
Penalized regression methods for compositional data

Taciana Kisaki Oliveira Shimizu

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

Taciana Kisasi Oliveira Shimizu

Penalized regression methods for compositional data

Thesis submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics.

Advisor: Prof. Dr. Francisco Louzada Neto

São Carlos
February 2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

Taciana Kisasi Oliveira Shimizu

Penalized regression methods for compositional data

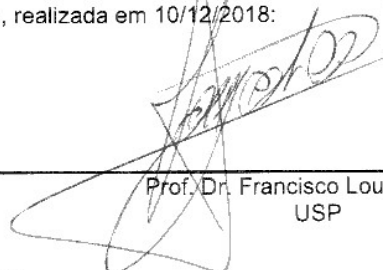
Thesis submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics.

Advisor: Prof. Dr. Francisco Louzada Neto

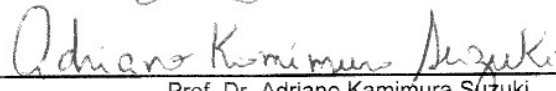
São Carlos
February 2019

Folha de Aprovação

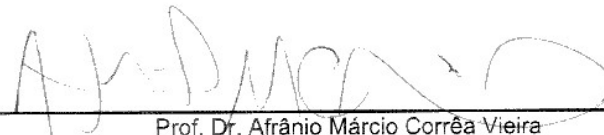
Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado da candidata Taciana Kasaki Oliveira Shimizu, realizada em 10/12/2018:




Prof. Dr. Francisco Louzada Neto
USP



Prof. Dr. Adriano Kamimura Suzuki
USP



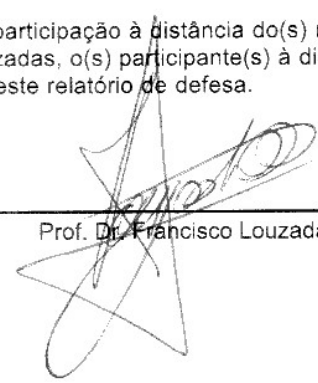
Prof. Dr. Afrânio Márcio Corrêa Vieira
UFSCar



Profa. Dra. Teresa Cristina Martins Dias
UFSCar

Prof. Dr. Paulo Henrique Ferreira da Silva
UFBA

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Paulo Henrique Ferreira da Silva e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ão) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.



Prof. Dr. Francisco Louzada Neto

For my husband, Marcelo (Hiro), who has always been a constant source of support and encouragement during the challenges of my life!

ACKNOWLEDGEMENTS

Firstly, I would like to thank God for the gift of my life, providing protection and wisdom in all the moments of my life.

To my parents who devoted their love and care to my education. To my brothers who always demonstrated love and fraternal fellowship.

To my husband Hiro, who has always supported me unconditionally in my personal and professional projects with love and patience.

I am especially grateful to my supervisor Dr. Francisco Louzada for his patience, comprehension, encouragement, professionalism and support in this study. I am also grateful to Professor Adriano Kamimura Suzuki who always patient and willing to help in moments that I needed.

I am thankful to my family and my husband's family who helped us in crucial moments.

I express my sincere gratitude to Elizabeth Mie Hashimoto who always supported me and encouraged me to go ahead.

I would also like to thank the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - processo nº 2014/16147-3, for the financial support, which enabled me to carry out the project.

*“O destino é uma questão de escolha.”
(Augusto Cury)*

RESUMO

SHIMIZU, T. K. O. **Métodos de regressão penalizados para dados composicionais**. 2019. 95 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Dados composicionais consistem em vetores conhecidos como composições cujos componentes são positivos e definidos no intervalo $(0, 1)$ representando proporções ou frações de um “todo”, sendo que a soma desses componentes totalizam um. Tais dados estão presentes em diferentes áreas, como na geologia, ecologia, economia, medicina entre outras. Desta forma, há um grande interesse em ampliar os conhecimentos acerca da modelagem de dados composicionais, principalmente quando há a influência de covariáveis nesse tipo de dado. Nesse contexto, a presente tese tem por objetivo propor uma nova abordagem de modelos de regressão aplicada em dados composicionais. A ideia central consiste no desenvolvimento de um método balizado por regressão penalizada, em particular Lasso, do inglês *least absolute shrinkage and selection operator*, *elastic net* e Spike-e-Slab Lasso (SSL) para a estimação dos parâmetros do modelo. Em particular, visionamos o desenvolvimento dessa modelagem para dados composicionais, com o número de variáveis explicativas excedendo o número de observações e na presença de grandes bases de dados, e além disso, quando há restrição na variável resposta e nas covariáveis.

Palavras-chave: Dados composicionais, modelo de regressão, coordenadas log-razão isométricas, seleção de variáveis.

ABSTRACT

SHIMIZU, T. K. O. **Penalized regression methods for compositional data**. 2019. 95 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Compositional data consist of known vectors such as compositions whose components are positive and defined in the interval $(0, 1)$ representing proportions or fractions of a “whole”, where the sum of these components must be equal to one. Compositional data is present in different areas, such as in geology, ecology, economy, medicine, among many others. Thus, there is great interest in new modeling approaches for compositional data, mainly when there is an influence of covariates in this type of data. In this context, the main objective of this thesis is to address the new approach of regression models applied in compositional data. The main idea consists of developing a marked method by penalized regression, in particular the Lasso (least absolute shrinkage and selection operator), elastic net and Spike-and-Slab Lasso (SSL) for the estimation of parameters of the models. In particular, we envision developing this modeling for compositional data, when the number of explanatory variables exceeds the number of observations in the presence of large databases, and when there are constraints on the dependent variables and covariates.

