_{1D})$, $\mathbf{x}_2 = (x_{21},\ldots,x_{2D})$ be two D-part compositions in Σ^+. We have to check that for any $\lambda \in [0,\ 1]$ it holds that

$$(1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2 \in \Sigma^+.$$

Due to $\sum_{j=1}^D x_{1j} = \sum_{j=1}^D x_{2j} = 1$ then the equations $x_{1i} \geq k$ and $x_{2i} \geq k$ are respectively equivalent to

$$x_{1i} \geq k(x_{11} + \ldots + x_{1D}) \quad \text{and} \quad x_{2i} \geq k(x_{21} + \ldots + x_{2D}).$$

Using this equivalence, $\forall \lambda \in [0,\ 1]$ it holds

$$x_{1i}^{1-\lambda} x_{2i}^{\lambda} \geq k(x_{11} + \ldots + x_{1D})^{1-\lambda}(x_{21} + \ldots + x_{2D})^{\lambda} = k(\sum_{j=1}^{D} x_{1j})^{1-\lambda}(\sum_{j=1}^{D} x_{2j})^{\lambda},$$

where, with Hölder's inequality, the last term of the expression holds

$$k(\sum_{j=1}^{D} x_{1j})^{1-\lambda}(\sum_{j=1}^{D} x_{2j})^{\lambda} \geq k\sum_{j=1}^{D} x_{1j}^{1-\lambda} x_{2j}^{\lambda}.$$

That is, it holds that

$$x_{1i}^{1-\lambda} x_{2i}^{\lambda} \geq k\sum_{j=1}^{D} x_{1j}^{1-\lambda} x_{2j}^{\lambda}. \tag{7}$$

Finally, writing Eq. (7) in terms of the i^{th} part of the composition $(1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2$:

$$((1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2)_i = \frac{x_{1i}^{1-\lambda} x_{2i}^{\lambda}}{\sum_{j=1}^{D} x_{1j}^{1-\lambda} x_{2j}^{\lambda}} \geq k,$$

that is, $(1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2 \in \Sigma^+$. ■

Proposition 2 *Let f_1 and f_2 be two $\mathcal{A}$-convex functions on the $\mathcal{A}$-convex set $W \subset \mathcal{S}^D$. The function $f_1 + f_2$ is $\mathcal{A}$-convex on W.*

Proof. Let $\mathbf{x}_1,\ \mathbf{x}_2$ be two D-part compositions in W. For any $\lambda \in [0,1]$ it holds that

$$f_1((1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2) + f_2((1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2) \leq (1-\lambda)[f_1(\mathbf{x}_1) + f_2(\mathbf{x}_1)] + \lambda[f_1(\mathbf{x}_2) + f_2(\mathbf{x}_2)]. \tag{8}$$

■

Proposition 4 *Let f be an $\mathcal{A}$-convex function on the $\mathcal{A}$-convex set $W \subset \mathcal{S}^D$. The sublevel set $\mathcal{G}_\alpha^- = \{\mathbf{x}|\ \mathbf{x} \in W, f(\mathbf{x}) \leq \alpha\}$ is an $\mathcal{A}$-convex set for any real number α.*

Proof. Let $\mathbf{x}_1,\ \mathbf{x}_2$ be in $\mathcal{G}_\alpha^-$ and $\lambda \in [0,\ 1]$,

$$f((1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2) \leq (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2) \leq \alpha.$$

Consequently, $(1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2 \in \mathcal{G}_\alpha^-$. ■

Example 1 The function $f(\mathbf{x}) = \frac{x_i}{x_j},\ 1 \leq i, j \leq D$ is an $\mathcal{A}$-convex over its domain, $dom(f) = \mathcal{S}^D$.

Proof. Let $\mathbf{x}_1,\ \mathbf{x}_2$ be two D-part compositions and for any $\lambda \in [0,\ 1]$,

$$f((1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2) = \left(\frac{x_{1i}}{x_{1j}}\right)^{1-\lambda}\left(\frac{x_{2i}}{x_{2j}}\right)^{\lambda}.$$

We apply Young's inequality, $a^{1-\lambda}b^\lambda \leq (1-\lambda)a + \lambda b$ for any $a, b \geq 0$ and $\lambda \in [0,\ 1]$, to the values $a = \dfrac{x_{1i}}{x_{1j}}$ and $b = \dfrac{x_{2i}}{x_{2j}}$,

$$f((1-\lambda)\odot\mathbf{x}_1 \oplus \lambda \odot \mathbf{x}_2) = \left(\frac{x_{1i}}{x_{1j}}\right)^{1-\lambda}\left(\frac{x_{2i}}{x_{2j}}\right)^{\lambda} \leq$$
$$(1-\lambda)\frac{x_{1i}}{x_{1j}} + \lambda\frac{x_{2i}}{x_{2j}} = (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2). \quad (9)$$

■

Example 2 For any $i = 1, \ldots, D$, the function $f(\mathbf{x}) = x_i$ is an $\mathcal{A}$-quasiconcave function over all its domain, $dom(f) = \mathcal{S}^D$. Moreover, using the $\mathcal{A}$-convex function

$$\Phi_\alpha(\mathbf{x}) = \alpha\sum_{j=1}^{D}\frac{x_j}{x_i} - 1\,,$$

a superlevel $\mathcal{G}_\alpha^+ = \{\mathbf{x} \in \mathcal{S}^D|\ x_i \geq \alpha\}$ for any $\alpha \in (0,\ 1)$ can be represented by means of $\Phi_\alpha(\mathbf{x}) \leq 0$.

Proof. Through proposition 1, the set $\mathcal{G}_\alpha^+ = \{\mathbf{x} \in \mathcal{S}^D|\ x_i \geq \alpha\}$ is $\mathcal{A}$-convex for any real number α.

Because $\mathbf{x} \in \mathcal{S}^D$, the equality $\sum_{j=1}^D x_j = 1$ holds. So, $x_i \geq \alpha \iff \alpha\sum_{j=1}^D \frac{x_j}{x_i} \leq 1$. And the function $\Phi_\alpha(\mathbf{x}) = \alpha\sum_{j=1}^D \frac{x_j}{x_i} - 1$ is $\mathcal{A}$-convex because it is a positive linear combination of $\mathcal{A}$-convex functions $\frac{x_j}{x_i}$. ■

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, London. Reprinted 2003 with additional material by The Blackburn Press, London, UK.

Barceló-Vidal, C. and Martín-Fernández, J. A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics*, 45(4), 57-71.

Barceló-Vidal, C., Martín-Fernández, J. A. and Mateu-Figueras, G. (2011). Compositional differential calculus on the simplex. In: Pawlowsky, V. and Buccianti, A. (eds) *Compositional Data Analysis: Theory and Applications*. Chichester (UK), John Wiley & Sons, Chapter 13, 176-190.

Bates, S. and Tibshirani, R. (2019). Log-ratio lasso: Scalable, sparse estimation for log-ratio models. *Biometrics*, 75(2), 613-624.

Billheimer, D., Guttorp, P. and Fagan, W. F. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456), 1205-1214.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge university press.

Campbell, S. and Wong, T. (2022). Eficient convex pca with applications to wasserstein geodesic pca and ranked data. (in press).

Caussinus, H., Ettinger, P. and Tomassone, R. (2012). *COMPSTAT 1982 5th Symposium Held at Toulouse 1982: Part I: Proceedings in Computational Statistics*. Springer Science & Business Media.

Chastin, S.F.M., Palarea-Albaladejo, J., Dontje, M.L. and Skelton, D.A. (2015). Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic markers: A novel compositional data analysis approach. *PLoS ONE*, 10.

Chen, R., Zhang, Z., Feng, C., Hu, K., Li, M., Li, Y., Shimizu, K., Chen, N., and Sugiura, N. (2010). Application of simplex-centroid mixture design in developing and optimizing ceramic adsorbent for as(v) removal from water solution microporous and mesoporous materials. *Microporous and Mesoporous Materials*, 131, 115-121.

Coetzer, R. and Haines, L. (2017). The construction of d- and i-optimal designs for mixture experiments with linear constraints on the components. *Chemometrics and Intelligent Laboratory Systems*, 171, 112-124.

Dumuid, D., Pedi**v**sić, Z., Palarea-Albaladejo, J., Martín-Fernández, J. A., Hron, K., and Olds, T. (2020). Compositional data analysis in time-use epidemiology: what, why, how. *International Journal of Environmental Research and Public Health*, 17(2220).

Dumuid, D., Wake, M., Burgner, D., Tremblay, M., Okely, A., Edwards, B., Dwyer, T., and Olds, T. (2021). Balancing time use for children's fitness and adiposity: Evidence to inform 24-hour guidelines for sleep, sedentary time and physical activity. *PLoS ONE*, 16(1).

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37, 795-828.

Egozcue, J. J., Pawlowsky-Glahn, V. and Gloor, G. (2018). Linear association in compositional data analysis. *Austrian Journal of Statistics*, 47, 3.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3), 279-300.

Fairclough, S., Dumuid, D., Mackintosh, K., Stone, G., Dagger, R., Stratton, G., Davies, I. and Boddy, L. (2018). Adiposity, fitness, health-related quality of life and the reallocation of time between children's school day activity behaviours: A compositional data analysis. *Preventive medicine reports*, 11, 254-261.

Gupta, N., Rasmussen, C., Holtermann, A., and Mathiassen, S. (2020). Time-based data in occupational studies: The whys, the hows, and some remaining challenges in compositional data analysis (coda). *Annals of Work Exposures and Health*, 64(8):778–785.

Halim, N., Abidin, Z., Siajam, S., Hean, C. and Harun, M. (2021). Optimization studies and compositional analysis of subcritical water extraction of essential oil from citrus hystrix dc. leaves. *The Journal of Supercritical Fluids*, 178.

Harsh, A., Ball, J. and Wei, P. (2016). Onion-peeling outlier detection in 2-d data sets. *International Journal of Computer Applications*, 139(3), 26-31.

Kitano, N., Kay, Y., Jindo, T., Tsunoda, K. and Arao, T. (2020). Compositional data analysis of 24-hour movement behaviors and mental health in workers. *Preventive medicine reports*, 20.

Lo Huang, M.-N. and Huang, M.-K. (2009). ϕ_p-optimal designs for a linear log contrast model for experiments with mixtures. *Metrika*, 70, 239-256.

Luenberger, D. and Ye, Y. (2008). *Linear and Nonlinear Programming*. International Series in Operations Research & Management Science. Springer US.

Martín-Fernández, J. A. (2019). Comments on: Compositional data: the sample space and its structure. *TEST*, 28(3), 653-657.

Martín-Fernández, J. A., Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosona-Delgado, R. (2018). Advances in principal balances for compositional data. *Mathematical Geosciences*, 50(3), 273-298.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2011). The principle of working on coordinates. In *Compositional Data Analysis: Theory and Applications.*

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2013). The normal distribution in some constrained sample spaces. *SORT (Statistics and Operations Research Transactions)*, 37, 29-56.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15, 384-398.

Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons, Chichester.

R-Core-Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Small, C. (1990). A survey of multidimensional medians. *International Statistical Review*, 58, 263-277.

Susin, A., Wang, Y., Lê Cao, K.-A. and Calle, M. L. (2020). Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, 2(2), lqaa029.

Tolosana-Delgado, R., von Eynatten, H. and Karius, V. (2011). Constructing modal mineralogy from geochemical composition: A geometric-bayesian approach. *Computers & Geosciences*, 37(5), 677-691.

van den Boogaart, K. and Tolosana-Delgado, R. (2008). "compositions": a unified r package to analyze compositional data. *Computers & Geosciences*, 34(4), 320-338.

van den Boogaart, K., Filzmoser, P., Hron, K., Templ, M. and Tolosana-Delgado, R. (2021). Classical and robust regression analysis with compositional data. *Mathematical Geosciences*, 53, 823-858.

Wang, X., Wang, H., Wang, Z. and Yuan, J. (2020). Convex clustering method for compositional data modeling. *Soft Computing*, 25, 2965-2980.

4.2 Journal of Geochemical Exploration

El segon article introdueix com a novetat la norma L^1-*plr*, una norma basada en el valor absolut dels *logpairwise* de la composició $\mathbf{x}$ (O2 2.1). Es presenten i demostren les propietats més importants de la norma. Aquesta norma s'utilitza en el terme de penalització de la regressió LASSO, i s'obté com a resultat un model lineal regularitzat que discrimina entre balanços influents i balanços no influents sobre la variable resposta.

L'article ha estat publicat a la revista Journal of Geochemical Exploration.
Volum: 255, número: -, pàgines: -
Enviat: agost 2023, acceptat: octubre 2023
DOI: 10.1016/j.gexplo.2023.107327
Factor d'impacte: 3.4 (Q1).

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Lasso regression method for a compositional covariate regularised by the norm L^1 pairwise logratio

Jordi Saperas-Riera *, Glòria Mateu-Figueras, Josep Antoni Martín-Fernández

Universitat de Girona, C/de la Universitat de Girona, 6, Girona 17003, Spain

ARTICLE INFO

Keywords:
Aitchison's geometry
Compositional data
Norm L^1
Balance selection

ABSTRACT

Lasso regression methods include a penalty function expressed in terms of a norm defined in the space of model coefficients. The norm plays a key role as regards the way coefficients can become irrelevant in the model. For models with a compositional covariate, the norm should be coherent with the Aitchison geometry. The proposed method is based on a newly-defined compositional norm called L^1 pairwise logratio. The novel approach allows one to construct an appropriate basis through a sequential binary partition for discriminating between balances that influence the response variable and those that have no effect. This generalised Lasso regression scheme is illustrated with the analysis of a geochemical data set.

1. Introduction

One of the goals of linear regression analysis is to identify a subset of explanatory variables that are associated with the response variable. For example, in geochemistry it may be of interest to identify which chemical elements have an important effect on the soil pH in a particular region. To address this, Lasso regression methods, introduced by Tibshirani (1996), are a popular option for variable selection. Lasso regression applies an L^1-norm penalisation to the model coefficients (*slopes*), where the L^1-norm is the sum of the absolute value of the coefficients. The standard regression models assume the independence of the covariates, having each one its own slope. Importantly, these assumptions do not apply to a compositional explanatory variable, that is, in the case of compositional data (CoDa).

CoDa analysis (Aitchison, 1986) has become increasingly important in various fields such as environmental science, geochemistry, microbiology, and economics. However, CoDa poses unique challenges, especially when compositions are used as covariates in regression models. Indeed, following the *principle of working on coordinates* (Mateu-Figueras et al., 2011), the D-part composition in the explanatory part of the model should be expressed in terms of at least $D-1$ logarithms of ratios of raw variables (*logratios*). Recently, a number of papers provided tools and methods for regression model simplification with CoDa. First works on penalised regression with compositional covariates are Lin et al. (2014); Shi et al. (2016); Lu et al. (2019), later extended to robust regression in Monti and Filzmoser (2021, 2022). Some of them are focused on considering all possible pairwise logratios in a penalised regression model (Bates and Tibshirani, 2019; Susin et al., 2020; Calle and Susin, 2022a, 2022b; Calle et al., 2023) with the usual L^1-norm (Lasso) or L^2-norm (Ridge) or a linear combination of both (Elastic net) applied to the model coefficients. Other works use supervised learning methods to select pairwise logratios in a generalised linear model (Coenders and Greenacre, 2022). A pairwise logratio approach in CoDa analysis is based on comparing the logarithm of the ratios between two parts of a composition. This approach allows one to analyse the relative information between different parts while avoiding issues of scale dependence and spurious correlation (Aitchison, 1986). Importantly, an approach based on balances can be considered as a generalisation because a balance is a logarithm of the ratios between the average of two groups of parts (Egozcue and Pawlowsky-Glahn, 2005). Balances in CoDa analysis are useful for identifying geochemical relationships and gaining insights into geological processes. By examining the ratios of different elements within samples, researchers can determine patterns and potential causes of variation. This approach can be applied to a range of materials, from rocks and minerals to soils and sediments, and can inform our understanding of issues (Buccianti and Grunsky, 2014). In the context of linear regression models for CoDa, Rivera-Pinto et al. (2018) propose a stepwise algorithm for selecting balances but the global optimum is not guaranteed. A more efficient algorithm identifying a sequence of balances is introduced by Gordon-Rodriguez et al. (2022). In addition, Nesrstová et al. (2023) introduce a Partial Least Squares procedure to construct principal balances (Martín-Fernández

* Corresponding author.
E-mail addresses: jordi.saperas@udg.edu (J. Saperas-Riera), gloria.mateu@udg.edu (G. Mateu-Figueras), josepantoni.martin@udg.edu (J.A. Martín-Fernández).

https://doi.org/10.1016/j.gexplo.2023.107327
Received 6 August 2023; Received in revised form 21 September 2023; Accepted 29 September 2023
Available online 18 October 2023
0375-6742/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

et al., 2018) that maximise the explained variability of the response variable.

To our knowledge, none of these recent works deals with identifying a particular structure of the parts in a composition for selecting a subset of parts (*subcomposition*) in a linear regression model. On the other hand, Boogaart et al. (2021) deal with compositional part selection by introducing the concepts of *internal and external subcompositional independence*. In a linear regression model, a subcomposition is internal independent when changes in values within the parts of the subcomposition have no effect on the explained variable. Whereas external independence further assumes that the balance between the parts of the subcomposition and the rest of the parts in the composition also does not influence the response variable. Once one detects a subcomposition both internally and externally independent then the subset of parts can be removed from the model. The key element of the method proposed in this article is the new norm called L^1 pairwise logratio (L^1-plr) taking part in the penalty term of the Lasso regression model. Using this norm, the Lasso method is able to automatically identify internal independent subcompositions, that is, it detects which pairwise logratios are influential in the response variable and which are not. Once the independent subcompositions have been identified, the model can be simplified by considering an adequate set of balances involving the corresponding parts. Furthermore, a Bootstrap scheme is proposed for checking the external independence and, in such a case, indicating the parts that can be removed from the model.

This article is organised as follows. In Section 2 the basic concepts of CoDa and standard penalty regression are described. In Section 3 the new norm L^1-plr is introduced, and its compositional properties are provided. In Section 4 the Lasso regression model using the norm L^1-plr is formulated and its properties are explored. A geochemical case study is provided in Section 5 for illustration purposes. Finally, the last section concludes with some remarks.

The analyses discussed in this article were carried out in R (R-Core-Team, 2022) using the packages *ADMM* (You and Zhu, 2021) and *coda.base* (Comas-Cufí, 2022).