Keywords: Compositional data, regression model, isometric logratio coordinates, variable selection.

LIST OF FIGURES

Figure 1	– Estimation picture for the Lasso (left) and ridge regression (right).	34
Figure 2	– Spike and Slab distributions for $\lambda_0 = 1, 2, 3$ and $\lambda_1 = 0.1$	36
Figure 3	– ICMS series disaggregated in three economic sectors.	42
Figure 4	– The solution path SSL (A, B, C), Lasso (D) and elastic net (E) for $\text{ilr}(y_1)$. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model Lasso and elastic net (cross-validation), respectively.	45
Figure 5	– The solution path SSL (A, B, C), Lasso (D) and elastic net (E) for $\text{ilr}(y_2)$. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model Lasso and elastic net (cross-validation), respectively.	46
Figure 6	– The parameter estimation averaged over 1000 replicates assuming $\rho = 0.2$ for the covariance matrix ($n = 100, p = 1000$).	51
Figure 7	– The parameter estimation averaged over 1000 replicates assuming $\rho = 0.5$ for the covariance matrix ($n = 100, p = 1000$).	52
Figure 8	– The SSL solution paths (A, B, C).	53
Figure 9	– The Lasso solution path (D) and elastic net path (E).	58
Figure 10	– The solution path SSL (A, B, C), lasso (D) and elastic net (E) for healthy patients. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model lasso and elastic net (cross-validation), respectively.	59
Figure 11	– The solution path SSL (A, B, C), lasso (D) and elastic net (E) for patients with pathologies. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model lasso and elastic net (cross-validation), respectively.	60
Figure 12	– The SSL solution paths (A, B, C) (for $\text{ilr}(y_1)$).	68
Figure 13	– The lasso solution path (D) and elastic net path (E) (for $\text{ilr}(y_1)$).	69
Figure 14	– The SSL solution paths (A, B, C) (for $\text{ilr}(y_2)$).	70
Figure 15	– The lasso solution path (D) and elastic net path (E) (for $\text{ilr}(y_2)$).	71
Figure 16	– The solution path SSL (A, B, C), lasso (D) and elastic net (E) for $\text{ilr}(y_1)$. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model lasso and elastic net (cross-validation), respectively.	72

Figure 17 – The solution path SSL (A, B, C), lasso (D) and elastic net (E) for $\text{ilr}(y_2)$. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model lasso and elastic net (cross-validation), respectively. 73

LIST OF TABLES

Table 1	– Example: Volleyball game score.	23
Table 2	– Elementary logistic transformations of \mathbb{S}^D for \mathbb{R}^{D-1}	25
Table 3	– Averages of some performance measures for penalized methods with compositional response variable $(\text{ilr}(y_1))$	40
Table 4	– Averages of some performance measures for penalized methods with compositional response variable $(\text{ilr}(y_2))$	41
Table 5	– Averages of some performance measures for penalized methods with compositional covariates.	50
Table 6	– Averages of some performance measures for penalized methods with compositional dependent variables and covariates $(\text{ilr}(y_1))$	63
Table 7	– Averages of some performance measures for penalized methods with compositional dependent variables and covariates $(\text{ilr}(y_2))$	64

LIST OF ABBREVIATIONS AND ACRONYMS

alr	additive logratio
BMI	body mass index
clr	centered logratio
EM	Expectation-Maximization
FN	false negative
FP	false positive
HAM	Hamming
ilr	isometric logratio
LARS	Least Angle Regression
MCMC	Markov Chain Monte Carlo
MSE	Mean Square Error
OLS	Ordinary Least Squares
PA	phase angle
SCAD	Smoothly Clipped Absolute Deviations
SSL	Spike-and-Slab Lasso
WHO	World Health Organization

CONTENTS

1	INTRODUCTION	23
2	PRELIMINARIES	27
2.1	Basic Concepts for compositional data	27
2.1.1	<i>Principles of compositional analysis</i>	28
2.1.1.1	<i>Scale invariance</i>	28
2.1.1.2	<i>Permutation invariance</i>	28
2.1.1.3	<i>Subcompositional coherence</i>	28
2.1.2	<i>The Aitchison geometry</i>	28
2.1.3	<i>Logratio coordinates</i>	30
2.2	Shrinkage Methods	32
2.2.1	<i>Lasso</i>	33
2.2.1.1	<i>Orthonormal design</i>	34
2.2.1.2	<i>K-fold Cross-validation</i>	34
2.2.1.3	<i>Coordinate descent algorithm</i>	35
2.2.2	<i>Elastic net</i>	35
2.2.3	<i>Spike-and-Slab Lasso</i>	35
3	PENALIZED REGRESSION MODEL FOR COMPOSITIONAL RESPONSE VARIABLES	37
3.1	Simulation Analysis	38
3.2	Real data application - ICMS dataset	40
3.3	Discussion	43
4	PENALIZED REGRESSION MODEL FOR COMPOSITIONAL COVARIATES	47
4.1	Simulation Analysis	48
4.2	Artificial Data	53
4.3	Real data application - Brazilian children malnutrition dataset	54
4.4	Discussion	57
5	PENALIZED REGRESSION MODEL FOR COMPOSITIONAL RESPONSE VARIABLES AND COVARIATES	61
5.1	Simulation Analysis	62

5.2	Toy example	63
5.3	Real data application - ICMS dataset	65
5.4	Discussion	66
6	CONCLUSION	75
	BIBLIOGRAPHY	77
	APPENDIX A	81

INTRODUCTION

Appropriate study of the compositional data theory has been developed since the 1970s from the contributions of [Aitchison and Shen \(1980\)](#) and [Aitchison \(1982a\)](#). Since then, its applications have grown in different areas of knowledge, some examples include mineral compositions of rocks or sediment compositions such as (sand, silt, clay) compositions in geology, species compositions of biological communities in ecology, household budget compositions in economy, blood and urine compositions in medicine.

Compositional data are vectors of proportions that specify D fractions as a whole. Therefore, for $\mathbf{z} = (z_1, z_2, \dots, z_D)^\top$ to be a compositional vector, $z_i > 0$ for $i = 1, \dots, D$ and $z_1 + z_2 + \dots + z_D = 1$.

In order to exemplify, we can describe compositional data as follows in the Table 1. The vector \mathbf{z} in the simplex sample space as a composition (rows of the Table 1 - % attack, % block, % serve and % opponent's error of each match), the elements of such vector as components (columns of the Table 1) and the set of these vectors represent compositional data (Table 1) ([AITCHISON, 1982b](#)). Such data often result from the normalization of raw data or obtaining

Table 1 – Example: Volleyball game score.

Match	% attack	% block	% serve	% opponent's error
1	48.00	12.00	2.67	37.33
2	53.06	14.29	7.14	25.51
3	44.00	13.33	8.00	34.67
4	52.63	14.74	7.37	25.26
5	56.00	8.00	5.33	30.67
6	65.63	10.16	2.34	21.88
⋮	⋮	⋮	⋮	⋮

data as proportions of a certain heterogeneous quantity. Standard methods for multivariate data analysis under the usual assumption of multivariate normal distribution (see, for example,

([JOHNSON; WICHERN, 1998](#)) are not appropriate for compositional data, due to compositional restrictions.

Different models have been adopted for the analysis of compositional data analysis. The first was the Dirichlet distribution; however, it requires the correlation structure to be wholly negative, a fact that is not observed for compositional data, in which some correlations are positive (see, for example, ([AITCHISON, 1982a](#); [AITCHISON, 1986](#))).

An alternative for the analysis of compositional data was proposed by Aitchison [Aitchison \(1986\)](#), who considered suitable transformations from restricted sample space Simplex to well-defined real sample space. More specifically, Aitchison and Shen [Aitchison and Shen \(1980\)](#) developed the logistic-Normal class of distributions transforming the D component vector \mathbf{x} into a vector \mathbf{y} in \mathbb{R}^{D-1} and considering the additive logratio (alr) function.

The alr and centered logratio (clr) transformations were introduced by [Aitchison \(1986\)](#) in order to solve the constant sum constraint. These transformations are coordinates with respect to the Aitchison geometry ([PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015](#)). However, a remarkable disadvantage about the clr transformation is that the variance matrix of its transformed composition is unique. Furthermore, the alr coordinates are non-isometric and asymmetric and the clr coordinates are isometric and symmetric ([CHEN; ZHANG; LI, 2017](#)). Another transformation for compositional data is proposed by [Egozcue et al. \(2003\)](#) called isometric logratio (ilr), which is calculated with respect to a given orthogonal basis, allowing a simple manipulation of the geometric elements in the simplex sample space. Such a transformation preserves all metric properties in the real coordinates. According to the [Hron, Filzmoser and Thompson \(2012\)](#), the ilr transformation provides a way to obtain an interpretation of the unknown parameters for regression model without constraints on the parameters.

Table 2 presents other elementary transformations: multiplicative logistic and hybrid logistic, besides the alr mentioned above.

More recently, some contributions about the theory and applications of compositional data have been developed, for example, in [Pawlowsky-Glahn and Buccianti \(2011\)](#), [Boogaart and Tolosana-Delgado \(2013\)](#), [Pawlowsky-Glahn, Egozcue and Tolosana-Delgado \(2015\)](#). [Hijazi and Jernigan \(2009\)](#) made a comparison between the Dirichlet regression model and the alr transformation to verify which one is better in the presence of the covariate.

[Aitchison and Egozcue \(2005\)](#) reported a bibliography review about statistical modeling for compositional data in the last twenty years, where one of the tantalizing problems in compositional data was how to deal with the presence of components equal to zero. One of the few articles which addressed this situation was proposed by [Martín-Fernandez, Barceló-Vidal and Pawlowsky-Glahn \(2003\)](#) who considered non-parametric imputation. [Hijazi \(2011\)](#) proposed a novel technique based on the Expectation-Maximization (EM) algorithm to replace the components containing zeros.

Table 2 – Elementary logistic transformations of \mathbb{S}^D for \mathbb{R}^{D-1} .