2. Some basic concepts

2.1. Compositional data

CoDa conveys relative information because the variables describe relative contributions to a given total (Aitchison, 1986). These variables are called *parts* of a whole and are, usually, expressed in proportions, percentages or ppm. Historically (Aitchison, 1986), the sample space of CoDa is designed as the D-part unit simplex $\mathscr{S}^D = \{\mathbf{x} \in \mathbb{R}^D : x_j > 0; \sum x_j = 1; j = 1,\ldots,D\}$. The formal geometric framework for the analysis of CoDa first appeared in Pawlowsky-Glahn and Egozcue (2001) and Billheimer et al. (2001). This geometry was coined the *Aitchison geometry*, later formally established in Barceló-Vidal and Martn-Fernández (2016). The property of scale invariance of results in the analysis offers a broader understanding of compositions. According to this property, two vectors one multiple of the other are considered compositionally equivalent. Consequently, the set of vectors proportional to $\mathbf{x} \in \mathscr{S}^D$ ($\{k\cdot\mathbf{x};\ k > 0\}$) is called a composition and for simplicity denoted again by $\mathbf{x}$. While the compositional space is the set of all compositions and it is denoted for simplicity by $\mathscr{S}^D$.

The Aitchison geometry is based on two specific operations that induce a vector space structure on $\mathscr{S}^D$ called *perturbation* and *powering*, and defined as $\mathbf{x} \oplus \mathbf{y} = (x_1y_1, x_2y_2, \ldots, x_Dy_D)$ and $\alpha \odot \mathbf{x} = (x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)$ for $\mathbf{x},\mathbf{y} \in \mathscr{S}^D$, $\alpha \in \mathbb{R}$. In order to interpret the results of these operations, one can *closure* the result, that is, to normalise the resulting vector to a unit sum by dividing each component by its total sum. Note that the closure operation provides a vector compositionally equivalent.

Once we have a vector space structure, a metric structure is easily defined using the clr-scores of a composition $\mathbf{x}$ (Aitchison, 1986):

$$\mathrm{clr}(\mathbf{x}) = \big(\mathrm{clr}(\mathbf{x})_1, \ldots, \mathrm{clr}(\mathbf{x})_D\big) = \left(ln\frac{x_1}{g(\mathbf{x})}, \ldots, ln\frac{x_D}{g(\mathbf{x})}\right),$$

where $g(\cdot)$ is the geometric mean of the composition. Indeed, the basic metric elements of the Aitchison geometry: inner product ($<\cdot,\cdot>_{\mathscr{A}}$), L^2-norm ($\|\cdot\|_{\mathscr{A}}$), and distance ($d_{\mathscr{A}}(\cdot,\cdot)$) are

$$<\mathbf{x},\mathbf{y}>_{\mathscr{A}} = <\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y})>_E \ , \quad \|\mathbf{x}\|^2_{\mathscr{A}} = <\mathbf{x},\mathbf{x}>_{\mathscr{A}} \ , \quad d_{\mathscr{A}}(\mathbf{x},\mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{\mathscr{A}}, \tag{1}$$

where "$\mathscr{A}$" means the Aitchison geometry, "E" means the typical Euclidean geometry, and "$\ominus$ " is the perturbation difference $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$.

An important scale invariant function is the *logcontrast* because it plays the typical role of the linear combination of variables. Given a composition $\mathbf{x} = (x_1, \ldots, x_D)$, a logcontrast is defined as any linear combination of the logarithms of the compositional parts:

$$\sum_{j=1}^{D} a_j ln x_j, \quad \text{with} \quad \sum_{j=1}^{D} a_j = 0, \quad a_j \in \mathbb{R}. \tag{2}$$

Note that each clr-score $\mathrm{clr}(\mathbf{x})_j; j = 1,\ldots,D$, is a logcontrast, and, on the other side, any logcontrast can be expressed as a logratio

$$\sum_{j=1}^{D} a_j ln x_j = ln\frac{\prod_{a_j>0} x_j^{a_j}}{\prod_{a_j<0} x_j^{|a_j|}}.$$

In fact, parts with $a_j > 0$ in (2) appear in the numerator, and parts with $a_j < 0$ appear in the denominator. If a part has no contribution, then $a_j = 0$.

The metric elements defined in Eq. (1) can be used to construct an orthonormal logratio (*olr*) basis and to calculate the corresponding olr-coordinates of a composition ($\mathrm{olr}(\mathbf{x})$, formerly known as ilr-coordinates) (Egozcue et al., 2003; Martín-Fernández, 2019). The expression of these olr-coordinates depends on the basis selected. Following Egozcue and Pawlowsky-Glahn (2005), one can define particular olr-coordinates, called *balances*. A balance involves two groups of parts of a composition and is expressed as the logratio of the geometric mean of each group of parts multiplied by a constant to guarantee the unit length of the vectors of the basis.

A sequential binary partition (SBP) of a composition $\mathbf{x} = (x_1, \ldots, x_D)$ provides balances associated with a specific olr-basis. In the first step of an SBP, the full composition $\mathbf{x} = (x_1, \ldots, x_D)$ is split into two groups of parts: one for the numerator (coded with $+1$) and the other for the denominator (with code -1). According to this partition, the first olr-coordinate is obtained as the logarithm of the geometric mean of the parts in the numerator divided by the geometric mean of the parts in the denominator, multiplied by a scaling factor that depends on the number of parts (Eq. (3)). In the following steps, each group of parts is in turn split into two groups and the following olr-coordinates are obtained. In step k when the $\mathrm{olr}(\mathbf{x})_k$-coordinate is created, the r_k parts $(x_{n1_k}, \ldots, x_{nr_k})$ in the first group are placed in the numerator (code $+1$); the s_k parts $(x_{d1_k}, \ldots, x_{ds_k})$ in the second group will appear in the denominator (code -1); and the rest of $D - (r_k + s_k)$ parts are not involved in the logratio (code 0). As a result, the $\mathrm{olr}(\mathbf{x})_k$ is:

$$\mathrm{olr}(\mathbf{x})_k = \sqrt{\frac{r_k \cdot s_k}{r_k + s_k}} ln \frac{\left(x_{n1_k} \cdots x_{nr_k}\right)^{1/r_k}}{\left(x_{d1} \cdots x_{ds_k}\right)^{1/s_k}}, \quad k = 1, \ldots, D-1, \tag{3}$$

where $\sqrt{\frac{r_k \cdot s_k}{r_k + s_k}}$ is the factor for normalising vectors of the basis. Note that the $\mathrm{olr}(\mathbf{x})_k$ coordinate, being a logcontrast that involves two groups of parts, informs us of, on average, the relative importance of one group of parts with regard to the other.

Relating the clr-scores with any olr-coordinates by means of a matrix relationship is straightforward. Indeed, $\mathrm{olr}_\Psi(\mathbf{x}) = \Psi \mathrm{clr}(\mathbf{x})$ and $\mathrm{clr}(\mathbf{x}) =$

$\Psi^T \text{olr}_\Psi(\mathbf{x})$, with $\Psi \in \mathbb{R}^{(D-1)\times D}$ a matrix where the $D-1$ rows are the clr-scores of compositions forming the olr-basis. Consequently, all compositional operations and compositional metric elements (Eq. (1)) are translated into ordinary operations between the corresponding olr-coordinates.

2.2. Linear model with compositional covariates

Given a dependent variable y and an explanatory D-part composition $\mathbf{x}$, the definition of a linear regression model in terms of a logcontrast (Aitchison and Bacon-Shone, 1984; Hron et al., 2012) is:

$$y = \alpha_0 + \sum_{j=1}^{D} \alpha_j ln x_j, \quad \text{with} \quad \sum_{j=1}^{D} \alpha_j = 0, \quad a_j \in \mathbb{R}, \tag{4}$$

whereas, in terms of metric elements, the model formulation is (Boogaart and Tolosana, 2013):

$$y = \beta_0 + \langle \boldsymbol{\beta},\ \mathbf{x} \rangle_{\mathscr{A}} = \beta_0 + \langle \text{clr}(\boldsymbol{\beta}),\ \text{clr}(\mathbf{x}) \rangle_E = \beta_0 + \langle \text{olr}_\Psi(\boldsymbol{\beta}),\ \text{olr}_\Psi(\mathbf{x}) \rangle_E, \tag{5}$$

where $\boldsymbol{b}$ is the compositional gradient vector. Note that $\Psi \in \mathbb{R}^{(D-1)\times D}$ is the matrix associated to any olr-basis in the clr-space, for example, a basis created using an SBP. Considering the expression in terms of clr-scores, the coefficients could be estimated using a statistical toolbox but the use of the generalised inversion for the covariance matrix of the clr-scores is required (Boogaart et al., 2021), which it is not necessary when working with the model based on olr-coordinates.

2.3. Penalty regression

The Lasso regression model is formulated as the combination of the L^2-norm cost function and the L^1-norm regularisation term. For a real data set $\mathbf{X}$ with n observations and D predictors and a real response vector $\mathbf{Y}$ of length n, the Lasso regression model can be formulated as (Tibshirani, 1996)

$$min \left\{ \frac{1}{2} \| \mathbf{Y} - \beta_0 - \mathbf{X}\cdot\boldsymbol{\beta} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_1 \right\}, \tag{6}$$

where a is the intercept, $\mathbf{b}$ is the gradient, and λ is the penalty parameter that controls the amount of regularisation. For $\lambda = 0$, the Lasso regression model (Eq. (6)) provides the classical least squares regression model. The larger the value of λ, the greater the number of coefficients in $\mathbf{b}$ forced to be zero. The *optimal* value of λ can be chosen based on cross-validation techniques or other methods (James et al., 2021).

It is possible to generalise the Lasso problem by taking the L^1-norm of a linear transformation of gradient $\mathbf{b}$ as a penalty function. The generalised Lasso regression model is

$$min \left\{ \frac{1}{2} \| \mathbf{Y} - \beta_0 - \mathbf{X}\cdot\boldsymbol{\beta} \|_2^2 + \lambda \| \mathbf{F}\cdot\boldsymbol{\beta} \|_1 \right\}, \tag{7}$$

where $\mathbf{F}$ is the matrix associated to an arbitrary linear transformation. Note that $\mathbf{F} = Id$ corresponds to the simple Lasso problem.

The metric elements (Eq. (1)) used to define the regression model (Eq. (5)) facilitate the formulation of the cost function in a Lasso model Eqs. (6) and (7) for compositional covariates. However, the definition of an appropriate L^1-norm for CoDa requires the supplementary concepts described in the following section.

3. Norm L^1 pairwise logratio

The Aitchison norm $\| \mathbf{x} \|_{\mathscr{A}}$ (Eq. (1)) is defined as the Euclidean L^2-norm of the clr-scores ($L^2 - \text{clr}$):

$$\| \mathbf{x} \|_{\mathscr{A}}^2 = \| \mathbf{x} \|_{2-\text{clr}}^2 = \| \text{clr}(\mathbf{x}) \|_2^2 = \sum_{j=1}^{D} \left(ln\left(\frac{x_j}{g(\mathbf{x})} \right) \right)^2, \tag{8}$$

that is, $\| \mathbf{x} \|_{\mathscr{A}}$ can be interpreted as the restriction of a L^2 Euclidean norm on the clr-space. Following this idea, the norm $L^1 - \text{clr}$ ($\| \mathbf{x} \|_{1-\text{clr}}$) for a Lasso regression model can be defined as (Bates and Tibshirani, 2019; Susin et al., 2020)

$$\| \mathbf{x} \|_{1-\text{clr}} = \| \text{clr}(\mathbf{x}) \|_1 = \sum_{j=1}^{D} \left| ln\left(\frac{x_j}{g(\mathbf{x})} \right) \right|. \tag{9}$$

Note that a regularisation term $\| \boldsymbol{\beta} \|_{1-\text{clr}}$ forces some components $\text{clr}(\boldsymbol{\beta})_j$, $j = 1, \ldots, D$, to take small values, suggesting the corresponding parts $\mathbf{x}_j$ could be removed from the model. However, the presence of the removed parts in the geometrical mean of the non-removed parts ($g(\boldsymbol{\beta})$ and $g(\mathbf{x})$) is a difficulty for the regression model simplification.

Importantly, the Aitchison distance (Eq. (1)) can also be defined in terms of pairwise logratios (Aitchison et al., 2000). Consequently, the Aitchison norm can be expressed as $\| \mathbf{x} \|_{\mathscr{A}}^2 = \frac{1}{D} \sum_{i<j} \left(ln\left(\frac{x_i}{x_j} \right) \right)^2$. Based on this expression, a new L^1-norm on the simplex $\mathscr{S}^D$ is defined as:

Definition 1. The L^1-plr norm of a composition $\mathbf{x} \in \mathscr{S}^D$ is

$$\| \mathbf{x} \|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| ln\left(\frac{x_i}{x_j} \right) \right|. \tag{10}$$

Proposition 2. $\| \mathbf{x} \|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| ln\left(\frac{x_i}{x_j} \right) \right|$ *verifies the properties of a norm. That is:*

- *Positive definiteness:* $\forall \mathbf{x} \in \mathscr{S}^D$, $\| \mathbf{x} \|_{1-plr} \geq 0$. *Moreover,* $\| \mathbf{x} \|_{1-plr} = 0$ *if and only if* $\mathbf{x} = (1, \ldots, 1)$.
- *Absolute homogeneity:* $\forall \mathbf{x} \in \mathscr{S}^D$ *and* $\forall \lambda \in \mathbb{R}$, $\| \lambda \odot \mathbf{x} \|_{1-plr} = |\lambda| \, \| \mathbf{x} \|_{1-plr}$.
- *Subadditivity:* $\forall \mathbf{x},\ \mathbf{y} \in \mathscr{S}^D$, $\| \mathbf{x} \oplus \mathbf{y} \|_{1-plr} \leq \| \mathbf{x} \|_{1-plr} + \| \mathbf{y} \|_{1-plr}$.

See Appendix for the proof.

Importantly, the coefficient accompanying the sum of squared pairwise logratios in the Aitchison norm is $1/D$, whereas in the norm L^1-plr is $1/(D-1)$ (Eq. (10)). Using this factor, the norm L^1-plr is endowed with the property of subcompositional dominance among other compositional properties:

Proposition 3. *The* L^1*-plr norm on* $\mathscr{S}^D$, $\| \mathbf{x} \|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| ln\left(\frac{x_i}{x_j} \right) \right|$ *verifies the properties*

- *Scale invariance:* $\| \mathbf{x} \|_{1-plr} = \| \lambda \mathbf{x} \|_{1-plr}$, $\lambda > 0$.
- *Permutation invariance:* $\| (x_1, \ldots, x_i, \ldots, x_j, \ldots, x_D) \|_{1-plr} = \| (x_1, \ldots, x_j, \ldots, x_i, \ldots, x_D) \|_{1-plr}$.
- *Subcompositional dominance:* $\| \mathbf{x} \|_{1-plr} \geq \| \text{sub}(\mathbf{x}) \|_{1-plr}$ *where* $\text{sub}(\mathbf{x})$ *denotes any subcomposition of* $\mathbf{x}$.

See Appendix for the proof.

The norm L^1-plr can be expressed in terms of the clr-scores as

$$\| \mathbf{x} \|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| ln\left(\frac{x_i / g(\mathbf{x})}{x_j / g(\mathbf{x})} \right) \right| = \frac{1}{D-1} \sum_{i<j} \left| \text{clr}(\mathbf{x})_i - \text{clr}(\mathbf{x})_j \right|, \tag{11}$$

suggesting that, in general, $\| \mathbf{x} \|_{1-plr} \neq \| \mathbf{x} \|_{1-clr}$.

Fig. 1 shows the shape of the unit balls (i.e., set of points that have distance 1 from the origin) measured by the norms L^1-clr (Eq. (9), green), L^1-plr (Eq. (10), orange), and Aitchison (Eq. (1), blue) in the 3-part compositional space. To represent it, the olr-coordinates $\text{olr}_1(\mathbf{x}) = \frac{\sqrt{2}}{2} ln \frac{x_1}{x_2}$, and $\text{olr}_2(\mathbf{x}) = \sqrt{\frac{2}{3}} ln \frac{\sqrt{x_1 x_2}}{x_3}$ are used. Because the norms are calculated using clr-scores and pairwise logratios, the value of the norms is

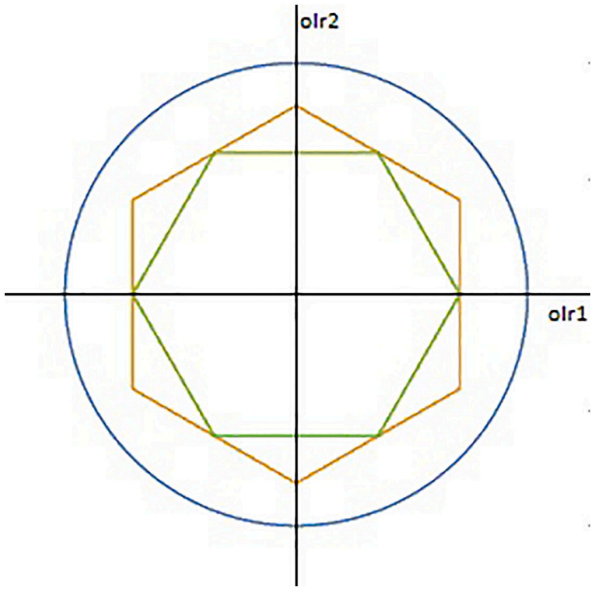


Fig. 1. Unit balls in the olr-coordinates space of 3-part compositions using the norms: L^2-clr (blue), L^1-clr (green), and L^1-plr (orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

invariant under a change of olr-basis. When one takes a different olr-basis then the shape of the unit balls is only affected by a rotation. As expected, the unit ball of the L^2 Aitchison norm (blue) shows the typical shape of a circle, which includes the L^1 unit balls as it is well known for norms L^p in Euclidean spaces. Interestingly, both norms L^1-clr (green) and L^1-plr (orange) create a hexagon, latter being the biggest. That is, the unit ball with norm L^1-plr includes the unit ball of norm L^1-clr because it holds $\| \mathbf{x}\|_{1-plr} \leq \| \mathbf{x}\|_{1-clr}$, for $\mathbf{x} \in \mathscr{S}^D$ (see Appendix for the proof). The points of contact between the unit balls correspond to $\mathbf{x} \in \mathscr{S}^3$ that $\mathrm{clr}(\mathbf{x}) \in \left\{ \left(\pm\frac{1}{2}, \mp\frac{1}{2}, 0\right); \left(\pm\frac{1}{2}, 0, \mp\frac{1}{2}\right); (0, \pm\frac{1}{2}, \mp\frac{1}{2}) \right\}$, where $\| \mathbf{x}\|_{1-\mathrm{plr}} = \| \mathbf{x}\|_{1-\mathrm{clr}} = 1$.

4. Generalised Lasso regression with the norm L^1 pairwise logratio

To our knowledge, the compositional Lasso regression methods introduced in the literature (Bates and Tibshirani, 2019; Susin et al., 2020; Calle and Susin, 2022a, 2022b; Calle et al., 2023) aim to separate the parts into two groups: parts that influence the response variable, and parts that do not. However, the methods do not analyse the *external independence* of parts that do not affect the response variable (Boogaart et al., 2021). That is, they do not explore the balance between the non-influential subcomposition and the rest of the parts.

We will show that the Lasso regression using our new norm L^1-plr aims to identify and separate the balances (i.e., pairwise logratios) into two groups: the balances that influence the response variable, and those that do not. Therefore this method permits the analyst to deal with both types of subcompositional independence: *internal* and *external* (Boogaart et al., 2021).

Definition 4. Given y_i, $i = 1, \ldots, n$ the sample of the response variable, $\mathbf{X}$ the $n \times D$ matrix whose rows, $\mathbf{X}_i = (x_{i1}, \ldots, x_{iD})$ for $i = 1, \ldots, n$, contains the compositional sample, and $\mathrm{clr}(\mathbf{X})_i$ the i-th row of matrix $\mathrm{clr}(\mathbf{X})$. The L^1-plr Lasso estimator is defined as

$$\boldsymbol{\beta} \in \underset{\boldsymbol{\beta}}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \mathrm{clr}(\boldsymbol{\beta}),\ \mathrm{clr}(\mathbf{X})_i \rangle_E \right)^2 + \lambda \| \boldsymbol{\beta}\|_{1-plr} \right\}. \tag{12}$$

Following Eq. (11), it holds that $\| \boldsymbol{\beta}\|_{1-plr} = \| \mathbf{F} \cdot \mathrm{clr}(\boldsymbol{\beta})\|_1$, where $\mathbf{F}$ is the matrix $\frac{D(D-1)}{2} \times D$ associated to the linear transformation $F(z_1, \ldots, z_D) = \frac{1}{D-1}(z_1 - z_2, z_1 - z_3, \ldots, z_1 - z_D, z_2 - z_3, \ldots, z_2 - z_D, \ldots, z_{D-1} - z_D)$. Consequently, Definition 4 can be generalised to:

$$\boldsymbol{\beta} \in \underset{\boldsymbol{\beta}}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \mathrm{clr}(\boldsymbol{\beta}),\ \mathrm{clr}(\mathbf{X})_i \rangle_E \right)^2 + \lambda \| \mathbf{F} \cdot \mathrm{clr}(\boldsymbol{\beta})\|_1 \right\}. \tag{13}$$

The matrix $\mathrm{clr}(\mathbf{X})$ is not a full rank matrix, thus causing troubles when solving the convex optimisation problem (Saperas-Riera et al. (2023)) in Eq. (13). To avoid these troubles the problem can be solved in terms of olr_Ψ-coordinates, $\mathrm{olr}_\Psi(\mathbf{x}) = \Psi \cdot \mathrm{clr}(\mathbf{x})$, that is, the L^1-plr Lasso estimator (Eq. (12)) in olr_Ψ-coordinates is

$$\boldsymbol{\beta} \in \underset{\boldsymbol{\beta}}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \mathrm{olr}_\Psi(\boldsymbol{\beta}),\ \mathrm{olr}_\Psi(\mathbf{X})_i \rangle_E \right)^2 + \lambda \| \mathbf{F} \cdot \Psi^T \cdot \mathrm{olr}_\Psi(\boldsymbol{\beta})\|_1 \right\}. \tag{14}$$

The relationship between Eq. (12) and Eq. (14) can be used for analyzing the relations between coefficients $\mathrm{clr}(\boldsymbol{\beta})_j$; $j = 1, \ldots, D$ and the coefficients of the balances in the generalised Lasso model (Eq. (14)). Importantly, the penalty term in Eq. (13) forces the sum of the absolute value of the differences of the clr-scores of the gradient vector to be less than a fixed number, which forces some pairwise differences of clr-scores to be zero ($\mathrm{clr}(\boldsymbol{\beta})_i - \mathrm{clr}(\boldsymbol{\beta})_j = 0$), that is, forces some pairs of clr-scores to be equal. This means that the corresponding pairwise logratios ($ln\frac{x_i}{x_j}$) do not influence the response variable. For example, without loss of generality, suppose that the pairwise logratio $ln\frac{x_1}{x_2}$ does not influence the response variable y. That is, the coefficient of the balance in the model is equal to zero. In this case, taking an adequate matrix Ψ one can obtain a model in clr-scores with $\mathrm{clr}(\boldsymbol{\beta})_1 = \mathrm{clr}(\boldsymbol{\beta})_2$ (i.e., $\beta_1 = \beta_2$). And vice-versa, if $\boldsymbol{\beta}$ fulfils $\beta_1 = \beta_2$, an olr-basis including the unit vector $\frac{\sqrt{2}}{2}(1, -1, 0, \ldots, 0)$ provides a model where the coefficient of the pairwise logratio $ln\frac{x_1}{x_2}$ is equal to zero, that is, the pairwise logratio $ln\frac{x_1}{x_2}$ does not influence the response variable y. The above reasoning can be extended to a linear model with gradient $\boldsymbol{\beta}$ fulfilling $\beta_1 = \beta_2 = \ldots = \beta_k$, $2 < k < D$. In this case, with an adequate matrix Ψ, one detects that the any pairwise logratio $ln(x_i/x_j)$, $1 \leq i < j \leq k$ do not influence the response variable y. In addition, any balance involving some of the parts in the subcomposition $(x_1, \ldots, x_k)$ does not influence the response variable y. That is, the subcomposition is *internal independent* (Boogaart et al., 2021).

Following the idea described above, a general algorithm for a L^1-plr Lasso method can be formulated as:

Algorithm 1. L^1-plr Lasso.

1. Fit the L^1-plr Lasso model with tuning parameter λ (Eq. (14)).
2. Express the L^1-plr Lasso model in terms of clr-scores (Eq. (5)).
3. For each string detected being $\left\{ \mathrm{clr}(\boldsymbol{\beta})_{j_1} = \ldots = \mathrm{clr}(\boldsymbol{\beta})_{j_k} \right\}$, built an orthonormal basis for the subcomposition $(\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k})$ in $\mathscr{S}^D$ (Eq. (3)).
4. Put together the bases created above for the subcompositions. Complete until an olr-basis basis for the full composition $\mathbf{x} \in \mathscr{S}^D$ is reached. Write the L^1-plr Lasso model in terms of the olr-coordinates (Eq. (5)).

When fitting the model in $\mathrm{olr} - \textit{coordinates}$ (step 1), any matrix Ψ can be used. Once fitted, the relationship between clr-scores and olr-coordinates ($\mathrm{clr}(\mathbf{x}) = \Psi^T \mathrm{olr}_\Psi(\mathbf{x})$) is used for detecting the subcompositions of the gradient vector $\boldsymbol{\beta}$ fulfiling $\beta_{j_1} = \ldots = \beta_{j_k}$, $2 \leq k < D$ (step 2). The

balances forming the olr-bases of these subcompositions do not influence the variable response y (step 3). That is, in this step, internal independent subcompositions are detected. The rest of the parts of the composition $\mathbf{x} \in \mathscr{S}^D$ form an influential subcomposition. When completing the olr-basis for the full composition $\mathbf{x}$, any olr-basis can be created for the influential subcomposition (step 4). Importantly, the corresponding balances linking the different subcompositions detected must be included in the olr-basis. Moreover, the significance of the coefficient of a linking balance between a non-influential subcomposition and the rest of the parts provides information about the *external independence*. Consequently, any subcomposition $\{\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k}\}$, $2 \leq k < D$ both internal and external independent can be removed from the linear regression model. A routine written in R code (R-Core-Team, 2022) has been developed by us to perform the steps involved in carrying out the algorithm. The routine is freely available from the leading author.

5. Case study

Following Boogaart et al. (2021), a total of $n = 2095$ samples of the data set of project GEMAS ("Geochemical Mapping of Agricultural and grazing land Soil") were analyzed. For further information about the data set, you can consult Reimann et al. (2014a, 2014b). The analyzed data set contains information on the *soil pH*, as a real response variable y. The compositional covariate is the 11-part composition $\mathbf{x}$ of the major oxides and LOI (loss on ignition): (SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5, LOI).

To select the optimal λ parameter for our model in Eq. (14), a 10-fold cross-validation was performed. Each iteration involves dividing the data into 10 equal parts, training the model on nine of them, and then evaluating it on the remaining part to produce the lowest Mean Squared Error (MSE). Fig. 2, shows a plot with the MSE curve and the values of $lambda.min = 4.197$ and the $lambda.1se = 52.436$. With the value $lambda.min$, one obtains the minimum mean cross-validated error, whereas $lambda.1se$ is the largest value of the tuning parameter λ such that the error is within one standard error of the cross-validated errors for $lambda.min$. Because the larger value of an optimal λ, the larger the number of regularised coefficients, the value $lambda.1se = 52.436$ was considered for the L^1 Lasso model (James et al., 2021).

The linear penalised model for $\lambda = lambda.1se = 52.436$ expressed in terms of clr-scores (Eq. (5)) has the following coefficients: intercept $\beta_0 = 5.938$ and gradient

$$\mathrm{clr}(\boldsymbol{\beta}) = (0.046, 0.046, 0.046, 0.046, 0.096, 0.083, 0.550, -0.489, 0.096, -0.182, -0.339).$$

In this case, the method detects two strings: $\beta_1 = \beta_2 = \beta_3 = \beta_4$ and $\beta_5 = \beta_9$. Consequently, the associated balances and/or pairwise logratios within the subcompositions $(SiO_2, TiO_2, Al_2O_3, Fe_2O_3)$ and (MnO, K_2O) do not have any influence on the response variable *soil pH*. The first 4-part subcomposition forms a 3-dimensional logratio subspace where all the pairwise logratios and balances involving the major oxides SiO_2, TiO_2, Al_2O_3, and Fe_2O_3 are non-influential. For example, the pairwise logratio $ln\frac{SiO_2}{Al_2O_3}$ or the balance $ln\frac{(SiO_2Al_2O_3)^{1/2}}{(TiO_2Fe_2O_3)^{1/2}}$ are non-influential. The second subcomposition defines a 1-dimensional space, that is, changes in the values of the pairwise logratio $ln\frac{MnO}{K_2O}$ have no effect on the soil pH. On the other hand, pairwise logratios and balances mixing major oxides of the first group $(SiO_2, TiO_2, Al_2O_3, Fe_2O_3)$ with major oxides of the second group (MnO, K_2O) could be influential. In particular, the coefficient for the linking balance between both subcompositions could be significant. The linking balance is the second balance (olr_2, blue) in Table 1. The other balances that can be influential are those involving the rest of parts forming the subcomposition $(MgO, CaO, Na_2O, P_2O_5, LOI)$ and the linking balance between the two types of subcompositions (Table 1, green).

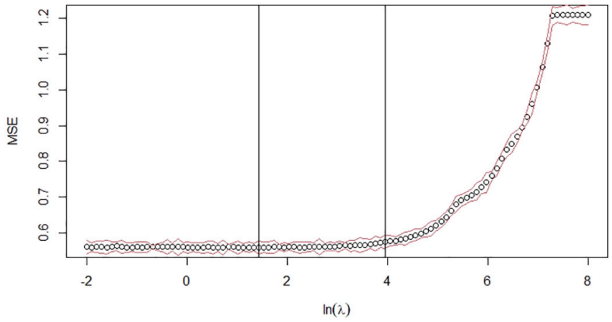


Fig. 2. Cross-validation MSE curve for different log-transformed values of the penalty parameter ($ln(\lambda)$). The circle (∘) is the arithmetical mean of the 10-fold CV. The red lines (above and below the mean) are respectively the value $mean \pm stdev$, where $stdev$ is the standard deviation of the 10-fold CV. Vertical lines are the log-transformed values of $lambda.min = 4.197$ and $lambda.1se = 52.436$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1 summarises the subcompositional structure suggested by the L^1-plr Lasso method when one creates an appropriate olr-basis using an SBP. Note that the code "+ 1"in the SBP means that the part is in the numerator of the balance, whereas for the parts in the denominator, the label is " − 1". The code "0"is reserved for the parts non-involved in the balance. In green, the first row is the full balance between the major oxides involved in the non-influential subcompositions and the remaining major oxides ($MgO, CaO, Na_2O, P_2O_5, LOI$). In blue, olr_2 balances the two non-influential subcompositions. In red, the three rows form the basis of the subcomposition $(SiO_2, TiO_2, Al_2O_3, Fe_2O_3)$. In purple, the sixth row is the vector forming basis of the subcomposition (MnO, K_2O). In black, the last rows show an SBP for creating an olr-basis of the subcomposition with the remaining major oxides.

The L^1-plr Lasso gradient $\boldsymbol{\beta}$ expressed in terms of olr-coordinates associated to the SBP defined in Table 1 is

$$\mathrm{olr}_\Psi(\boldsymbol{\beta}) = (0.228, -0.058, 0, 0, 0, 0, 0.194, 0.735, 0.236, 0.187),$$

where coefficients equal to zero correspond to non-influential balances. Accordingly, the linear regression model is

$$y = 5.938 + 0.228\ \mathrm{olr}_1(\mathbf{x}) - 0.058\ \mathrm{olr}_2(\mathbf{x}) + 0.194\ \mathrm{olr}_7(\mathbf{x}) + 0.735\ \mathrm{olr}_8(\mathbf{x}) + 0.236\ \mathrm{olr}_9(\mathbf{x}) + 0.187\ \mathrm{olr}_{10}(\mathbf{x}). \tag{15}$$

Note that the largest coefficient is $\mathrm{olr}_\Psi(\boldsymbol{\beta})_8 = 0.7345$, suggesting that the ratio $\frac{CaO}{Na_2O}$ concentrates most of the predicting power for pH. That is, one can interpret that when increasing the ratio $\frac{CaO}{Na_2O}$ while keeping all other predictor balances constant the soil pH increases (Coenders and Pawlowsky-Glahn, 2020). This feature coincides with the evaluation of the model presented in Boogaart et al. (2021). Among the other coefficients, it is of particular interest to test if the coefficient of the balance $\mathrm{olr}_2(\mathbf{x})$ is zero ($\mathrm{olr}_\Psi(\boldsymbol{\beta})_2 = -0.058$). Because olr_2 is the linking balance between the two non-influential subcompositions (Table 1), removing this balance would indicate a non-influential 6-part

Table 1
SBP and balances for the olr-basis suggested by the L^1-plr Lasso method. Colours are associated with subcompositions (see text for details).

$D = 11$											
SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5	LOI	*Balances*
+1	+1	+1	+1	+1	−1	−1	−1	+1	−1	−1	$\text{olr}_1 = \sqrt{\frac{30}{11}} \ln \frac{(SiO_2\ TiO_2\ Al_2O_3\ Fe_2O_3\ MnO\ K_2O)^{1/6}}{(MgO\ CaO\ Na_2O\ P_2O_5\ LOI)^{1/5}}$
+1	+1	+1	+1	−1	0	0	0	−1	0	0	$\text{olr}_2 = \sqrt{\frac{4}{3}} \ln \frac{(SiO_2\ TiO_2\ Al_2O_3\ Fe_2O_3)^{1/4}}{(MnO\ K_2O)^{1/2}}$
+1	+1	−1	−1	0	0	0	0	0	0	0	$\text{olr}_3 = \ln \frac{(SiO_2\ TiO_2)^{1/2}}{(Al_2O_3\ Fe_2O_3)^{1/2}}$
+1	−1	0	0	0	0	0	0	0	0	0	$\text{olr}_4 = \frac{\sqrt{2}}{2} \ln \frac{SiO_2}{TiO_2}$
0	0	+1	−1	0	0	0	0	0	0	0	$\text{olr}_5 = \frac{\sqrt{2}}{2} \ln \frac{Al_2O_3}{Fe_2O_3}$
0	0	0	0	+1	0	0	0	−1	0	0	$\text{olr}_6 = \frac{\sqrt{2}}{2} \ln \frac{MnO}{K_2O}$
0	0	0	0	0	−1	+1	+1	0	−1	−1	$\text{olr}_7 = \sqrt{\frac{6}{5}} \ln \frac{(CaO\ Na_2O)^{1/2}}{(MgO\ P_2O_5\ LOI)^{1/3}}$
0	0	0	0	0	0	+1	−1	0	0	0	$\text{olr}_8 = \frac{\sqrt{2}}{2} \ln \frac{CaO}{Na_2O}$
0	0	0	0	0	+1	0	0	0	+1	−1	$\text{olr}_9 = \sqrt{\frac{2}{3}} \ln \frac{(MgO\ P_2O_5)^{1/2}}{LOI}$
0	0	0	0	0	+1	0	0	0	−1	0	$\text{olr}_{10} = \frac{\sqrt{2}}{2} \ln \frac{MgO}{P_2O_5}$

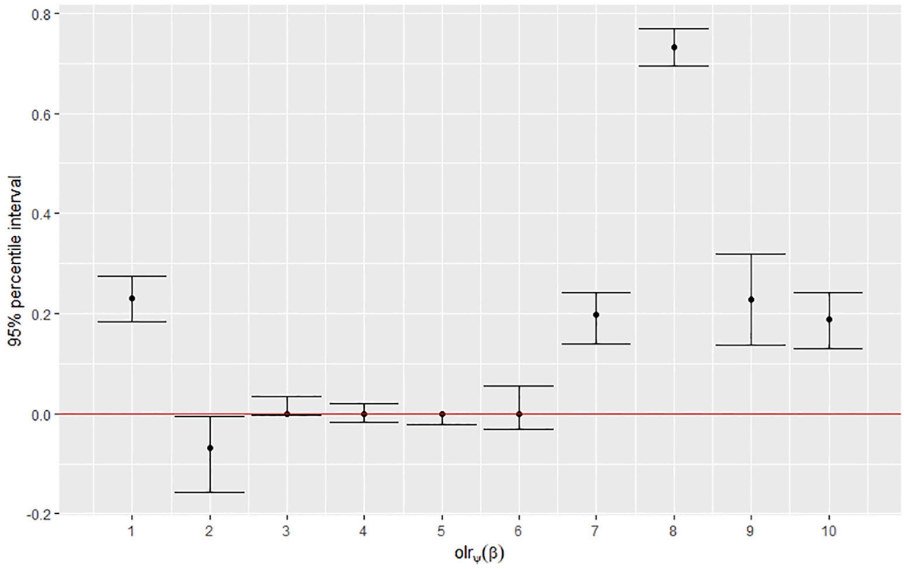


Fig. 3. Bootstrapping 95 % percentile intervals for the coefficients $\text{olr}_\Psi(\boldsymbol{\beta})$ of the linear regression model.

subcomposition (SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, K_2O), generating a 5-dimensional space. In addition, eliminating the balance $\text{olr}_1(\mathbf{x})$ would simplify the model because the subcompositional external independence of the subcomposition would permit removing major oxides SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, and K_2O from the model. In this case, the coefficient involved is $\text{olr}_\Psi(\boldsymbol{\beta})_1 = 0.228$.