Transformations	Inverses
alr	$y_i = \ln \left(\frac{z_i}{z_D} \right)$
multiplicative logistic	$y_i = \ln \left(\frac{z_i}{1 - \sum_{k=1}^i z_k} \right)$
hybrid logistic	$y_1 = \ln \left(\frac{z_1}{1 - z_1} \right);$ $y_i = \ln \left[\frac{z_i}{\left(1 - \sum_{k=1}^{i-1} z_k \right) \left(1 - \sum_{k=1}^i z_k \right)} \right], \quad i = 2, \dots, D - 1$

On the other hand, the increase in large datasets whose dimensionality is much larger than the sample size establishes new challenges for current methodology of compositional data. The fact that there is a situation of a low relation between the number of dependent variables and the sample size makes the standard analysis unsuitable for a regression model with compositional data. According to this situation, the high collinearity among the covariates can also be seen, which is restricted by the dependence on each other. Whenever one of the above situations is considered, a problem of poor conditioning is observed, and a contraction can be considered in order to overcome this problem.

The advance of regularization techniques for variable selection and estimation in linear regression have received much attention from many authors to handle high-dimensional datasets and colinearity between the covariates of the model. Among the most popular penalization approaches are the Lasso (TIBSHIRANI, 1996), the elastic net (ZOU; HASTIE, 2005), the Smoothly Clipped Absolute Deviations (SCAD) (FAN; LI, 2001), among others and more recently, the SSL (ROCKOVÁ; GEORGE, 2018).

The l_1 regularization or Lasso, and its extensions, has become a popular method because it achieves a sparse solution (TIBSHIRANI, 1996). This method shrinks many coefficients exactly to zero through the fast optimization algorithms called Least Angle Regression (LARS) and the cyclic coordinate descent proposed by Efron *et al.* (2004) and Friedman, Hastie and Tibshirani (2010), respectively. Recently, Lin *et al.* (2014) have proposed an l_1 regularization method for variable selection and estimation in high-dimensional linear models with constraints in the covariates, combining coordinate descent with the method of multipliers. Similar to Lasso, Zou and Hastie (2005) proposed a new regularization and variable selection method that deal with strong correlations among the covariates. The best advantage of the elastic net is that it incorporates the ridge penalty with the Lasso penalty. This combination performs feature

selection and works with multicollinearity in the dataset together, which reveals important attributes for the analysis where the number of observations is smaller than the number of covariates in the model. The combination of ridge and Lasso performs feature selection and handles multicollinearity within the dataset, which are important characteristics for analyzing datasets with large numbers of features (many of which could be collinear) and relatively smaller number of observations.

On the other hand, a Bayesian alternative is to adapt the amount of shrinkage applied to the hierarchical model with mixture Spike-and-Slab priors (GEORGE; MCCULLOCH, 1993). In this context, the Spike-and-Slab prior has been an important tool for most Bayesian variable selection (CHIPMAN, 1996; ROCKOVÁ; GEORGE, 2014). Some studies have applied Spike-and-Slab variable selection approaches using the mixture normal priors on coefficients by Markov Chain Monte Carlo (MCMC) algorithms (see for example, Ishwaran and Rao (2005); Shelton *et al.* (2015); among others). Although widely practical, the MCMC methods have a high computational cost. Moreover, such methods cannot perform a variable selection, and the mixture of normal priors does not shrink coefficients towards zero. An EM algorithm was developed by Rocková and George (2014) to apply in large-scale linear models with the mixture normal priors. Recently, Rocková and George (2018) developed a new structure for high-dimensional normal linear models; the so-called SSL. Under this model, a new prior was applied to the coefficients, that is, the Spike-and-Slab mixture double-exponential distribution. The SSL is a fast-computable approximation to mode detection under the Spike-and-Slab mixture of a point mass at 0 and ensures significant theoretical and practical properties. The SSL method applied to Cox models and generalized linear models have received some attention in the literature, as can be seen in Tang *et al.* (2017b) and Tang *et al.* (2017a), respectively. However, to the best of our knowledge, the shrinkage methodology for compositional data needs to be developed further.

In this context, the main objective of this thesis is to introduce a proposal for regression models based on regularization methods such as Lasso, elastic net and SSL, where responses and/or covariates have a compositional character. It is worth pointing out that the study with using these data is challenging due to the dependence and absence of parametric classes in the simplex sample space.

The remainder of this work is organized as follows. Chapter 2 introduces the preliminaries of some important topics of compositional data and the shrinkage methods adopted for the analysis. Chapter 3 presents the penalized regression model for compositional dependent variables under the application of Lasso, elastic net and SSL methods for penalization. Chapter 4 presents a penalized regression model when the restriction of compositional data exists in the covariates. Chapter 5 presents the case of compositional constraints on both dependent variables and covariates where the regularization methods applied are Lasso, elastic net and SSL. Finally, Chapter 6 draws some general conclusions and possible extensions of this current work.

PRELIMINARIES

In this Chapter, we present a literature review of some important topics and properties about basic operations in the methodology of compositional data, logratio coordinates and some regularization methods which were applied in the proposed models.

2.1 Basic Concepts for compositional data

Initially, we start by defining compositional data. According to [Pawlowsky-Glahn, Egozcue and Tolosana-Delgado \(2015\)](#), a column vector $\mathbf{z} = (z_1, z_2, \dots, z_D)^\top$ is a D -part composition when all the components are positive real numbers and carry only relative information.