Following Hesterberg et al. (2012), one can add the uncertainty associated with the coefficient using a bootstrap technique. Fixed $\lambda = 52.436$ and the SBP (Table 1), 1.000 random samplings with replacement were performed in the data set and the L^1-plr Lasso model was fitted. As a result, the 95 % percentile interval for each coefficient in $\text{olr}_\Psi(\beta)$ was calculated. Fig. 3 shows that the coefficient $\text{olr}_\Psi(\beta)_1 = 0.228$ cannot be considered equal to zero, indicating that the subcomposition (SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, K_2O) cannot be removed from the model because it is not external independent. In addition, because the 95 % percentile interval for the coefficient $\text{olr}_\Psi(\beta)_2 \in (-0.158, -0.005)$ does not contain the zero then one can assume that the balance $\text{olr}_2(\mathbf{x})$ is influential, i.e., it cannot be removed from the model. Finally, as expected, percentile intervals of coefficients $\text{olr}_\Psi(\beta)_3$, $\text{olr}_\Psi(\beta)_4$, $\text{olr}_\Psi(\beta)_5$, and $\text{olr}_\Psi(\beta)_6$ contain the zero because the internal independence of subcomposition (SiO_2, TiO_2, Al_2O_3, Fe_2O_3) and (MnO, K_2O). Consequently, the model in Eq. (15) does not admit further simplification.

Despite L^1-plr Lasso removing four balances from the model (Eq. (15)), all the major oxides participate in the remaining balances. That is, no parts have been removed from the model. On the other hand, when one applies the L^1-clr Lasso method (Lin et al., 2014; Bates and Tibshirani, 2019) the goal is selecting influential parts. In the case of the GEMAS data set, the optimal λ tuning parameter was selected ($lambda.1se = 38.728$) using a 10-fold cross-validation evaluating the MSE. The linear regression model created has intercept $\beta_0 = 6.440$ and

the clr-gradient:

$$\mathrm{clr}(\boldsymbol{\beta}) = (0,\ 0,\ 0,\ 0,\ 0.162, 0.081, 0.555, -0.489, 0.170, -0.142, -0.336).$$

That is, the 4-part subcomposition of major oxides $(SiO_2, TiO_2, Al_2O_3, Fe_2O_3)$ has been removed from the linear regression model:

$$y = 6.44 + 0.162lnMnO + 0.081lnMgO + 0.555lnCaO - 0.489lnNa_2O + 0.170lnK_2O - 0.142lnP_2O_5 - 0.336lnLOI.$$

Note that the coefficients of this model have a lack of interpretation (Coenders and Pawlowsky-Glahn, 2020). For example, despite the coefficient of CaO being 0.555, it is not possible to interpret that one should expect an increase in soil pH when CaO is increased while the other concentrations are kept constant. Because concentrations are relative data (CoDa), the proportion of one part cannot increase if the other proportions are kept. Consequently, an adequate interpretation of the model requires being expressed in terms of balances or pairwise logratios (Coenders and Pawlowsky-Glahn, 2020).

6. Final remarks

Because concentrations of geochemical elements are compositional, any statistical analysis has to consider their relative nature. This article fills the gap for methods for automatic recognition of internal independent subcompositions in linear regression models. We introduced a new norm for CoDa, the norm L^1 pairwise logratio. This norm verifies the desirable properties, such as scale invariance and subcompositional dominance, in order to be coherent with the Aitchison geometry. By using the norm L^1 pairwise logratio in a Lasso regression model, we can determine the importance of the relative information between the compositional parts in explaining a response variable. We use this information to create an olr-basis taking into account the structure defined by the internal independent subcompositions. The basis created includes linking balances between the internal independent subcompositions and the rest of the parts of the composition. We test the linking balances for analysing the subcompositional external independence because in such a case, the parts involved can be removed from the regression model. In other words, because the method identifies the pairwise logratios and balances that are less relevant, we can simplify the model and improve its interpretation while maintaining a high level of predictive power. This methodology provides a more nuanced understanding of the relationship between the compositional parts, which can lead to better insights and decision-making in various fields. Still pending is analyzing how one can improve the model when introducing a penalty term based on a convex linear combination of the norm L^1-plr and the Aitchison norm (Elastic net). The development of these types of models is one of the more interesting challenges in current CoDa analysis. Moreover, in order to implement the algorithm on high-dimensional data sets is necessary to increase the speed when fitting the model. One option to explore could be to change the ADMM algorithm by a faster one.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was supported by the Ministerio de Ciencia e Innovación under the project "CODA-GENERA" (Ref. PID2021-123833OB-I00) and the grant PRE2019-090976; and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project "COSDA" (Ref. 2021SGR01197).

Appendix A. Proofs

Proposition. 2. *$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right|$ verifies the properties of a norm.*

Proof.
- Positive definiteness: $\forall \mathbf{x} \in \mathscr{S}^D,\ \| \mathbf{x}\|_{1-plr} \geq 0$. Moreover, $\| \mathbf{x}\|_{1-plr} = 0$ if and only if $\mathbf{x} = (1, \ldots, 1)$. Immediate from definition 1.
- Absolute homogeneity: $\forall \mathbf{x} \in \mathscr{S}^D$ and $\forall \lambda \in \mathbb{R},\ \| \lambda \odot \mathbf{x}\|_{1-plr} = |\lambda| \| \mathbf{x}\|_{1-plr}$. Immediate from definition 1.
- Subadditivity: $\forall \mathbf{x},\ \mathbf{y} \in \mathscr{S}^D,\ \| \mathbf{x} \oplus \mathbf{y}\|_{1-plr} \leq \| \mathbf{x}\|_{1-plr} + \| \mathbf{y}\|_{1-plr}$.

Because the absolute value is a convex function, for all $i,\ j$, it verifies $\left|ln\left(\frac{x_iy_i}{x_jy_j}\right)\right| = \left|ln\left(\frac{x_i}{x_j}\right) + ln\left(\frac{y_i}{y_j}\right)\right| \leq \left|ln\left(\frac{x_i}{x_j}\right)\right| + \left|ln\left(\frac{y_i}{y_j}\right)\right|$. Thus, $\frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_iy_i}{x_jy_j}\right)\right| \leq \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| + \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{y_i}{y_j}\right)\right|$. □

Proposition. 3. *The L^1-plr norm on $\mathscr{S}^D$, $\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right|$ verifies the properties scale invariance, permutation invariance, and subcompositional dominance.*

Proof. .

- Scale invariance: $\| \mathbf{x}\|_{1-plr} = \| \lambda\mathbf{x}\|_{1-plr}$. Immediate from definition 1.
- Permutation invariance: $\| (x_1, \ldots, x_i, \ldots, x_j, \ldots, x_D)\|_{1-plr} = \| (x_1, \ldots, x_j, \ldots, x_i, \ldots, x_D)\|_{1-plr}$. Immediate from definition 1.
- Subcompositional dominance: $\| \mathbf{x}\|_{1-plr} \geq \| \mathrm{sub}(\mathbf{x})\|_{1-plr}$.

Without loss of generality we will prove that $\| (x_1, \ldots, x_D)\|_{1-plr} \geq \| (x_2, \ldots, x_D)\|_{1-plr}$.

$$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| = \frac{1}{D-1}\left(\sum_{j=2}^{D}\left|ln\left(\frac{x_1}{x_j}\right)\right| + \sum_{2\leq i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right|\right).$$

We write the first summation, $\sum_{j=2}^{D}\left|ln\left(\frac{x_1}{x_j}\right)\right|$, in a double summation form:

$$\sum_{j=2}^{D}\left|ln\left(\frac{x_1}{x_j}\right)\right| = \frac{1}{D-2}\sum_{j=2}^{D}(D-2)\left|ln\left(\frac{x_1}{x_j}\right)\right| = \frac{1}{D-2}\sum_{j=2}^{D}\sum_{k=3}^{D}\left|ln\left(\frac{x_1}{x_k}\right)\right| =$$

$$\frac{1}{D-2}\sum_{2\le j<k}\left(\left|ln\left(\frac{x_1}{x_j}\right)\right| + \left|ln\left(\frac{x_1}{x_k}\right)\right|\right)$$

Renaming index j by i, and index k by j in the above summation, we can write the L^1-plr norm as follows:

$$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{2\le i<j}\left(\frac{1}{D-2}\left|ln\left(\frac{x_1}{x_i}\right)\right| + \frac{1}{D-2}\left|ln\left(\frac{x_1}{x_j}\right)\right| + \left|ln\left(\frac{x_i}{x_j}\right)\right|\right) \ge$$

$$\frac{1}{D-1}\sum_{2\le i<j}\left(\frac{1}{D-2}\left|ln\left(\frac{x_i}{x_j}\right)\right| + \left|ln\left(\frac{x_i}{x_j}\right)\right|\right) = \frac{1}{D-1}\sum_{2\le i<j}\frac{D-1}{D-2}\left|ln\left(\frac{x_i}{x_j}\right)\right|.$$

Therefore,

$$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| \ge \frac{1}{D-1}\frac{D-1}{D-2}\sum_{2\le i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| = \| (x_2,\ldots,x_D)\|_{1-plr}.$$

Note the importance of using the factor $\frac{1}{D-1}$ in the norm L^1-plr instead of the factor $\frac{1}{D}$ used in the Aitchison norm. □

Proposition. . *For all* $\mathbf{x}\in\mathscr{S}^D$*, it holds that* $\| \mathbf{x}\|_{1-\text{plr}} \le \| \mathbf{x}\|_{1-\text{clr}}$.

Proof. .

$$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i/g(\mathbf{x})}{x_j/g(\mathbf{x})}\right)\right| = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{g(\mathbf{x})}\right) - ln\left(\frac{x_j}{g(\mathbf{x})}\right)\right| \le$$

$$\frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{g(\mathbf{x})}\right)\right| + \left|ln\left(\frac{x_j}{g(\mathbf{x})}\right)\right|$$

For each $k=1,\ldots,D$, the term $ln\left(\frac{x_k}{g(\mathbf{x})}\right)$ appears $D-1$ times in the last summation, then it holds that

$$\frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{g(\mathbf{x})}\right)\right| + \left|ln\left(\frac{x_j}{g(\mathbf{x})}\right)\right| = \| \mathbf{x}\|_{1-\text{clr}}.$$

Therefore, it holds that $\mathbf{x}\in\mathscr{S}^D$, $\| \mathbf{x}\|_{1-\text{plr}} \le \| \mathbf{x}\|_{1-\text{clr}}$.

Note that for any $\mathbf{x}\in\mathscr{S}^D$ that $\text{clr}(\mathbf{x}) = \left(0,\ldots,0,\underbrace{\pm\frac{a}{2}}_{i},0,\ldots,0,\underbrace{\mp\frac{a}{2}}_{j},0,\ldots,0\right)$ then it holds that $\| \mathbf{x}\|_{1-\text{plr}} = \| \mathbf{x}\|_{1-\text{clr}} = a$. □

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London. Reprinted 2003 with additional material by The Blackburn Press, London, UK.

Aitchison, J., Bacon-Shone, J., 1984. Log contrast models for experiments with mixtures. Biometrika 71, 323–330.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Math. Geol. 32, 271–275.

Barceló-Vidal, C., Martn-Fernández, J.A., 2016. The mathematics of compositional analysis. Austrian J. Stat. 45, 57–71.

Bates, S., Tibshirani, R., 2019. Log-ratio lasso: scalable, sparse estimation for log-ratio models. Biometrics 75, 613–624.

Billheimer, D., Guttorp, P., Fagan, W.F., 2001. Statistical interpretation of species composition. J. Am. Stat. Assoc. 96, 1205–1214. https://doi.org/10.1198/016214501753381850 arXiv:https://doi.org/10.1198/016214501753381850.

Boogaart, K.G.v.d., Tolosana, R., 2013. Analyzing Compositional Data with R. Use R! Springer.

Boogaart, K., Filzmoser, P., Hron, K., Templ, M., Tolosana-Delgado, R., 2021. Classical and robust regression analysis with compositional data. Math. Geosci. 53, 823–858.

Buccianti, A., Grunsky, E., 2014. Compositional data analysis in geochemistry: are we sure to see what really occurs during natural processes? J. Geochem. Explor. 141 https://doi.org/10.1016/j.gexplo.2014.03.022.

Calle, M., Susin, A., 2022a. coda4microbiome: compositional data analysis for microbiome studies. bioRxiv doi:https://doi.org/10.1101/2022.06.09.495511.

Calle, M., Susin, A., 2022b. Identification of dynamic microbial signatures in longitudinal studies. bioRxiv doi:https://doi.org/10.1101/2022.04.25.489415.

Calle, M., Pujolassos, M., Susin, A., 2023. coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. BMC Bioinform. https://doi.org/10.1186/s12859-023-05205-3.

Coenders, G., Greenacre, M., 2022. Three approaches to supervised learning for compositional data with pairwise logratios. J. Appl. Stat. https://doi.org/10.1080/02664763.2022.2108007.

Coenders, G., Pawlowsky-Glahn, V., 2020. On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT-Stat. Oper. Res. Trans. 44, 201–220. URL: https://raco.cat/index.php/SORT/article/view/371189 https://doi.org/10.2436/20.8080.02.100.

Comas-Cufí, M., 2022. coda.base: A Basic Set of Functions for Compositional Data Analysis. URL: https://CRAN.R-project.org/package=coda.base. r package version 0.5.2.

Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37, 795–828.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.

Gordon-Rodriguez, E., Quinn, T.P., Cunningham, J.P., 2022. Learning sparse log-ratios for high-throughput sequencing data. Bioinformatics 38, 157–163.

Hesterberg, T., Moore, D.S., Monaghan, S., Clipson, A., Epstein, R., Craig, B.A., McCabe, G., 2012. Bootstrap Methods and Permutation Tests, 7th ed. W. H. Freeman, New York, p. 657. Chapter 16 of Introduction to the Practice of Statistics.

Hron, K., Filzmoser, P., Thompson, K., 2012. Linear regression with compositional explanatory variables. J. Appl. Stat. 39, 1–14. https://doi.org/10.1080/02664763.2011.644268.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. Introduction to Statistical Learning, 2nd edition. Springer, New York.

Lin, W., Shi, R., Feng, R., Li, H., 2014. Variable selection in regression with compositional covariates. Biometrika 101, 785–797.

Lu, J., Shi, P., Li, H., 2019. Generalized linear models with linear constraints for microbiome compositional data. Biometrics 75, 235–244.

Martín-Fernández, J.A., 2019. Comments on: compositional data: the sample space and its structure. TEST 28, 653–657.

Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R., 2018. Advances in principal balances for compositional data. Math. Geosci. 50, 273–298.

Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2011. The Principle of Working on Coordinates. Chapter 3 of Compositional Data Analysis: Theory and Applications, pp. 29–42. https://doi.org/10.1002/9781119976462.ch3.

Monti, G., Filzmoser, P., 2021. Sparse least trimmed squares regression with compositional covariates for high-dimensional data. Bioinformatics 37, 3805–3814.
Monti, G., Filzmoser, P., 2022. Robust logistic zero-sum regression for microbiome compositional data. ADAC 16, 301–324.
Nesrstová, V., Wilms, I., Palarea-Albaladejo, J., Filzmoser, P., Martín-Fernández, J., Friedecký, D., Hron, K., 2023. Principal balances of compositional data for regression and classification using partial least squares. J. Chemom., e3518 https://doi.org/10.1002/cem.3518.
Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stoch. Env. Res. Risk A. 15, 384–398. https://doi.org/10.1007/s004770100077.
R-Core-Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), 2014a. Chemistry of Europe's Agricultural Soils—Part A: Methodology and Interpretation of the GEMAS Data Set, Geologisches Jahrbuch (Reihe B 102). Schweizerbarth, Hannover.
Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), 2014b. Chemistry of Europe's Agricultural Soils—Part B: General Background Information and Further Analysis of the GEMAS Data Set, Geologisches Jahrbuch (Reihe B 103). Schweizerbarth, Hannover.
Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., Calle, M.L., 2018. Balances: a new perspective for microbiome analysis. MSystems 3, e00053–18.
Saperas-Riera, J., Martín-Fernández, J., Mateu-Figueras, G., 2023. Fundamentals of convex optimization for compositional data. SORT-Stat. Oper. Res. Trans. 47.
Shi, P., Zhang, A., Li, H., 2016. Regression analysis for microbiome compositional data. Ann. Appl. Stat. 10, 1019–1040.
Susin, A., Wang, Y., Lê Cao, K.A., Calle, M.L., 2020. Variable selection in microbiome compositional data analysis. NAR Genom. Bioinform. 2, lqaa029.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
You, K., Zhu, X., 2021. ADMM: Algorithms Using Alternating Direction Method of Multipliers. URL: https://CRAN.R-project.org/package=ADMM. r package version 0.3.3.

4.3 Mathematics

El tercer article adapta amb rigor la definició de norma L^p a l'espai composicional (O2 2.1) a través de l'isomorfisme logarítmic. Seguidament s'estudien les propietats de les normes en coherència amb la geometria composicional. En el cas particular de les normes L^1-CoDA, L^2-CoDA i L^∞-CoDa es reescriuen i s'interpreten en termes de balanços. Finalment es condueix un procés de regularització LASSO amb tres normes diferents: L^1-CoDA, L^1-*plr* i L^1-clr, cosa que permet una comparació de l'impacte sobre la simplificació dels respectius models lineals. Els diferents processos de regularització parteixen d'un problema d'optimització del tipus LASSO generalitzat (O3 2.1).

L'article ha estat publicat a la revista Mathematics.
Volum: 12, número: 9, pàgines: -
Enviat: gener 2024, acceptat: abril 2024
DOI: 10.3390/math12091388
Factor d'impacte: 2.3 (Q1).

mathematics

MDPI

Article

L^p-Norm for Compositional Data: Exploring the CoDa L^1-Norm in Penalised Regression

Jordi Saperas-Riera, Glòria Mateu-Figueras and Josep Antoni Martín-Fernández *

Department of Computer Science, Applied Mathematics and Statistics, University of Girona, 17003 Girona, Spain; jordi.saperas@udg.edu (J.S.-R.); gloria.mateu@udg.edu (G.M.-F.)
* Correspondence: josepantoni.martin@udg.edu

Abstract: The Least Absolute Shrinkage and Selection Operator (LASSO) regression technique has proven to be a valuable tool for fitting and reducing linear models. The trend of applying LASSO to compositional data is growing, thereby expanding its applicability to diverse scientific domains. This paper aims to contribute to this evolving landscape by undertaking a comprehensive exploration of the L^1-norm for the penalty term of a LASSO regression in a compositional context. This implies first introducing a rigorous definition of the compositional L^p-norm, as the particular geometric structure of the compositional sample space needs to be taken into account. The focus is subsequently extended to a meticulous data-driven analysis of the dimension reduction effects on linear models, providing valuable insights into the interplay between penalty term norms and model performance. An analysis of a microbial dataset illustrates the proposed approach.