An important operation called closure assigns a constant sum representative to a composition. It divides each component of a vector by the sum of the components, rescaling the initial vector to the constant sum k . In mathematical terms, the definition is given by

Definition 1 (Closure). For any vector of D strictly positive real components, $\mathbf{z} = [z_1, \dots, z_D] \in \mathbb{R}_+^D$, $z_i > 0$ for all $i = 1, \dots, D$, the closure of \mathbf{z} to $c > 0$ is defined as

$$C(\mathbf{z}) = \left[\frac{cz_1}{\sum_{i=1}^D z_i}, \dots, \frac{cz_D}{\sum_{i=1}^D z_i} \right],$$

where c is an arbitrary positive real number and is usually 1 (proportions) or 100 (%) depending on the units of measurement.

An appropriate scaling factor can be used to represent compositional data as proportions. Consequently, we can assume compositional data as proportions, that is, as vectors of constant sum c .

The sample space of compositional data called simplex is denoted by

$$\mathbb{S}^D = \{\mathbf{z} = (z_1, z_2, \dots, z_D)^\top : z_i > 0, i = 1, 2, \dots, D; \sum_{i=1}^D z_i = c\},$$

More specifically, we can define a vector \mathbf{z} in the simplex sample space as a composition, the elements of such vector as components and the set of these vectors represent compositional data (AITCHISON, 1982a).

2.1.1 Principles of compositional analysis

The definition of compositional data follows the natural principles of compositions and they are called: scale invariance, permutation invariance and subcompositional coherence.

2.1.1.1 Scale invariance

Scale invariance refers to when a composition has information only about relative values. According to Aitchison and Egozcue (2005), the concept is easily formalized into a statement that all meaningful functions of a composition can be expressed in terms of a set of component ratios. In other words, if a composition changes from parts per unit to percentages, for example, the information carried is completely equivalent (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015).

Definition 2 (Scale invariance). Let $f(\cdot)$ be a function defined on \mathbb{R}_+^D . Such a function is scale invariant if for any positive real value $v \in \mathbb{R}_+$ and for any composition $\mathbf{z} \in \mathbb{S}^D$ it satisfies $f(v\mathbf{z}) = f(\mathbf{z})$, that is, it yields the same result for all compositionally equivalent vectors.

2.1.1.2 Permutation invariance

The concept of permutation invariance is that it provides the same results when the components in the composition are changed (see Pawlowsky-Glahn, Egozcue and Tolosana-Delgado (2015) for a detailed discussion).

2.1.1.3 Subcompositional coherence

Finally, a definition for subcomposition is given by a subset of components or parts of a composition. Thus, the subcompositional coherence can be summarized as: if we have two compositions, in which one has full compositions and the other one a subcomposition of these full compositions, the inference about the relations within the common parts should be the same results, i.e., the scale invariance of the results is preserved within arbitrary subcompositions, that is, the ratios between any parts in the subcomposition are equal to the corresponding ratios in the original composition.

2.1.2 The Aitchison geometry

In Euclidian geometry, we work with operations in vectors in real space. This geometry is familiar with its geometric structure because the real space is a linear vector space with a metric structure. However, this geometry is not suitable for analyzing compositional data. A way

to illustrate this statement is to consider four compositions $[5, 55, 40]$, $[15, 45, 40]$, $[40, 30, 30]$, $[50, 20, 30]$. The difference between $[5, 55, 40]$ and $[15, 45, 40]$ is not the same as the difference between $[40, 30, 30]$ and $[50, 20, 30]$. The Euclidean distance between them is the same, there is a difference of 10 units both between the first and second components respectively. While in the first case, the proportion in the first component is triplicated, in the second case, the relative increase is about 25%. To describe compositional variability, it is more interesting to consider this relative difference.

This is one of the reasons for dispensing the Euclidian geometry as an appropriate tool for analyzing compositional data. Other problems might occur, such as those where outcomes finish up outside the sample space simplex, or when translating compositional vectors, or determining joint confidence regions for random compositions under assumptions of normality.

A wise geometry is needed to deal with compositional data. Indeed, it is possible to obtain two operations that provide the simplex of a vector space structure. They are defined as perturbation and powering. The first one is like an addition in real space and the second one is like a multiplication by a scalar in real space. These basic operations required for a vector space structure of the simplex are defined below.

Definition 3 (Perturbation). Consider the compositions $\mathbf{z}, \mathbf{y} \in \mathbb{S}^D$. The perturbation of \mathbf{z} with \mathbf{y} is given by

$$\mathbf{z} \oplus \mathbf{y} = C[z_1 y_1, z_2 y_2, \dots, z_D y_D] \in \mathbb{S}^D.$$

where $C[\cdot]$ is defined in Definition 1.

Definition 4 (Powering). Consider the compositions $\mathbf{z}, \mathbf{y} \in \mathbb{S}^D$. The powering of \mathbf{z} by a constant $\alpha \in \mathbb{R}$ as the composition is given by

$$\alpha \odot \mathbf{z} = C[z_1^\alpha, z_2^\alpha, \dots, z_D^\alpha] \in \mathbb{S}^D.$$

To obtain a Euclidian vector space structure (PAWLOWSKY-GLAHN; EGOZCUE, 2001), we take the following inner product with its related norm $\|\cdot\|_a$ and Aitchison distance (the subindex a stands for Aitchison).