Keywords: Aitchison's geometry; compositional data; L^p-norm; balance selection

MSC: 62J07; 62P10; 62H99

check for updates

Citation: Saperas-Riera, J.; Mateu-Figueras, G.; Martín-Fernández, J.A. L^p-Norm for Compositional Data: Exploring the CoDa L^1-Norm in Penalised Regression. *Mathematics* **2024**, *12*, 1388. https://doi.org/10.3390/math12091388

Academic Editors: Antonio Di Crescenzo and Francisco Chiclana

Received: 29 January 2024
Revised: 5 April 2024
Accepted: 29 April 2024
Published: 1 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Linear regression serves as a powerful framework for modelling relationships between variables, as it aims to capture the underlying patterns that govern the variability in the response variable. For instance, in the microbiome domain, there is a particular interest in identifying which taxa are associated with a variable of interest, for example, the inflammatory parameter sCD14. To address such complex problems, adopting LASSO regression methods [1] has emerged as a popular choice for variable selection. LASSO regression strategically applies the Euclidean L^1-norm penalisation to the model coefficients, wherein the L^1-norm represents the sum of the absolute values of these coefficients. The penalised term shrinks some regression parameters towards zero, facilitating variable selection.

While conventional regression models assume independence among covariates, this assumption fails when dealing with compositional explanatory variables. These variables are called *parts* of a whole, and are usually expressed in proportions, percentages, or ppm. Historically [2], the sample space of the compositional data (CoDa) is designed as the D-part unit simplex $\mathcal{S}^D = \{\mathbf{x} \in \mathbb{R}^D : x_j > 0; \sum x_j = 1; j = 1, \ldots, D\}$. The fundamental idea in the analysis of CoDa is that the information is relative, and is primarily contained in the ratios between parts, not the absolute amounts of the parts. Therefore, the use of log-ratios is advocated. The analysis of CoDa, pioneered by [2], has witnessed increasing significance across such diverse fields as environmental science, geochemistry, microbiology, and economics. However, the integration of CoDa as covariates in regression models introduces particular challenges. The existing literature addresses these challenges, providing methodologies for regression model simplification with CoDa. The first works on penalised regression with compositional covariates [3–6] restricted the Euclidean L^1-norm on the centered log-ratio (clr) subspace when defining the penalty term. Saperas et al. [7]

Mathematics **2024**, *12*, 1388

introduced a new norm, called the pairwise log-ratio (L^1-plr), as a part of a methodology on penalised regression to simplify the log-ratios on the explanatory side of the model. These log-ratios are also known as balances [8].

The primary objective of this article lies in comprehensive comparison of the effects of different norms on the penalty term within LASSO regression with different compositional explanatory variables. The choice of a norm in the penalty term is a pivotal aspect that significantly influences the regularization mechanism, and consequently the characteristics of the resulting models. To accomplish this, a precise and rigorous definition of the induced L^p-norms for CoDa (CoDa L^p-norms) within the compositional space is necessary. A comparison between the CoDa L^1-norm and other norms for compositions established in the literature is provided. Through this analysis, we seek to contribute valuable insights into the characteristics and implications of these norms in penalised regression.

The rest of this article is organised as follows. In Section 2, fundamental concepts related to the geometric structure of CoDa are outlined. In addition, some popular measures of central tendency are written as the solution of a variational problem using L^p-norms in real space. Section 3 is devoted to defining the CoDa L^p-norms on the compositional space. In Section 4, after describing the basic concepts of standard penalty regression, we analyse LASSO regression with compositional covariates using three different L^1-norms in the penalty term, with the CoDa L^1-norm among them. A comparison of the different norms is illustrated in Section 5 using a microbiome dataset. Finally, Section 6 concludes with some closing remarks.

2. Some Basic Concepts

2.1. Elements of the Aitchison Geometry

CoDa conveys relative information because the variables describe relative contributions to a given total [2]. The formal geometrical framework for the analysis of CoDa, coined the *Aitchison geometry*, first appeared in [9,10]. The Aitchison geometry is based on two specific operations on $\mathcal{S}^D$, called *perturbation* and *powering*, respectively defined as $\mathbf{x} \oplus \mathbf{y} = (x_1y_1, x_2y_2, \ldots, x_Dy_D)$ and $\alpha \odot \mathbf{x} = (x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)$ for $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$. In order to interpret the results of these operations, one can perform *closure* on the result, that is, normalise the resulting vector to a unit sum by dividing each component by its total sum. Note that the closure operation provides a compositionally equivalent vector. With a vector space structure, a metric structure can be easily defined using the clr-scores of a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)$ [2]:

$$\operatorname{clr}(\mathbf{x}) = (\operatorname{clr}(\mathbf{x})_1, \ldots, \operatorname{clr}(\mathbf{x})_D) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})} \right),$$

where $g(\cdot)$ is the geometric mean of the composition. Note that clr-scores are collinear, because it holds that $\sum_{j=1}^{D} \operatorname{clr}(\mathbf{x})_j = 0$.

The basic metric elements of the Aitchison geometry are the inner product ($< \cdot, \cdot >_{\mathcal{A}}$), L^2-norm ($|| \cdot ||_{\mathcal{A}}$), and distance ($d_{\mathcal{A}}(\cdot, \cdot)$), defined as follows:

$$< \mathbf{x}, \mathbf{y} >_{\mathcal{A}} = < \operatorname{clr}(\mathbf{x}), \operatorname{clr}(\mathbf{y}) >_E , \quad \|\mathbf{x}\|_{\mathcal{A}}^2 = < \mathbf{x}, \mathbf{x} >_{\mathcal{A}} , \quad d_{\mathcal{A}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{\mathcal{A}}, \text{ for } \mathbf{x}, \mathbf{y} \in \mathcal{S}^D, \tag{1}$$

where "$\mathcal{A}$" means the Aitchison geometry, "E" the typical Euclidean geometry, and "$\ominus$" the perturbation difference $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$. Log-ratios, like clr-scores, have become a cornerstone of CoDa analysis; nevertheless, in the literature the concept of balance between two non-overlapping groups of parts is frequently used. A balance is defined as the log-ratio between the geometric means of the parts within each group multiplied by a constant that depends on the number of parts in each group [8].

An important scale-invariant function is the *log-contrast*, which plays the typical role of the linear combination of variables. Given a D-part composition $\mathbf{x}$, a log-contrast is defined as any linear combination of the logarithms of the compositional parts

$$\sum_{j=1}^{D} a_j \ln x_j, \quad \text{with} \quad \sum_{j=1}^{D} a_j = 0, \quad a_j \in \mathbb{R}.$$

Given a dependent variable y and an explanatory D-part composition $\mathbf{x}$, the definition of a linear regression model in terms of a log-contrast [11] is

$$y = \alpha_0 + \sum_{j=1}^{D} \alpha_j \ln x_j, \quad \text{with} \quad \sum_{j=1}^{D} \alpha_j = 0, \quad \alpha_j \in \mathbb{R}, \tag{2}$$

whereas in terms of metric elements the model formulation [12] is

$$y = \beta_0 + \langle \boldsymbol{\beta}, \mathbf{x} \rangle_{\mathcal{A}} = \beta_0 + \langle \operatorname{clr}(\boldsymbol{\beta}), \operatorname{clr}(\mathbf{x}) \rangle_E, \tag{3}$$

where $\boldsymbol{\beta}$ is the compositional gradient vector. The expressions in both Equations (2) and (3) are equivalent when one considers $\alpha_0 = \beta_0$ and $\boldsymbol{\alpha} = \operatorname{clr}(\boldsymbol{\beta})$. Because the sum of the clr-scores is zero ($\sum_{j=1}^{D} \operatorname{clr}(\beta)_j = \sum_{j=1}^{D} \operatorname{clr}(\mathbf{x})_j = 0$), the inner product of clr transformed vectors is equal to

$$\langle \operatorname{clr}(\boldsymbol{\beta}), \operatorname{clr}(\mathbf{x}) \rangle_E = \langle \operatorname{clr}(\boldsymbol{\beta}), \ln(\mathbf{x}) \rangle_E = \langle \ln \boldsymbol{\beta}, \operatorname{clr}(\mathbf{x}) \rangle_E. \tag{4}$$

For simplicity and to avoid overloading the notation, we denote $\boldsymbol{\beta}^* = \ln \boldsymbol{\beta}$, and write the linear regression model in terms of the Euclidean inner product as $y = \beta_0 + \langle \boldsymbol{\beta}^*, \operatorname{clr}(\mathbf{x}) \rangle_E$.

2.2. Norms and Measures of Central Tendency

The most popular measures of the central tendency of a real variable are the median and the arithmetic mean. Both can be defined as solving a variational problem [13]; indeed, the median $Med(\mathbf{z})$ of a dataset $\mathbf{z} = \{z_1, \ldots, z_D | \, z_i \in \mathbb{R}\}$ is the value that minimises the average absolute deviation $Med(\mathbf{z}) = \arg\min_{\lambda} \frac{1}{D} \sum_{j=1}^{D} |z_j - \lambda|$. The arithmetic mean $\overline{\mathbf{z}}$ of a dataset $\mathbf{z} = \{z_1, \ldots, z_D | \, z_i \in \mathbb{R}\}$ is the value that minimises the mean squared deviation $\overline{\mathbf{z}} = \arg\min_{\lambda} \frac{1}{D} \sum_{j=1}^{D} (z_j - \lambda)^2$. In addition, the mid-range $MR(\mathbf{z})$ of a dataset $\mathbf{z}$ is the value that minimises the maximum absolute deviation $MR(\mathbf{z}) = \arg\min_{\lambda}(\max_{j} |z_j - \lambda|)$. These definitions can be generalised to any L^p-norm [13] (Chapter 3).

Definition 1. *Let $\mathbf{z} = \{z_1, \ldots, z_D\}$ be a dataset and let $p \geq 1$; furthermore, let μ_p be the **p-measure of central tendency** that minimises the total p-deviation function $TD_p(\lambda)$:*

$$\mu_p = \arg\min_{\lambda} \left\{ TD_p(\lambda) = \|\mathbf{z} - \boldsymbol{\Lambda}\|_p^p = \frac{1}{D} \sum_{j=1}^{D} (z_j - \lambda)^p \right\},$$

where $\boldsymbol{\Lambda} = (\lambda, \ldots, \lambda)$, $\lambda \in \mathbb{R}$.

With this definition, the median ($\mu_1 = Med(\mathbf{z})$), arithmetic mean ($\mu_2 = \overline{\mathbf{z}}$), and mid-range ($\mu_\infty = MR(\mathbf{z})$) follow as special cases for the norms L^1, L^2, and L^∞, respectively.

Remark 1. *Convexity of the total p-deviation function $TD_p(\lambda)$:*

- *For $p = 1$, the average absolute deviation $TD_1(\lambda)$ is a convex function of λ; however, it is not strictly convex. Thus, the median may be a non-unique value.*

- *For $p > 1$, the total p-deviation function $TD_p(\lambda)$ is strictly convex; thus, if $\mu_p(\mathbf{Z})$ exists, this is unique.*

3. L^p-Norms on the Compositional Space

To define induced L^p-norms on the compositional space (CoDa L^p-norms) in a compatible way with the Aitchison geometry, one must capture the geometric structure of the $\mathcal{S}^D$ [14]. To achieve this objective, we initially define the induced L^p-norm within the quotient space $\mathcal{L}^D = \{\mathbf{z} + \lambda \mathbf{1}_D | \, \mathbf{z} \in \mathbb{R}^D, \, \mathbf{1}_D = (1, \ldots, 1)\}$. Following Brezis [15] (Chapter 11.2), an induced L^p-norm on the quotient space $\mathcal{L}^D$ can be defined by inducing the Euclidean L^p-norm in $\mathbb{R}^D$ on $\mathcal{L}^D$. The underlying idea is to assign to an equivalence class the minimum value of the L^p-norm among the elements belonging to the same equivalence class.

Definition 2. *Let $\mathbf{z} \in \mathcal{L}^D$ be a log-composition. The* **induced L^p-norm**, *denoted by $\|\mathbf{z}\|_{p,\mathcal{L}^D}$, is*

$$\|\mathbf{z}\|_{p,\mathcal{L}^D} = \min_{\lambda} \|\mathbf{z} + \lambda \mathbf{1}_D\|_p,$$

where $\mathbf{1}_D = (1, \ldots, 1)$ and $\|\cdot\|_p$ denotes the typical L^p-norm in the real space.

Using the logarithmic isomorphism [14], the L^p-norm can be extended to the compositional space.

Definition 3. *Let $\mathbf{x} \in \mathcal{S}^D$ be a composition. The* **CoDa L^p-norm**, *denoted by $\|\mathbf{x}\|_{p,\mathcal{S}^D}$, is*

$$\|\mathbf{x}\|_{p,\mathcal{S}^D} = \|\ln \mathbf{x}\|_{p,\mathcal{L}^D} = \min_{\lambda} \|\ln \mathbf{x} + \lambda \mathbf{1}_D\|_p.$$

Proposition 1. *The CoDa L^p-norm on $\mathcal{S}^D$ is $\|\mathbf{x}\|_{p,\mathcal{S}^D}$, and verifies the properties of the Aitchison geometry [16]:*

- *Scale invariance: $\|\mathbf{x}\|_{p,\mathcal{S}^D} = \|k\mathbf{x}\|_{p,\mathcal{S}^D}$, $k > 0$.*
- *Permutation invariance: $\|(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_D)\|_{p,\mathcal{S}^D} = \|(x_1, \ldots, x_j, \ldots, x_i, \ldots, x_D)\|_{p,\mathcal{S}^D}$.*
- *Subcompositional dominance: $\|\mathbf{x}\|_{p,\mathcal{S}^D} \geq \|\text{sub}(\mathbf{x})\|_{p,\mathcal{S}^d}$, where $\text{sub}(\mathbf{x}) \in \mathcal{S}^d$ denotes any subset formed by d parts of $\mathbf{x}$.*

Proof of Proposition 1. The proof directly follows from the Definition 3. □

Following Definition 3 and the measures of central tendency described in Section 2.2, the CoDa L^p-norms L^1, L^2, and L^∞ can be developed:

- The CoDa L^1-norm on $\mathcal{S}^D$ is

$$\|\mathbf{x}\|_{1,\mathcal{S}^D} = \|\ln \mathbf{x} - Med(\ln \mathbf{x})\mathbf{1}_D\|_1 = \|\ln \frac{\mathbf{x}}{Med(\mathbf{x})}\|_1 = \sum_{j=1}^{D} \left| \ln \frac{x_j}{Med(\mathbf{x})} \right|,$$

  where $Med(\ln \mathbf{x})$ and $Med(\mathbf{x})$ are the median of the sets $\{\ln x_1, \ldots, \ln x_D\}$ and $\{x_1, \ldots, x_D\}$, respectively. As the logarithm function is strictly increasing, as per Definition 1, the set of points that serve as solutions to the variational problem TD_1 when applied to log-transformed values $\mu_1 = Med(\ln(x))$ precisely corresponds to the log-transformed set of points that are solutions to the variational problem TD_1 when applied to the raw data, that is, $\ln(Med(x))$.
  Wu et al. [17] proposed the median of a D-part composition as an alternative denominator to the geometric mean in an attempt to extend the definition of clr-scores. In general, the performance of the median as a robust estimator of the midpoint of a dataset is better when the data have high asymmetry. The CoDa L^1-norm captures the distance between two points when movement is restricted to paths that run parallel to the clr-axes ($\ln(\frac{x_i}{g(\mathbf{x})})$), as is the case in a grid or city street network (Manhattan distance, Figure 1). The CoDa L^1-norm has an equivalent expression that captures

the information about the ratio between the components of a composition; indeed, the median is the central point that divides a set into two equal parts, with half of the values falling below this central position and half above it. Therefore, half of the log-ratios $\ln(\frac{x_j}{Med(\mathbf{x})})$ are positive and the other half are negative. If we rearrange the parts of a composition in increasing order (small to large), i.e., $x_{(1)} \leq \ldots \leq x_{(D)}$, then the CoDa L^1-norm can be written in the following manner:

* $\|\mathbf{x}\|_{1,\mathcal{S}^D} = \ln\left(\dfrac{x_{(n+1)} \cdot \ldots \cdot x_{(2n)}}{x_{(1)} \cdot \ldots \cdot x_{(n)}}\right)$ if $D = 2n$;
* $\|\mathbf{x}\|_{1,\mathcal{S}^D} = \ln\left(\dfrac{x_{(n+1)} \cdot \ldots \cdot x_{(2n-1)}}{x_{(1)} \cdot \ldots \cdot x_{(n-1)}}\right)$ if $D = 2n-1$.

Thus, the CoDa L^1-norm is a *balance* between the large parts and small parts.

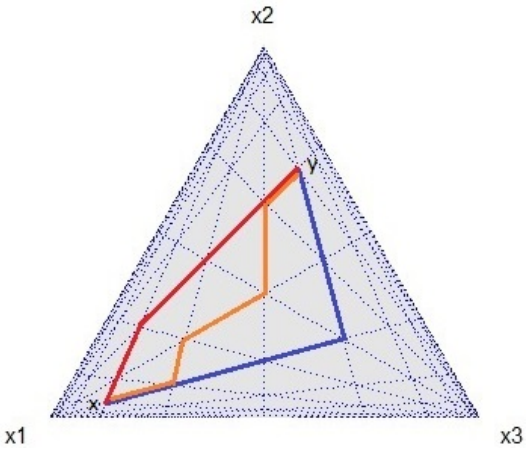


Figure 1. The Manhattan distance based on the CoDa L^1-norm in the simplex $\mathcal{S}^3$: the distance between two points $\mathbf{x} = \mathcal{C}[e^3, 1, e]$ and $\mathbf{y} = \mathcal{C}[1, e^2, e]$ in a grid-based system, where $\mathcal{C}[\cdot]$ is the closure operation, is represented by three paths (red, orange, and blue) of the same length (five units).

- The CoDa L^2-norm on $\mathcal{S}^D$ is

$$\|\mathbf{x}\|_{2,\mathcal{S}^D} = \|\ln\mathbf{x} - \overline{\ln(\mathbf{x})}\mathbf{1}_D\|_2 = \|\ln\mathbf{x} - \ln(g(\mathbf{x}))\mathbf{1}_D\|_2 = \left\|\ln\frac{\mathbf{x}}{g(\mathbf{x})}\right\|_2 = \left(\sum_{j=1}^{D}\left(\ln\frac{x_j}{g(\mathbf{x})}\right)^2\right)^{\frac{1}{2}},$$

where $g(\mathbf{x})$ is the geometric mean of the set $\{x_1, \ldots, x_D\}$. Because $\ln\frac{\mathbf{x}}{g(\mathbf{x})} \in$ clr-subspace, the CoDa L^2-norm is the restricted Euclidean L^2-norm on the clr-subspace. This norm is commonly referred to as Aitchison's norm $\|\mathbf{x}\|_{\mathcal{A}}$ [18].