Definition 5 (Aitchison inner product). Inner product of $\mathbf{z}, \mathbf{y} \in \mathbb{S}^D$,

$$\langle \mathbf{z}, \mathbf{y} \rangle_a = \sum_{i=1}^D \ln \left(\frac{z_i}{g_m(\mathbf{z})} \right) \left(\ln \frac{y_i}{g_m(\mathbf{y})} \right),$$

where $g_m(\mathbf{z})$ denotes the geometric mean of the components of \mathbf{z} .

Definition 6 (Aitchison norm).

$$\|\mathbf{z}\|_a = \sqrt{\langle \mathbf{z}, \mathbf{z} \rangle_a}.$$

Definition 7 (Aitchison distance).

$$d_a(\mathbf{z}, \mathbf{y}) = \|\mathbf{z} \ominus \mathbf{y}\|_a,$$

where $\mathbf{z} \ominus \mathbf{y}$ is equal to the perturbation $\mathbf{z} \oplus ((-1) \odot \mathbf{y})$.

2.1.3 Logratio coordinates

Aitchison (1986) proposed transformations based on ratios, including alr transformation and clr transformation. By the Aitchison's approach, it is possible to give an algebraic-geometric foundation and based on this framework, a transformation of coefficients is equivalent to express observations in a different coordinate system (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015). The principal logratio coordinates (alr, clr and ilr) are defined below. The alr transformation is defined as follows.

Definition 8 (Additive logratio coordinates). Let $\mathbf{z} = [z_1, z_2, \dots, z_D]$ be a composition in \mathbb{S}^D and consider z_D as a reference part. Its alr transformation into \mathbb{R}^{D-1} is

$$\text{alr}(\mathbf{z}) = \left[\ln \frac{z_1}{z_D}, \ln \frac{z_2}{z_D}, \dots, \ln \frac{z_{D-1}}{z_D} \right] = \boldsymbol{\zeta}.$$

To recover \mathbf{z} from $\boldsymbol{\zeta} = [\zeta_1, \zeta_2, \dots, \zeta_{D-1}]$, the inverse alr transformation is given through by closure definition

$$\mathbf{z} = \text{alr}^{-1}(\boldsymbol{\zeta}) = C[\exp(\zeta_1), \exp(\zeta_2), \dots, \exp(\zeta_{D-1}), 1].$$

As the reference part z_D is in the denominator of the components logratio, the alr transformation is not symmetric in the components. Another option of the reference part can be chosen, conducting it to different alr-transformations. On the other hand, alr coordinates cannot compute the Aitchison inner products or distances in the standard Euclidean way, that is, the alr does not supply an isometry between \mathbb{S}^D and \mathbb{R}^{D-1} . Each part of the composition except for the part in the denominator of the alr is

$$z_i = \frac{\exp(\zeta_i)}{1 + \sum_{j=1}^{D-1} \exp(\zeta_j)},$$

where the denominator is the effect of the closure. Its term *additive* comes from the denominator, which is the sum of the exponentials.

The clr coordinates give the expression of a composition in centered logratio coefficients

$$\text{clr}(\mathbf{z}) = \left[\ln \frac{z_1}{g_m(\mathbf{z})}, \ln \frac{z_2}{g_m(\mathbf{z})}, \dots, \ln \frac{z_D}{g_m(\mathbf{z})} \right] = \boldsymbol{\xi}.$$

The clr transformation is symmetric in the components, but the sum of the components is zero. In addition, the covariance matrix of $\text{clr}(\mathbf{z})$ is singular, that is, the determinant is zero.

Moreover, the clr coefficients are not subcompositionally coherent, because the geometric mean of the parts of a subcomposition $g_m(\mathbf{z})$ is not necessarily equal to that of the full composition. A formal definition of the clr coefficients is given as follows.

Definition 9 (Centered logratio coefficients). For a composition $\mathbf{z} \in \mathbb{S}^D$, the clr coefficients are the components of the unique vector $\boldsymbol{\xi} = [\xi_1, \xi_2, \dots, \xi_D] = \text{clr}(\mathbf{z})$, satisfying the two conditions

$$\mathbf{z} = \text{clr}^{-1}(\boldsymbol{\xi}) = C(\exp(\boldsymbol{\xi})) \quad \text{and} \quad \sum_{i=1}^D \xi_i = 0.$$

The i th clr coefficient is

$$\xi_i = \ln \frac{z_i}{g_m(\mathbf{z})}.$$

The more recent logratio coordinates are ilr coordinates. The main idea of ilr coordinates is to obtain an orthonormal basis on the simplex, and to apply the new coordinates in a linear regression model. There are many ways to construct such a basis (CHEN; ZHANG; LI, 2017). Specifically, an example of a basis for compositional data is called sequential binary partitioning (HRON; FILZMOSER; THOMPSON, 2012). We obtain coordinates which are interpreted in terms of the included compositional parts. For a given matrix

$$\mathbf{W}_{D \times (D-1)} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{D-1}) = \begin{pmatrix} \sqrt{\frac{D-1}{D}} & 0 & \dots & 0 \\ -\frac{1}{\sqrt{D(D-1)}} & \sqrt{\frac{D-2}{D-1}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{\sqrt{D(D-1)}} & -\frac{1}{\sqrt{(D-1)(D-2)}} & \dots & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{D(D-1)}} & -\frac{1}{\sqrt{(D-1)(D-2)}} & \dots & -\frac{1}{\sqrt{2}} \end{pmatrix}, \quad (2.1)$$

and

$$\mathbf{e}_i = C(\exp \mathbf{w}_i), \quad i = 1, \dots, D-1,$$

is the corresponding orthonormal basis $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ and then the transformation of the composition $\mathbf{z} = (z_1, \dots, z_D)^\top \in \mathbb{S}^D$ to the ilr coordinates $\text{ilr}(\mathbf{z}) = (\text{ilr}(\mathbf{z})_1, \dots, \text{ilr}(\mathbf{z})_{D-1})^\top \in \mathbb{R}^{D-1}$ is obtained by

$$v_i = \text{ilr}(\mathbf{z})_i = \sqrt{\frac{D-i}{D-i+1}} \ln \left(\frac{z_i}{\sqrt[D-i]{\prod_{j=i+1}^D z_j}} \right), \quad i = 1, \dots, D-1. \quad (2.2)$$

The inverse ilr transformation of (2.2) is given by

$$\begin{aligned} z_1 &= \exp \left\{ \sqrt{\frac{D-1}{D}} v_1 \right\}, \\ z_i &= \exp \left\{ -\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} v_j + \sqrt{\frac{D-i}{D-i+1}} v_i \right\}, \quad i = 2, \dots, D-1, \\ z_D &= \exp \left\{ -\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} v_j \right\}. \end{aligned} \quad (2.3)$$

Considering the scale invariance property, the composition \mathbf{z} (2.3) can be represented by vectors with a chosen constant sum constraint. An important relationship between ilr and clr coordinates of composition \mathbf{z} (CHEN; ZHANG; LI, 2017) is defined by

$$\text{ilr}(\mathbf{z}) = \mathbf{W}_D^\top \text{clr}(\mathbf{z}) = \mathbf{W}_D^\top \log(\mathbf{z}),$$

where \mathbf{W}_D is defined in (2.1). Specifically, the first coordinates of $\text{ilr}(\mathbf{z})$ and $\text{clr}(\mathbf{z})$ present a linear relation as

$$\text{ilr}(\mathbf{z})_1 = \sqrt{\frac{D}{D-1}} \text{clr}(\mathbf{z})_1 = \frac{1}{D(D-1)} \left(\ln\left(\frac{z_1}{z_2}\right) + \dots + \ln\left(\frac{z_1}{z_D}\right) \right).$$

The coordinate $\text{ilr}(\mathbf{z})_1$ extracts all relative information regarding z_1 and obtains the relative contribution of z_1 respecting all the other parts (HRON; FILZMOSE; THOMPSON, 2012). Some properties of the ilr coordinates are expressed below to explain their potential application and computation. Let the function $\text{ilr}: \mathbb{S}^D \rightarrow \mathbb{R}^{D-1}$ an isometry of vector spaces and the asterisk (*) denotes coordinates in an orthonormal basis.

Property 1. Consider $\mathbf{z}_h \in \mathbb{S}^D$, $h = 1, 2$ and real constants α, γ ,

- (a) $\text{ilr}(\alpha \odot \mathbf{z}_1 \oplus \gamma \odot \mathbf{z}_2) = \alpha \cdot \text{ilr}(\mathbf{z}_1) + \gamma \cdot \text{ilr}(\mathbf{z}_2) = \alpha \cdot \mathbf{z}_1^* + \gamma \cdot \mathbf{z}_2^*$;
- (b) $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle_a = \langle \text{ilr}(\mathbf{z}_1), \text{ilr}(\mathbf{z}_2) \rangle = \langle \mathbf{z}_1^*, \mathbf{z}_2^* \rangle$;
- (c) $\|\mathbf{z}_1\|_a = \|\text{ilr}(\mathbf{z}_1)\| = \|\mathbf{z}_1^*\|$;
- (d) $d_a(\mathbf{z}_1, \mathbf{z}_2) = d(\text{ilr}(\mathbf{z}_1), \text{ilr}(\mathbf{z}_2)) = d(\mathbf{z}_1^*, \mathbf{z}_2^*)$.

In this thesis, we applied the ilr coordinates to the compositional data in order to avoid numerical problems in the context of linear models. For clr and ilr coordinates, the scalar product is preserved and they are isometric, but this fact is different with alr coordinates, that is,

$$\langle \mathbf{z}, \mathbf{y} \rangle = \text{clr}(\mathbf{z}) \cdot \text{clr}^\top(\mathbf{y}) = \text{ilr}(\mathbf{z}) \cdot \text{ilr}^\top(\mathbf{y}) \neq \text{alr}(\mathbf{z}) \cdot \text{alr}^\top(\mathbf{y}).$$

A disadvantage of using the alr coordinates is that they should not be applied when there are distances, angles and shapes involved. In addition, the clr coordinates provide singular covariance matrices, a problem for estimation in linear models (BOOGAART; TOLOSANA-DELGADO, 2013).

2.2 Shrinkage Methods

The current section considers the three penalization methods used to develop the proposed models for compositional data. This involves the Lasso, elastic net and SSL.

First, we consider the generic and classical linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2.4)$$

where $\mathbf{y} \in \mathbb{R}^n$ is a vector of responses, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a regression matrix of p predictors, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ is a vector of unknown regression coefficients and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the independent noise vector distributed as $N_n(0, \sigma^2 \mathbf{I}_n)$ being \mathbf{I}_n an identity matrix with dimension n . To solve the estimation problem, the ordinary least squares (OLS) method is often used where the parameters are estimated by the minimization of the residual sum of squares $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$. This method can be applied under specific conditions, that is, $\mathbf{X}^\top \mathbf{X}$ is nonsingular and consequently we obtain $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. However, problems with high-dimensional regression are common in a wide range of applications, that is, when we have a large number of covariates p to a response of interest, which exceeds the number of observations n , $p > n$ or even $p \gg n$.