- The CoDa L^∞-norm on $\mathcal{S}^D$ is

$$\|\mathbf{x}\|_{\infty,\mathcal{S}^D} = \|\ln\mathbf{x} - MR(\ln\mathbf{x})\mathbf{1}_D\|_\infty = \left\|\ln\frac{\mathbf{x}}{GR(\mathbf{x})}\right\|_\infty = \max_j\left\{\left|\ln\frac{x_j}{GR(\mathbf{x})}\right|\right\},$$

where $MR(\ln\mathbf{x})$ and $GR(\mathbf{x})$ are respectively the mid-range and geometric mid-range of the sets $\{\ln x_1, \ldots, \ln x_D\}$ and $\{x_1, \ldots, x_D\}$. Note that $MR(\ln(\mathbf{x})) = \ln(GR(\mathbf{x}))$; thus, $GR(\mathbf{x}) = \left(\max_i\{x_i\} \cdot \min_j\{x_j\}\right)^{\frac{1}{2}}$, $i, j = 1, \ldots, D$. The CoDa L^∞-norm can be interpreted as a form of log-pairwise, as the CoDa L^∞-norm represents half of the log-pairwise between the largest part against the smallest part. This log-pairwise is the greatest among all log-pairwise in the composition:

$$\|\mathbf{x}\|_{\infty,\mathcal{S}^D} = \frac{1}{2}\ln\left(\frac{\max\limits_i\{x_i\}}{\min\limits_j\{x_j\}}\right) = \frac{1}{2}\max_{i,j}\left\{\ln\frac{x_i}{x_j}\right\}.$$

4. Penalised Regression with a Compositional Covariate

The LASSO regression model is formulated as the combination of the L^2-norm cost function and the L^1-norm regularisation term [1]. For a real dataset $\mathbf{Z}$ with n observations and D predictors along with a real response vector $\mathbf{Y}$ of length n, the LASSO regression model can be formulated as follows:

$$\min\left\{\frac{1}{2}\|\mathbf{Y} - \beta_0 - \langle\boldsymbol{\beta},\ \mathbf{Z}\rangle_E\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right\} \tag{5}$$

where β_0 is the intercept, the vector $\boldsymbol{\beta}$ is the gradient, and λ is the penalty parameter that controls the amount of regularisation. Note that $\|\cdot\|_2$ and $\|\cdot\|_1$ refer to the Euclidean L^2 and L^1 norms in real space, respectively. For $\lambda = 0$, the LASSO regression model (Equation (5)) provides the classical least squares regression model. The larger the value of λ, the greater the number of coefficients in $\boldsymbol{\beta}$ that is forced to be zero. The *optimal* value of λ can be chosen based on cross-validation techniques and related methods [19].

In the case of CoDa, additional considerations must be taken into account in order to respect the compositional nature of both the covariate $\mathbf{X}$ and the intercept $\boldsymbol{\beta}$. In variable selection, [3,20] wrote the LASSO model in terms of $\boldsymbol{\beta}^* = \ln\boldsymbol{\beta}$ and the log-transformed data instead of the clr-scores; consequently, a linear constraint on the compositional gradient coefficient $\boldsymbol{\beta}^*$ is necessary:

$$\min\left\{\frac{1}{2}\|\mathbf{Y} - \beta_0 - \langle\boldsymbol{\beta}^*,\ \ln(\mathbf{X})\rangle_E\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1\right\}, \text{subject to } \sum_{j=1}^{D}\beta_j^* = 0. \tag{6}$$

Most of the literature addressing the topic of penalised regression with a compositional covariate has predominantly employed the Euclidean L^1-norm in the penalty term, leading to a clr-variable selection ([3–6,20–22]).

In Equation (6), the constraint $\sum_{j=1}^{D}\beta_j^* = 0$ can be incorporated in the minimising function. The constraint $\sum_{j=1}^{D}\beta_j^* = 0$ forces the $\boldsymbol{\beta}^*$ parameter to be an element in the clr-subspace. Therefore, per Equation (4), the inner product $\langle\boldsymbol{\beta}^*,\ \ln(\mathbf{X})\rangle_E$ is equivalent to $\langle\mathrm{clr}(\boldsymbol{\beta}),\ \mathrm{clr}(\mathbf{X})\rangle_E = \langle\boldsymbol{\beta},\ \mathbf{X}\rangle_{\mathcal{A}}$. Thus, the constrained LASSO (Equation (6)) is equivalent to the following definition.

Definition 4. *Given y_i, $i = 1,\ldots,n$, the sample of the response variable $\mathbf{X}$, and the $n \times D$ matrix whose rows $\mathbf{X}_i = (x_{i1},\ldots,x_{iD})$ for $i = 1,\ldots,n$ contain the compositional sample, the L^1-clr LASSO estimator is defined as*

$$\boldsymbol{\beta} \in \underset{\beta_0,\,\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\frac{1}{2}\|\mathbf{Y} - \beta_0 - \langle\boldsymbol{\beta},\ \mathbf{X}\rangle_{\mathcal{A}}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{1-clr}\right\}, \tag{7}$$

where $\|\boldsymbol{\beta}\|_{1-clr} = \sum_{j=1}^{D}|\mathrm{clr}(\boldsymbol{\beta})_j| = \sum_{j=1}^{D}\left|\ln\frac{\beta_j}{g(\boldsymbol{\beta})}\right|$.

The key innovation here is that the linear constraint becomes embedded in the penalty term through the L^1-clr norm. This change in approach is not merely an algebraic or formal change; rather it implies a deeper understanding of the variable selection process in CoDa. The penalty term imposes a constraint on the sum of the absolute values of clr-scores within the gradient vector $\boldsymbol{\beta}$. This constraint compels the model to shrink or eliminate certain clr-scores, effectively driving them to zero. Consequently, this results in a balance selection. Without loss of generality, let us assume that the balance $\ln\frac{\beta_1}{g(\boldsymbol{\beta})}$ is zero. This implies that the corresponding balance $\ln\frac{x_1}{g(\mathbf{x})}$ has no influence on the response variable y. Therefore,

the maximum variation in y is concentrated in the subspace orthogonal to the balance $\ln \frac{x_1}{g(\mathbf{x})}$, i.e., the subspace of balances among the subcomposition $(x_2, \ldots, x_D)$. This selective regularization process facilitates variable selection, as x_1 does not influence the response variable y.

In order to establish a unified framework, the generalised LASSO problem [23] can be adapted to penalised linear models with a compositional covariate:

$$\min \left\{ \frac{1}{2} \|\mathbf{Y} - \beta_0 - \langle \boldsymbol{\beta}^*, \operatorname{clr}(\mathbf{X}) \rangle_E \|_2^2 + \lambda \|\mathbf{D} \cdot \boldsymbol{\beta}^*\|_1 \right\}, \tag{8}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ respectively refer to the Euclidean L^2 and L^1 norms in real space. The generalised LASSO problem allows for a broader range of applications by considering a matrix $\mathbf{D}$ associated with the penalty term. The matrix $\mathbf{D}$ is related to the L^1-norm considered in the penalty term. The choice of one norm over another determines the type of regularization. Different models can be formulated within the framework of the generalised LASSO problem and addressed through convex optimization algorithms. Solving each of these different penalised regression models yields distinct coefficients, each characterised by unique properties. Indeed, Definition 4 can be expressed as a generalised LASSO problem in the following manner:

$$\boldsymbol{\beta}^* \in \underset{\beta_0,\, \boldsymbol{\beta}^*}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \langle \boldsymbol{\beta}^*, \operatorname{clr}(\mathbf{X})_i \rangle_E)^2 + \lambda \|\mathbf{C}_D \boldsymbol{\beta}^*\|_1 \right\}, \tag{9}$$

where $\boldsymbol{\beta}^* = \ln \boldsymbol{\beta}$ and $\mathbf{D} = \mathbf{C}_D$ is the centering matrix on the clr-subspace, with $\mathbf{C}_D \boldsymbol{\beta}^* = \boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}^*} \mathbf{1}_D$.

On the other hand, following [7], it is possible to consider the matrix $\mathbf{D}$ equal to $\mathbf{F}$, that is, the matrix associated with the linear transformation $F(\beta_1^*, \ldots, \beta_D^*) = \frac{1}{D-1}(\beta_1^* - \beta_2^*, \beta_1^* - \beta_3^*, \ldots, \beta_1^* - \beta_D^*, \beta_2^* - \beta_3^*, \ldots, \beta_2^* - \beta_D^*, \ldots, \beta_{D-1}^* - \beta_D^*)$. Note that $\beta_i^* - \beta_j^* = \ln\left(\frac{\beta_i}{\beta_j}\right)$, which is a log-pairwise. In this case, the penalty term in a generalised LASSO problem can be written as $\|\mathbf{F} \cdot \boldsymbol{\beta}^*\|_1$, meaning that the generalised LASSO problem results in the following:

$$\boldsymbol{\beta}^* \in \underset{\beta_0,\, \boldsymbol{\beta}^*}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \langle \boldsymbol{\beta}^*, \operatorname{clr}(\mathbf{X})_i \rangle_E)^2 + \lambda \|\mathbf{F} \boldsymbol{\beta}^*\|_1 \right\}. \tag{10}$$

The model can be defined as follows.

Definition 5. *Given y_i, $i = 1, \ldots, n$, the sample of the response variable $\mathbf{X}$, and the $n \times D$ matrix whose rows, $\mathbf{X}_i = (x_{i1}, \ldots, x_{iD})$ for $i = 1, \ldots, n$ contain the compositional sample, the L^1-plr LASSO estimator is defined as*

$$\boldsymbol{\beta} \in \underset{\beta_0,\, \boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{Y} - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{X} \rangle_{\mathcal{A}} \|_2^2 + \lambda \|\boldsymbol{\beta}\|_{1-plr} \right\}, \tag{11}$$

where $\|\boldsymbol{\beta}\|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| \ln\left(\frac{\beta_i}{\beta_j}\right) \right|$.

Importantly, because $\ln \frac{\beta_i}{\beta_j} = \operatorname{clr}(\boldsymbol{\beta})_i - \operatorname{clr}(\boldsymbol{\beta})_j$, the penalty term shrinks the absolute value of the differences of the clr-scores within the gradient vector, which forces some pairwise differences of clr-scores to be zero, i.e., it forces equality on some clr-scores. Therefore, each set of equal clr-scores defines a subcomposition with non-influential balances within its parts. This selective regularization process facilitates balanced selection [7].

Finally, using the CoDa L^1-norm introduced in Section 3, it is possible to define another generalised LASSO problem.

Mathematics **2024**, *12*, 1388

Definition 6. *Given y_i, $i = 1, \ldots, n$, the sample of the response variable* $\mathbf{X}$*, and the $n \times D$ matrix whose rows $\mathbf{X}_i = (x_{i1}, \ldots, x_{iD})$ for $i = 1, \ldots, n$ contain the compositional sample, the CoDa L^1-norm LASSO estimator is defined as*

$$\boldsymbol{\beta} \in \underset{\beta_0,\, \boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2} \| \mathbf{Y} - \beta_0 - \langle \boldsymbol{\beta},\, \mathbf{X} \rangle_{\mathcal{A}} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_{1,\mathcal{S}^D} \right\}, \tag{12}$$

where $\| \boldsymbol{\beta} \|_{1,\mathcal{S}^D} = \sum_{j=1}^{D} \left| \ln \frac{\beta_j}{Med(\boldsymbol{\beta})} \right|$.

In this case, the penalty term compels certain parts β_j to be equal to the median of the parts, ensuring equality among them in particular. Consequently, the effect produced is also a balance selection, as in the previous case; however, unlike the L^1-plr LASSO estimator, with the CoDa L^1-norm estimator there is only one set of equal clr-scores, and all non-influential balances belong to a single subcomposition.

As there is no algebraic formula to express the median, $Med(\boldsymbol{\beta})$, it is necessary to include a new variable $m \in \mathbb{R}$ in the penalty term when formulating the minimization problem (Equation (12)) as a generalised LASSO problem:

$$\boldsymbol{\beta}^* \in \underset{\beta_0,\, \boldsymbol{\beta}^*,\, m}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \langle \boldsymbol{\beta}^*,\, \operatorname{clr}(\mathbf{X})_i \rangle_E)^2 + \lambda \| \mathbf{G} \cdot (\boldsymbol{\beta}^*, m) \|_1 \right\}, \tag{13}$$

where $\boldsymbol{\beta}^* = \ln \boldsymbol{\beta}$, $\mathbf{D} = \mathbf{G}$, and the matrix associated with the transformation $G(\boldsymbol{\beta}^*, m) = (\boldsymbol{\beta}_1^* - m, \boldsymbol{\beta}_2^* - m, \ldots, \boldsymbol{\beta}_D^* - m)$.

5. Study Case

We used the microbial dataset analyzed in [24,25] to compare the different L^1-norms in a CoDa LASSO regression problem. The dataset, collected and explained in [24], comprises the compositions of $D = 60$ taxa spanning various taxonomic levels (e.g., g for genus, f for family, o for order, and k for kingdom) within a set of n =151 individuals. The dependent variable y is an inflammatory parameter, specifically, the levels of soluble CD14 (sCD14 variable) measured for each individual. These data are available in the R package [26]. An individual having a zero value recorded for some parts indicates that certain taxa were not detected. A zero value prevents the application of the log-ratio methodology. Following a more analogous procedure than in [26], the genus *Thalassospira*, *unclassified* genus of the class *Alphaproteobacteria*, and *unclassified* genus of the family *Porphyromonadaceae*, all with more than 80% of zeros, were removed. The rest of the zeros recorded in the remaining 57 taxa were replaced by a small value using an imputation method [27,28]. Because the zeros are of count type, it is appropriate to apply methods based on Dirichlet-multinomial duality [29].

To solve the convex optimizations problems in Equations (9), (10), and (13), we first select the optimal λ parameter for the penalised model by performing a ten-fold cross-validation. Each iteration involves dividing the data into ten equal parts, training the model on nine of them, and then evaluating it on the remaining part to produce the lowest Mean Squared Error (MSE). With the parameter λ selected, we proceeded to solve the optimization problem in order to find the parameters β_0 and $\boldsymbol{\beta}^*$. The CVXR package in R version 4.3.2 [30] offers an interface for defining and solving convex optimization problems. CVXR utilises a domain-specific language, making it user-friendly and allowing users to express optimization problems. The package supports various solvers, enabling users to choose the one that best suits their needs. In our case, we opted for the Operator Splitting Quadratic Program (OSQP). The OSQP is a solver for quadratic programming problems and employs an operator-splitting method [31]. This solver is highly efficient even in cases where the matrices are not full-rank, such as our situation, because the clr-scores are used. Referring to the procedure detailed above, we have outlined an Algorithm 1 for a generalised LASSO method below.

The algorithm is applied in the three cases discussed in the previous section, namely, when considering the three different L^1 norms in the penalty term, i.e., the L^1-clr (Definition 4), L^1-plr (Definition 5), and CoDa L^1-norms (Definition 6).

For the L^1-clr estimator, the LASSO regression algorithm is applied iteratively within the cross-validation framework. Figure 2 illustrates the cross-validated MSE across different λ values. The optimal λ is determined by selecting the point on the curve where the mean squared error is minimised: λ_{min} = 35,769.42.

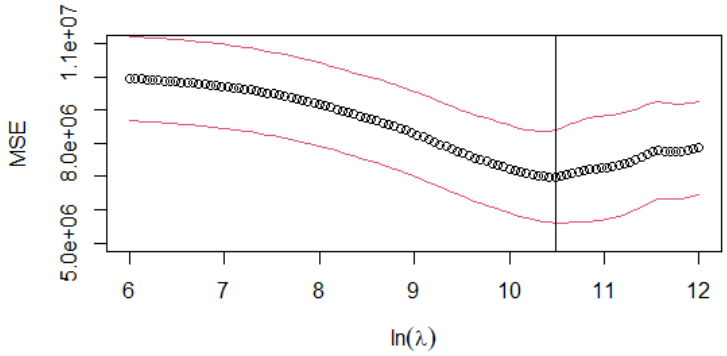


Figure 2. L^1-clr: cross-validation MSE curve for different log-transformed values of the penalty parameter ($\ln(\lambda)$). The circle ($\circ$) is the arithmetical mean of the ten-fold CV. The red lines (above and below the mean) indicate the mean $\pm$ stdev value, where stdev is the standard deviation of the ten-fold CV. The vertical line represents the log-transformed values of lambda.min = 35,769.42.

Algorithm 1 Generalised LASSO for CoDa

1. Fit the generalised LASSO model with tuning parameter λ (Equations (9), (10) or (13)).
2. Calculate the clr-representative: $\boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}^*}$
3. Express the generalised LASSO model in terms of clr-scores.

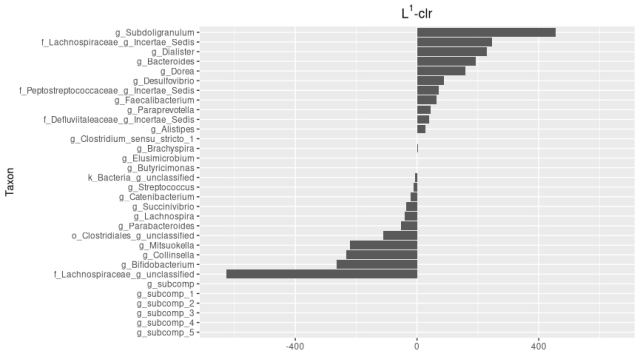
For λ_{min} = 35,769.42, the generalised LASSO (Equation (9)) identifies which ones among all the β_i^* are set to $\overline{\boldsymbol{\beta}^*}$, particularly ensuring equality among them. Importantly, when computing the representative $\boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}^*}\mathbf{1}_D \in$ clr-subspace, we find that some clr-scores are equal to zero. Therefore, the regularization process effectively splits the composition into two subcompositions. The first subcomposition represents the 33 non-influential parts, where coefficients $\text{clr}(\boldsymbol{\beta})_k$, $k = 1, \ldots, 33$ are driven to zero, contributing to model simplicity. The second subcomposition identifies the 24 parts that actively contribute to the influential balances on the response variable y (see Table A1). The intercept β_0 is equal to 6563.19. Figure 3a shows the non-zero clr-scores for parameter $\boldsymbol{\beta}$. We highlight that the most influential pairwise is formed by the genus *Subdoligranulum* and the unclassified genus of the family *Lachnospiraceae*.

For the CoDa L^1-norm estimator (Equation (13)), the LASSO regression algorithm is applied with the same cross-validation partition used in the L^1-clr estimator. Figure 4 illustrates the model's performance across different regularization parameters λ. The optimal value is λ_{min} = 45,582.21

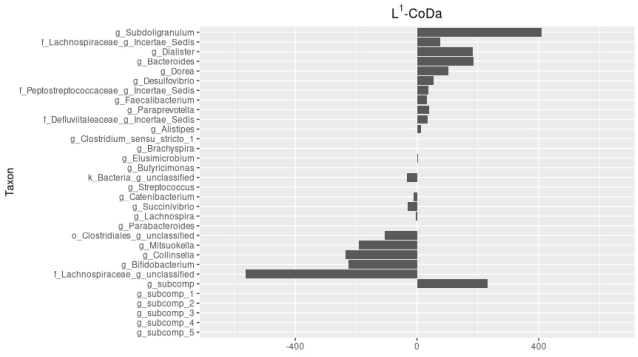
For λ_{min} = 45,582.21, the generalised LASSO (Equation (13)) identifies which ones among all the β_i^* are set to the median of $\boldsymbol{\beta}^*$, particularly ensuring equality among them. This equality among some $\boldsymbol{\beta}_i^*$ indicates that the balances involving their respective parts x_i have a non-influential role in the response variable y. However, in contrast to the L^1-clr scenario when computing the representative $\boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}^*}$, in general, all clr-scores are non-zero. Consequently, variable selection cannot be performed in this case. The regularization process effectively splits the composition into two subcompositions. The first subcomposition

Mathematics **2024**, *12*, 1388

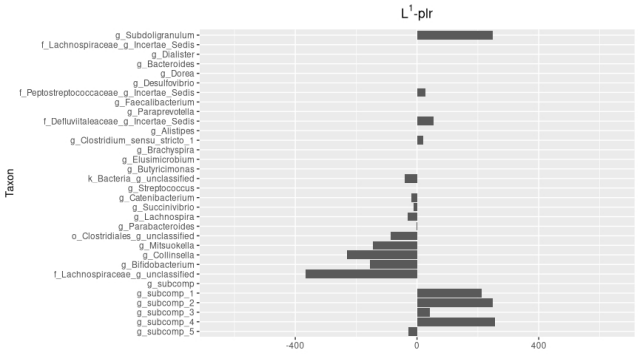
represents the 36 *internally independent* parts, that is, a subcomposition in which the balances between the respective parts do not influence y [32]. The coefficients $\text{clr}(\boldsymbol{\beta})_k$, $k = 1, \ldots, 36$ are driven to the median of $\boldsymbol{\beta}^*$ (6.44), contributing to model simplicity. The second subcomposition identifies the 21 parts that actively contribute to the influential balances on the response variable y (see Table A1).



(**a**)



(**b**)



(**c**)

Figure 3. Comparison of the $\text{clr}(\boldsymbol{\beta})$ parameter, with the taxon order maintained on the vertical axis to facilitate comparison: (**a**) clr-scores for the L^1-clr LASSO estimator, (**b**) clr-scores for the L^1-CoDa LASSO estimator, and (**c**) clr-scores for the L^1-plr LASSO estimator.

To highlight the model's simplicity, it is crucial to accurately summarise the information contained in the first subcomposition. Without loss of generality, let $(x_1, \ldots, x_k)$ be an internally independent subcomposition. The linear model in clr-scores is

$$y = \beta_0 + \sum_{j=1}^{k} \mathrm{clr}(\beta)_j \ln x_i + \sum_{j=k+1}^{D} \mathrm{clr}(\beta)_j \ln x_j,\ \mathrm{clr}(\beta)_1 = \ldots = \mathrm{clr}(\beta)_k.$$

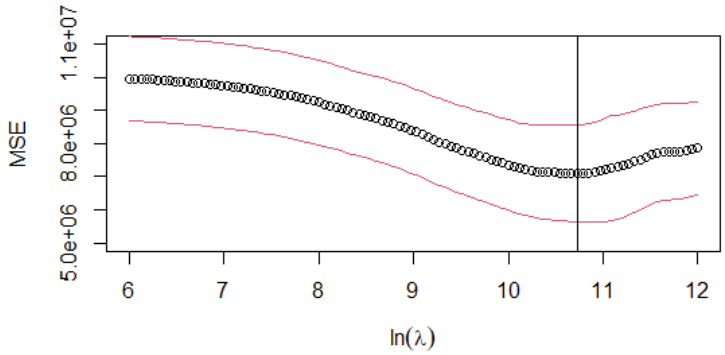


Figure 4. CoDa L^1-norm: cross-validation MSE curve for different log-transformed values of the penalty parameter ($\ln(\lambda)$). The circle ($\circ$) is the arithmetical mean of the ten-fold CV. The red lines (above and below the mean) represent the value mean $\pm$ stdev, where stdev is the standard deviation of the ten-fold CV. The vertical line represents the log-transformed values of lambda.min = 45,582.21.

As explained by [16] in Chapter 4, the best approach to represent a subcomposition is through its geometric mean, denoted as $g_{sub} = (x_1 \cdot \ldots \cdot x_k)^{\frac{1}{k}}$. Therefore, the linear model is

$$y = \beta_0 + k\mathrm{clr}(\beta_{sub}) \ln g_{sub} + \sum_{j=k+1}^{D} \mathrm{clr}(\beta)_j \ln x_j,$$

where $\mathrm{clr}(\beta_{sub}) = \mathrm{clr}(\beta)_1 = \ldots = \mathrm{clr}(\beta)_k$. This model has $D - k$ degrees of freedom, as opposed to the $D - 1$ degrees of freedom of the general linear model. The intercept value is $\beta_0 = 7023.879$, and Figure 3b shows the clr-scores of $\boldsymbol{\beta}$.

L^1-clr regularization creates a subcomposition that is both internally and externally independent [32], that is, both the balances within the parts of the subcomposition and the full balance between the parts of the subcomposition and the rest of the parts are all non-influential. In contrast, CoDa L^1-norm regularization relaxes the conditions and establishes only one subcomposition that is internally independent. In this context, the CoDa L^1-norm is somewhat more permissive. When comparing the results, we observe that both are quite similar; what stands out is the significance of the new variable g_{sub} in the CoDa L^1-norm penalised linear model. L^1-clr regularization eliminates the balance $\ln \frac{g_{sub}}{g(\mathbf{x})}$ without prior analysis. This observation prompts us to consider that the direct application of L^1-clr regularization might be premature. Furthermore, when dealing with a penalised model, it is always possible to subsequently test the nullity of any parameter [33].

L^1-clr and CoDa L^1-norm regularization share the fact that both shrink the difference between β_i^* coefficients and a central measure, respectively, the mean and the median; consequently, each regularization technique generates a unique subcomposition with certain properties related to its influence on the dependent variable y. Because the goal of a CoDa analysis is to describe the subcompositional structure of the data, the use of the L^1-clr and CoDa L^1-norms in the penalty term leads to a result that has to be considered as limited. To overcome this limitation, the L^1-plr norm enables the construction of more than one internally independent subcomposition, which can better capture the subcompositional structure of the data regarding the variable y [7].

Mathematics **2024**, *12*, 1388

With the same data partition as executed in previous cases, we performed cross-validation to find the optimal lambda value for the L^1-plr estimator. The optimal parameter is λ = 69,669.31 (Figure 5).

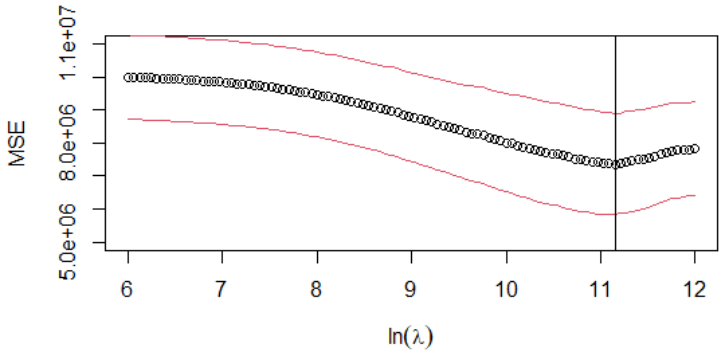


Figure 5. L^1-plr: cross-validation MSE curve for different log-transformed values of the penalty parameter ($\ln(\lambda)$). The circle ($\circ$) is the arithmetical mean of the ten-fold CV. The red lines (above and below the mean) represent the value mean $\pm$ stdev, where stdev is the standard deviation of the ten-fold CV. The vertical line represents the log-transformed values of lambda.min = 69,669.31.

For λ = 69,669.31, the generalised LASSO (Equation (13)) splits the composition into six distinct subcompositions: five internally independent subcompositions on response variable y and one subcomposition comprising 14 parts actively contributing to the influential balances on the response variable y (see Table A1 to compare L^1-plr estimator with L^1-clr and L^1-CoDa estimators, and Table A2 to explore its subcompositional structure). Each of the five internally independent subcompositions related to y contributes to reducing the dimension of the linear model. This reduction is achieved by substituting each subcomposition with its geometric mean ($g - subcomp_k$, $k = 1, \ldots, 5$), following the approach outlined in the CoDa L^1-norm estimator. The intercept value is $\beta_0 = 7023.879$ and Figure 3c shows the clr-scores of $\boldsymbol{\beta}$.

The L^1-plr estimator is the simplest and provides us with the most information about the subcompositional structure of the composition as regards the variable y.

6. Discussion

This paper has rigorously defined CoDa L^p-norms, providing a foundation for their application. The specific cases of the CoDa L^1, L^2, and L^∞ norms have been studied, interpreting these metrics in terms of log-ratios to enhance the reader's understanding. Additionally, a unified treatment of three distinct L^1-norms tailored for compositional data has been presented in the context of a generalised LASSO problem. Through a detailed examination of the regularization effects of each norm, we have uncovered valuable insights. The L^1-clr norm is well suited for variable selection, creating a unique subcomposition that is both internally and externally independent. The CoDa L^1-norm, on the other hand, emphasises internal independence. Lastly, the L^1-plr norm showcases a balance selection effect. Consequently, the L^1-plr norm enables more detailed study of the subcompositional structure of the compositional covariate $\mathbf{x}$ in relation to the explained variable y.

In this article, we have expanded the methodological toolkit for performing penalised regression with compositional covariates. For low dimensions, our recommendation is to run penalised regression with the L^1-plr norm. However, we cannot ignore that variable selection becomes imperative for higher dimensions. Therefore, we suggest conducting an initial examination using the CoDa L^1-norm or L^1-plr norm to gain insights into the subcompositional structure. Following this analysis, it is possible to proceed with penalised regression employing the L^1-clr norm.

As part of our future work, we aim to investigate penalised regression models that effectively integrate both the L^1-plr and L^1-clr norms into the penalty term. This research is expected to offer deeper insight into the underlying structure of compositional data, allowing for a more thorough understanding. Moreover, our aim is to improve the flexibility of modelling, especially in datasets with high dimensionality. This holistic approach will contribute to advancing the applicability and effectiveness of penalised regression techniques in the context of compositional data analysis.

Author Contributions: Conceptualization, J.A.M.-F., J.S.-R. and G.M.-F.; Formal analysis, J.A.M.-F., J.S.-R. and G.M.-F.; Methodology, J.A.M.-F., J.S.-R. and G.M.-F.; Software, J.S.-R.; Supervision, J.A.M.-F. and G.M.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Agency for Administration of University and Research grant number 2021SGR01197, and Ministerio de Ciencia e Innovación grant number PID2021-123833OB-I00, and Ministerio de Ciencia e Innovación grant number PRE2019-090976.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LASSO	Least Absolute Shrinkage and Selection Operator
CoDa	Compositional Data
clr	Centered Log-Ratio
L^1-plr	Pairwise Log-Ration norm
L^1-clr	Centered Log-Ration norm
TD	Total p-Deviation Function
MSE	Mean Squarred Error
OSQP	Operator-Splitting Quadratic Program

Appendix A

Table A1. clr-scores for the three LASSO estimators grouped into subcompositions.

Taxon	L1-clr	L1-CoDa	L1-plr
Intercept	6563.19	7023.88	7244.88
g_Subdoligranulum	455.51	409.93	249.27
f_Lachnospiraceae_g_Incertae_Sedis	245.69	77.35	
g_Dialister	229.23	182.53	
g_Bacteroides	193.46	186.23	
g_Dorea	159.97	104.06	
g_Desulfovibrio	88.32	53.57	
f_Peptostreptococcaceae_g_Incertae_Sedis	70.37	37.38	26.54
g_Faecalibacterium	63.74	33.07	
g_Paraprevotella	45.15	38.71	
f_Defluviitaleaceae_g_Incertae_Sedis	39.35	34.48	54.86
g_Alistipes	27.18	14.05	
g_Clostridium_sensu_stricto_1			20.52
g_Brachyspira	4.33		
g_Elusimicrobium		2.25	
g_Butyricimonas	0.16		
k_Bacteria_g_unclassified	−7.48	−33.79	−41.56
g_Streptococcus	−11.71		
g_Catenibacterium	−21.83	−12.10	−19.66
g_Succinivibrio	−34.64	−31.51	−10.83
g_Lachnospira	−40.80	−4.21	−29.65

Mathematics **2024**, *12*, 1388

Table A1. *Cont.*

Taxon	L1-clr	L1-CoDa	L1-plr
g_Parabacteroides	−52.41		−0.28
o_Clostridiales_g_unclassified	−111.82	−107.08	−85.66
g_Mitsuokella	−219.27	−192.13	−144.62
g_Collinsella	−233.08	−235.68	−228.94
g_Bifidobacterium	−263.23	−225.95	−154.34
f_Lachnospiraceae_g_unclassified	−626.20	−563.00	−365.84
g_subcomp		231.83	
g_subcomp_1			212.52
g_subcomp_2			248.34
g_subcomp_3			41.25
g_subcomp_4			255.78
g_subcomp_5			−27.66

Table A2. Details of the subcompositional structure for the L^1-plr LASSO estimator.

Taxon	clr(β)
g_Subdoligranulum	249.27
g_subcomp_1: g_Bacteroides, g_Dialister	212.52
f_Defluviitaleaceae_g_Incertae_Sedis	54.86
g_subcomp_2: f_Lachnospiraceae_g_Incertae_Sedis, g_Dorea, g_Faecalibacterium, g_Alistipes, g_Desulfovibrio, g_Paraprevotella	248.34
f_Peptostreptococcaceae_g_Incertae_Sedis	26.54
g_Clostridium_sensu_stricto_1	20.52
g_subcomp_3: g_Escherichia-Shigella, f_Ruminococcaceae_g_unclassified, g_Butyricimonas	41.25
g_subcomp_4: g_Brachyspira, g_Barnesiella, g_Blautia, f_Rikenellaceae_g_unclassified, g_Odoribacter, f_Erysipelotrichaceae_g_unclassified, g_Streptococcus, g_Anaerostipes, g_Phascolarctobacterium, g_Acidaminococcus, g_Anaerovibrio, g_Roseburia, g_Alloprevotella, f_Erysipelotrichaceae_g_Incertae_Sedis, g_Megasphaera, g_Coprococcus, g_Intestinimonas, g_Solobacterium, g_Oribacterium, g_Anaeroplasma, g_Victivallis, f_Ruminococcaceae_g_Incertae_Sedis, o_NB1-n_g_unclassified, g_Sutterella, o_Bacteroidales_g_unclassified, g_Prevotella, g_RC9_gut_group, f_Christensenellaceae_g_unclassified, g_Anaerotruncus	255.78
g_Parabacteroides	−0.28
g_subcomp_5: g_Ruminococcus, g_Elusimicrobium, f_vadinBB60_g_unclassified	−27.66
g_Succinivibrio	−10.83
g_Catenibacterium	−19.66
g_Lachnospira	−29.65
k_Bacteria_g_unclassified	−41.56
o_Clostridiales_g_unclassified	−85.66
g_Mitsuokella	−144.62
g_Bifidobacterium	−154.34
g_Collinsella	−228.94
f_Lachnospiraceae_g_unclassified	−365.84

References

1. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
2. Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman & Hall: London, UK, 1986.
3. Lin, W.; Shi, R.; Feng, R.; Li, H. Variable selection in regression with compositional covariates. *Biometrika* **2014**, *101*, 785–797. [CrossRef]
4. Shi, P.; Zhang, A.; Li, H. Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **2016**, *10*, 1019–1040. [CrossRef]
5. Lu, J.; Shi, P.; Li, H. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* **2019**, *75*, 235–244. [CrossRef] [PubMed]
6. Susin, A.; Wang, Y.; Lê Cao, K.A.; Calle, M.L. Variable selection in microbiome compositional data analysis. *NAR Genom. Bioinform.* **2020**, *2*, lqaa029. [CrossRef] [PubMed]
7. Saperas-Riera, J.; Martín-Fernández, J.; Mateu-Figueras, G. Lasso regression method for a compositional covariate regularised by the norm L^1 pairwise logratio. *J. Geochem. Explor.* **2023**, *255*, 107327. [CrossRef]
8. Egozcue, J.J.; Pawlowsky-Glahn, V. Groups of parts and their balances in compositional data analysis. *Math. Geol.* **2005**, *37*, 795–828. [CrossRef]
9. Pawlowsky-Glahn, V.; Egozcue, J.J. Geometric approach to statistical analysis on the simplex. *Stoch. Environ. Res. Risk Assess.* **2001**, *15*, 384–398. [CrossRef]
10. Billheimer, D.; Guttorp, P.; Fagan, W.F. Statistical Interpretation of Species Composition. *J. Am. Stat. Assoc.* **2001**, *96*, 1205–1214. [CrossRef]
11. Aitchison, J.; Bacon-Shone, J. Log contrast models for experiments with mixtures. *Biometrika* **1984**, *71*, 323–330. [CrossRef]
12. Van der Boogaart, K.G.; Tolosana, R. *Analyzing Compositional Data with R*; Use R!; Springer: Berlin/Heidelberg, Germany, 2013.
13. Dave, A. Measurement of Central Tendency. In *Applied Statistics for Economics*; Horizon Press: Toronto ON, Canada, 2014; Chapter 3.
14. Barceló-Vidal, C.; Martín-Fernández, J.A. The Mathematics of Compositional Analysis. *Austrian J. Stat.* **2016**, *45*, 57–71. [CrossRef]
15. Brezis, H. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*; Universitext; Springer: New York, NY, USA, 2011.
16. Pawlowsky-Glahn, V.; Egozcue, J.J.; Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*; John Wiley & Sons: Chichester, UK, 2015.
17. Wu, J.R.; Macklaim, J.M.; Genge, B.L.; Gloor, G.B. Finding the Centre: Compositional Asymmetry in High-Throughput Sequencing Datasets. In *Advances in Compositional Data Analisys*; Springer: Berlin/Heidelberg, Germany, 2021; Chapter 17, pp. 329–342. [CrossRef]
18. Martín-Fernández, J. Measures of Difference and Non-Parametric Classification of Compositional Data. Ph.D. Thesis, Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain, 2001.
19. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *Introduction to Statistical Learning*, 2nd ed.; Springer: New York, NY USA, 2021.
20. Bates, S.; Tibshirani, R. Log-ratio lasso: Scalable, sparse estimation for log-ratio models. *Biometrics* **2019**, *75*, 613–624. [CrossRef]
21. Monti, G.; Filzmoser, P. Sparse least trimmed squares regression with compositional covariates for high-dimensional data. *Bioinformatics* **2021**, *37*, 3805–3814. [CrossRef] [PubMed]
22. Monti, G.; Filzmoser, P. Robust logistic zero-sum regression for microbiome compositional data. *Adv. Data Anal. Classif.* **2022**, *16*, 301–324. [CrossRef]
23. Tibshirani, R.; Taylor, J. The solution path of the generalized lasso. *Ann. Statist.* **2011**, *39*, 1335–1371. [CrossRef]
24. Noguera-Julian, M.; Rocafort, M.; Guillén, Y.; Rivera, J.; Casadellà, M.; Nowak, P.; Hildebrand, F.; Zeller, G.; Parera, M.; Bellido, R.; et al. Gut Microbiota Linked to Sexual Preference and HIV Infection. *eBioMedicine* **2016**, *5*, 135–146. [CrossRef] [PubMed]
25. Rivera-Pinto, J.; Egozcue, J.J.; Pawlowsky-Glahn, V.; Paredes, R.; Noguera-Julian, M.; Calle, M.L. Balances: A new perspective for microbiome analysis. *mSystems* **2018**, *3*, e00053-18. [CrossRef] [PubMed]
26. Calle, M.; Susin, T.; Pujolassos, M. coda4microbiome: Compositional Data Analysis for Microbiome Studies; R Package Version 0.2.1. *BMC Bioinf.* **2023**, *24*, 82.
27. Palarea-Albaladejo, J.; Martín-Fernández, J.A. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemom. Intell. Lab. Syst.* **2015**, *143*, 85–96. [CrossRef]
28. Palarea-Albaladejo, J.; Martin-Fernandez, J. zCompositions: Treatment of Zeros, Left-Censored and Missing Values in Compositional Data Sets; R Package Version 1.5. 2023. Available online: https://cran.r-project.org/web/packages/zCompositions/zCompositions.pdf (accessed on 13 March 2024).
29. Martín-Fernández, J.; Hron, K.; Templ, M.; Filzmoser, P.; Palarea-Albaladejo, J. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat. Model.* **2015**, *15*, 134–158. [CrossRef]
30. Fu, A.; Narasimhan, B.; Boyd, S. CVXR: An R Package for Disciplined Convex Optimization. *J. Stat. Softw.* **2020**, *94*, 1–34. [CrossRef]
31. Stellato, B.; Banjac, G.; Goulart, P.; Bemporad, A.; Boyd, S. OSQP: An Operator Splitting Solver for Quadratic Programs. *Math. Program. Comput.* **2020**, *12*, 637–672. [CrossRef]

32. Boogaart, K.; Filzmoser, P.; Hron, K.; Templ, M.; Tolosana-Delgado, R. Classical and robust regression analysis with compositional data. *Math. Geosci.* **2021**, *53*, 823–858. [CrossRef]
33. Hyun, S.; G'Sell, M.; Tibshirani, R.J. Exact post-selection inference for the generalized lasso path. *Electron. J. Stat.* **2018**, *12*, 1053–1097. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Capítol 5

Resultats i discussió

En aquesta secció resumirem els principals resultats derivats de la tesi, centrant-nos especialment en l'objectiu principal: desenvolupar els fonaments matemàtics per a l'optimització convexa en l'espai composicional. Aquests progressos han culminat en diversos avenços significatius. Primerament, hem adaptat els elements que configuren la geometria convexa a l'espai composicional. En segon lloc, hem definit i interpretat les normes L^p en el símplex. A la definició de les normes L^p en l'espai composicional s'hi afegeix la definició de la norma L^1-plr. Aquest treball teòric s'ha vist complementat amb un treball metodològic. Concretament, s'han aplicat aquests coneixements a la regularització LASSO de models de regressió lineal amb covariable composicional, posant atenció en la selecció de balanços. Aquesta combinació de desenvolupament teòric i aplicació pràctica ha enriquit les eines disponibles per a l'estudi i anàlisi de models lineals amb composicions predictives.

5.1 Resultats

Presentem a continuació les principals contribucions d'aquesta tesi:

- En el primer article, Saperas-Riera *et al.* (2023) adapten les definicions de conjunt convex, envolupant convexa i funció convexa de manera compatible amb la geometria d'Aitchison. Aquest pas és indispensable per poder identificar correctament els problemes d'optimització convexa en CoDa. La manera i l'ordre en què es presenten les definicions fins arribar a la definició de problema d'optimització convexa segueixen el guió de llibres de referència en optimització convexa com ara Boyd

i Vandenberghe (2004) o Luenberger i Ye (2008). Es complementen aquestes definicions amb exemples que són d'interès en la bibliografia CoDa. L'article també discuteix, a través d'un exemple, la idoneïtat de la geometria composicional en la resolució del problema de definir la distància mínima a un conjunt convex en el símplex. Aquest exemple il·lustra com l'aplicació de la geometria d'Aitchison permet una resolució més coherent i precisa dels problemes d'optimització convexa per a dades composicionals, comparat amb l'ús de la geometria euclidiana convencional.

- En el segon article, Saperas-Riera *et al.* (2023), es defineix una nova norma basada en totes les parelles de relacions *pairwise*: L^1-*plr*. L'article demostra tant les propietats bàsiques com les propietats composicionals que ha de tenir tota norma en CoDa. Aquesta norma ha resultat ser de gran utilitat a l'hora d'estudiar l'estructura subcomposicional de la covariable composicional d'un model lineal. En concret, la mètrica L^1-*plr* permet seleccionar subcomposicions internament independents respecte a la variable resposta. Aquesta selecció és especialment flexible, ja que permet identificar més d'una subcomposició internament independent, i en resultat un model lineal senzill en termes de balanços.

- En el tercer article, Saperas-Riera *et al.* (2024), es defineixen amb rigor noves mètriques sobre l'espai composicional: les mètriques L^p-CoDa s'estudien i es caracteritzen en funció de balanços per tal de facilitar la seva interpretació en el context composicional. Aquest treball vol arrodonir el treball iniciat per Barceló-Vidal i Martín-Fernández (2016) en què es presenten els fonaments matemàtics per tal d'entendre les dades composicionals com a classes d'equivalència. Té especial interès l'estudi que es fa de la mètrica L^1-CoDa.

- La darrera contribució que cal destacar és l'aportació metodològica del segon i del tercer article. Aquesta tesi aprofundeix en la regularització LASSO amb covariable composicional, posant especial atenció en la selecció de balanços. Aquest avenç ha estat possible gràcies a la incorporació de normes pròpiament composicionals en el terme de penalització del problema de regularització LASSO. Així, s'obre la porta a un context que permet una anàlisi estadística més flexible i adequada dels models lineals amb covariable composicional.

5.2 Conclusions

Aquesta tesi omple el buit existent en la bibliografia sobre dades composicionals en els àmbits de la convexitat, l'optimització convexa i les mètriques L^p. S'han adaptat les definicions bàsiques de convexitat i s'han proporcionat exemples de conjunts convexos i funcions convexes d'interès en el context composicional. Entre aquestes funcions convexes, les funcions de distància tenen especial rellevància, ja que influeixen significativament en tècniques estadístiques com ara la detecció d'anomalies, el disseny experimental de mescles, l'anàlisi de clústers i la regressió LASSO.

A més, aquesta tesi omple un buit en l'estudi de les normes L^p per a dades composicionals. S'han definit les normes L^p-CoDa, s'han demostrat les seves propietats bàsiques i s'han establert algunes desigualtats entre normes. Aquest treball de fonament matemàtic s'ha aplicat a la regressió LASSO amb covariable composicional, amb l'objectiu d'aprofundir i ampliar les metodologies de regularització LASSO. A través de l'estudi de l'impacte d'aquestes mètriques en els models lineals, aquesta tesi també ha permès una millor comprensió de les mètriques en si mateixes.

5.3 Futures línies de recerca

Durant la realització de la tesi, han sorgit noves línies de recerca basades en els resultats obtinguts. Aquestes noves vies inclouen l'aplicació de les normes L^p-CoDa en altres mètodes estadístics amb elements mètrics, com ara són l'anàlisi de clústers, l'escalat multidimensional (Borg i Groenen, 2005) i el mapa autoorganitzat (Kohonen, 2001). A més, es preveu l'extensió de la metodologia desenvolupada per a la regularització LASSO a altres formes de regularització en regressió i l'estudi de la seva aplicabilitat en diversos camps com ara la bioestadística, la geologia i les ciències ambientals. Una tercera línia de recerca que ha emergit està relacionada amb l'optimització convexa, centrant-se en el disseny d'experiments uniformes en el símplex.

- Resta pendent ampliar el ventall d'exemples per tal d'enriquir i validar els resultats obtinguts fins ara. En particular, un dels aspectes clau és aplicar la regularització L^1-*plr* al model de regressió logística. Aquesta extensió és fonamental per consolidar la validesa del model i augmentar la seva aplicabilitat en àmbits reals.

- En la metodologia LASSO, queda pendent estudiar una mètrica composicional que permeti un procés de regularització equivalent al que

obtenim amb la norma L^1-*plr* i que permeti l'estudi de les subcomposicions externament independents.

- Seguint amb la regularització LASSO, un treball pendent és estudiar combinacions lineals convexes entre normes, de manera que puguem combinar els beneficis de diferents normes en un sol procés de regularització. Un exemple és la combinació de la mètrica L^1-*plr* i la mètrica L^1-clr. Aquest procés de regularització crea una subcomposició internament i externament independent (selecció de variable) i al mateix temps crea altres subcomposicions internament independents.

$$\min_{\beta_0,\beta}\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\beta_0-\langle\beta,\mathbf{X}_i\rangle_{\mathcal{A}}\right)^2+\lambda\left(\theta\|\beta\|_{1-\mathrm{clr}}+(1-\theta)\|\beta\|_{1-plr}\right)\right\}$$

  A més, tal i com es comenta en Saperas-Riera *et al.* (2023), queda pendent estudiar regularització coneguda com a Elastic Net, en la qual es combinen en el terme de penalització la norma L^1 amb la norma euclidiana L^2.

$$\min_{\beta_0,\beta}\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\beta_0-\langle\beta,\mathbf{X}_i\rangle_{\mathcal{A}}\right)^2 + \lambda_1\left(\theta\|\beta\|_{1-\mathrm{clr}}+(1-\theta)\|\beta\|_{1-plr}\right)+\lambda_2\|\beta\|_{\mathcal{A}}^2\right\}$$

  La regularització Elastic Net té un interès especial, ja que ajuda a mitigar els problemes de col·linealitat entre les variables predictores, permetent obtenir models més robustos en presència d'alta correlació entre aquestes.

- A Zou *et al.* (2021), es proposa una metodologia per a l'anàlisi de clúster per a dades composicionals que utilitza una forma de regularització per controlar la formació dels clústers. Aquesta metodologia, anomenada *clusterpath*, utilitza una funció de cost que combina una mesura de la distància entre punts de dades ($\mathbf{x}_i$) amb un terme de penalització que controla la formació de clústers ($\mathbf{u}_i$). A través de l'optimització d'aquesta funció, els punts de dades s'agrupen en clústers.

$$\min_{\mathbf{u}_i}\left\{\frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i\ominus\mathbf{u}_i\|_{\mathcal{A}}^2+\lambda\sum_{i<j}w_{ij}\|\mathbf{u}_i\ominus\mathbf{u}_j\|_*\right\}$$

L'objectiu d'aquesta línia de recerca és explorar el procés de formació de clúster quan es canvia la norma ($\|\cdot\|_*$ per L^1-clr, L^1-CoDa o L^2-CoDa) en el terme de penalització.

El treball fet en aquesta línia de recerca ja ha permès presentar una comunicació oral:

- 10th International Workshop on Compositional Data Analysis (CODAWORK 2024). **Convex Clustering Within the Simplex.** Paula de la Lama-Zubiran, Jordi Saperas-Riera i Valentino di Donato. Girona, Spain.

- L'objectiu principal d'un *disseny uniforme* és distribuir els punts de manera que cobreixin l'espai de forma homogènia. Això és especialment important quan es vol assegurar que totes les regions de l'espai estan representades de manera equitativa. Un avantatge dels dissenys uniformes en comparació amb els clàssics dissenys factorials és que l'experiment es pot completar amb un nombre reduït de proves, fins i tot quan el nombre de factors o el nombre de nivells de cada factor és elevat. En el cas particular dels experiments amb mixtures, en què els factors són les proporcions dels ingredients, el símplex esdevé l'espai de treball.

  Aproximacions al disseny uniforme per a mixtures com a Fang i Chan (2006), on la uniformitat en la distribució dels punts s'aconsegueix a partir de minimitzar una L_p-discrepància, no aconsegueixen captar la geometria d'Aitchison. L'objectiu d'aquesta línia de recerca és adaptar el disseny uniforme per a mixtures de manera compatible amb la geometria d'Aitchison. El treball fet en aquesta línia de recerca ja ha permès presentar un póster amb la definició de distribució uniforme per a dades composicionals i una comunicació oral en què s'introdueix el disseny factorial en el símplex de manera compatible amb la geometria d'Aitchison:

  - 9th International Workshop on Compositional Data Analysis (CODAWORK 2022). **The uniform distribution on the simplex.** Jordi Saperas-Riera i Glòria Mateu-Figueras. Toulouse, France.
  - 10th International Workshop on Compositional Data Analysis (CODAWORK 2024). **Introducing Uniform Design in Compositional Data Analysis.** Jordi Saperas-Riera i Glòria Mateu-Figueras. Girona, Spain.

Bibliografia

Aitchison, J. *Principal Component Analysis of Compositional Data.* Biometrika **70**(1),57-65 (1983)

Aitchison, J. *Log contrast models for experiments with mixtures.* Biometrika **71**,323–330 (1984)

Aitchison, J. *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (1986)

Barceló-Vidal, C., Martín-Fernández, J.A. i Mateu-Figueras, G. Compositional differential calculus on the simplex. *In: Compositional Data Analysis: Theory and Applications.* 176–190 Wiley, Chichester (2011)

Barceló-Vidal, C. i Martín-Fernández, J.A. The Mathematics of Compositional Data Analysis. *Austrian Journal of Statistics* **45**(4),57-71 (2016)

Bates, S. i Tibshirani, R. Log-Ratio Lasso: Scalable, Sparse Estimation for Log-Ratio Models. *Biometrics* **75**(2),613–624 (2018)

Billheimer, D., Guttorp, P. i Fagan, W.F. Statistical Interpretation of Species Composition *Journal of the American Statistical Association* **96**(456), 1205-1214 (2001)

Boogaart, K.G.v.d. i Tolosana, R. Analyzing Compositional Data with R. Use R!. Springer Berlin, Heidelberg (2013)

Boogaart, K.G.v.d., Filzmoser, P., Hron, K., Templ, M. i Tolosana-Delgado, R. Classical and Robust Regression Analysis with Compositional Data
newblock *Mathematical Geosciences* **53**,823–858 (2021)

Borg, I. i Groenen, P. Modern Multidimensional Scaling: Theory and Applications Springer, New York (2005)

Boyd, S. i Vandenberghe, L. Convex Optimization Cambridge university press., Cambridge (2004)

Brezis,H. Functional Analysis, Sobolev Spaces and Partial Differential Equations. Universitext; Springer, New York (2011)

Calle, M.L., Pujolassos, M. i Susin, A. coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinformatics* **24**(82), (2023)

Chatterjee, A. i Lahiri, S. N. Bootstrapping Lasso Estimators. *Journal of the American Statistical Association* **106**(494), 608–625 (2011)

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras,G. i Barceló-Vidal, C. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology* **35**(3), 279–300 (2003)

Egozcue, J.J. i Pawlowsky-Glahn, V. Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology* **37**, 795–828 (2005)

Fang, K.T. i Chan, L.Y. Uniform Design and Its Industrial Applications. *In: Springer Handbook of Engineering Statistics*. Springer, London (2006).

Gloor G.B., Macklaim J.M., Pawlowsky-Glahn V. i Egozcue J.J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8** (2017)

Hron, K., Filzmoser, P. i Thompson, K. Linear regression with compositional explanatory variables. *Journal of Applied Statistics* **39**(5), 1115-1128 (2012)

Hyun, S., G'sell, M. i Tibshirani, R. J. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics* **12**, 1053–1097 (2018)

James, G., Witten, D., Hastie, T. i Tibshirani, R. An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics; Springer, New York (2021).

Kohonen, T. Self-Organizing Maps Springer Berlin, Heidelberg (2001).

Lee, J., Sun, D., Sun, Y. and Taylor, J. *Exact post-selection inference, with application to the lasso.* Annals of Statistics **44**(3), 907–927 (2016)

Lin, W., Shi, P., Feng, R. i Li, H. *Variable selection in regression with compositional covariates.* Biometrika **101**(4),785-797 (2014)

Luenberger, D. and Ye, Y. Linear and Nonlinear Programming. International Series in Operations Research & Management Science. Springer, New York (2008)

Martín-Fernández, J.A. Comments on: Compositional data: the sample space and its structure. *TEST* **28**,653–657 (2019)

Pawlowsky-Glahn, V. i Egozcue, J.J. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment* **15**, 384–398 (2001)

Pawlowsky-Glahn, V. i Buccianti, A. Compositional Data Analysis: Theory and Applications. John Wiley, Chichester (2011)

Saperas-Riera,J., Martín-Fernández, J.A. i Mateu-Figuera, G. Fundamentals of convex optimization for compositional data. *Statistics and Operations Research Transactions (SORT)* **47**(2), 323-344 (2023)

Saperas-Riera,J., Mateu-Figuera, G. i Martín-Fernández, J.A. Lasso regression method for a compositional covariate regularised by the norm L^1-pairwise logratio. *Journal of Geochemical Exploration* **255**, (2023)

Saperas-Riera,J., Mateu-Figuera, G. i Martín-Fernández, J.A. L^p-Norm for Compositional Data: Exploring the CoDa L^1-Norm in Penalised Regression. *Mathematics* **12**, (2024)

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **58**(1), 267-288 (1996)

Tibshirani, R. i Taylor, J. The solution path of the generalized LASSO. *The Annals of Statistics,* **39**(3), 1335-1371 (2011)

Zou, H. i Hastie, T. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **67**(2), 301-320 (2005)

Wang, X., Wang, H., Wang, Z. i Yuan, J. Convex clustering method for compositional data modeling. *Soft Computing* **25**, 2965–2980 (2021)
Wu, J.R., Macklaim, J.M., Genge, B.L., Gloor, G.B.

Xie, Y., Neumann, A., Stanford, T., Rasmussen, C.L., Dumuid, D. i Neumann, F. Evolutionary Time-Use Optimization for Improving Children's Health Outcomes. *Lecture Notes in Computer Science* **13399**, 323-337 (2022)

Xie, Y., Neumann, A., Stanford, T., Rasmussen, C.L., Dumuid, D. i Neumann, F. Evolutionary Time-Use Optimization for Improving Children's Health Outcomes. *Lecture Notes in Computer Science* **13399**, 323-337 (2022)

Wang, X., Wang, H., Wang, Z. i Yuan, J. Convex clustering method for compositional data modeling. *Soft Comput* **25**, 2965–2980 (2021)