2.2.1 Lasso

Tibshirani (1996) proposed Lasso (least absolute shrinkage and selection operator) method or l_1 regularization which has become very popular for high-dimensional estimation problems taking into account its statistical accuracy for prediction and variable selection jointly with its computational feasibility. Furthermore, its theoretical properties in high-dimensional regression are well-understood (LIN *et al.*, 2014).

Lasso is known as a penalized likelihood approach that develops the methodology for l_1 -penalization in high-dimensional settings with desirable properties for $p \gg n$ problems. In such problems, lasso demonstrated its superiority compared to other existing methods.

Assuming the regression model (2.4), the convex optimization problem with the application the Lasso is defined as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right), \quad (2.5)$$

where λ is a tuning parameter dealing with the amount of shrinkage, and $\|\cdot\|_2$ and $\|\cdot\|_1$ are the l_2 and l_1 norms, respectively. Not only does the l_1 penalty shrink the coefficients toward zero, but it also has some advantages in relation to the classical Ordinary Least Squares (OLS) methods such as some criteria for model selection, resulting in convexity of the optimization problem and solving large problems efficiently (HASTIE; TIBSHIRANI; WAINWRIGHT, 2015).

When the lasso estimates regression coefficients to zero, it is creating a *sparse* solution, that is, only few of the regression coefficients are nonzero. This performance is important due to the variable selection that determines relevant covariates showing the strongest effects. The Lasso results in sparsity and ridge penalty is not sparse are presented in Figure ?? in a geometrical picture when there are only two parameters, which is given by the constraint interpretation of their penalties. We can observe that the Lasso estimate can be set to zero.

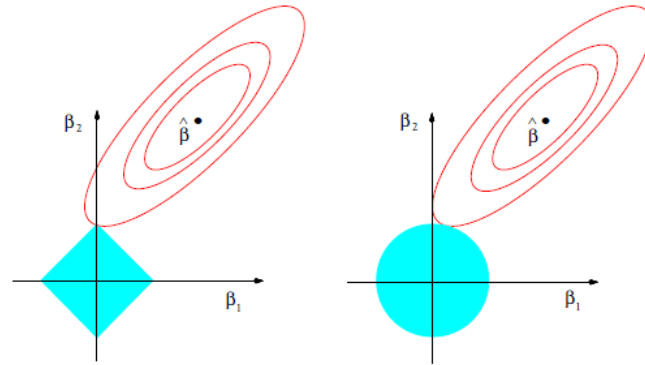


Figure 1 – Estimation picture for the Lasso (left) and ridge regression (right).

Source: (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.2.1.1 Orthonormal design

Following Buhlmann and Geer (2011), the lasso estimator can be derived for an orthonormal design case. Assuming uncorrelated variables implies that $\mathbf{X}_i^\top \mathbf{X}_j = 0$ for each $i \neq j$ and $\frac{1}{n} \mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$. Thus, the lasso estimator is given by

$$\hat{\boldsymbol{\beta}} = S(\hat{\boldsymbol{\beta}}^{LSE}; \lambda), \quad (2.6)$$

where $S(\cdot; \lambda)$ is the soft-thresholding operator

$$S(t; \lambda) = \text{sgn}(t)(|t| - \lambda)_+ = \begin{cases} t - \lambda, & \text{if } t > \lambda, \\ 0, & \text{if } |t| \leq \lambda, \\ t + \lambda, & \text{if } t < -\lambda, \end{cases}$$

where sgn denotes the sign of its argument (± 1), $(t)_+ = \max(t, 0)$ denotes the positive part and $\hat{\boldsymbol{\beta}}^{LSE}$ is the OLS estimator for $\boldsymbol{\beta}$.

2.2.1.2 K-fold Cross-validation

The k -fold cross-validation scheme is commonly used to select a reasonable tuning parameter λ for the Lasso estimator. First, we randomly divide the observations into k groups. One group is fixed as the test set, and the $k - 1$ groups are designated as a training set. The model is fitted to the training data for a range of values of λ in a grid, and we predict the responses in the test set based on each fitted model, saving the mean-squared prediction errors. We repeat this process k times, where the k groups have a chance to be test data, in relation to $k - 1$ groups used as a training set. Thus, we capture k different estimates of the prediction error over a range of values of λ (HASTIE; TIBSHIRANI; WAINWRIGHT, 2015). The choice of k is usually 5 or 10, where $k = 10$ is very common in the field of applied machine learning.

2.2.1.3 Coordinate descent algorithm

Based on the estimator (2.5), there is no closed form expression for the estimates for the lasso. Indeed, the optimization problem become a convex problem with inequality constraints (FRIEDMAN *et al.*, 2007). Since the seminal work of Tibshirani (1996), computational developments have been approached to obtain efficiency and solutions to solve the lasso problem.

The LARS algorithm was proposed by Efron *et al.* (2004), which is a useful and less greedy version of traditional forward selection methods.

On the other hand, another fast and popular approach used for estimation in regularization methods is the coordinate descent algorithm (FU, 1998), which has shown to be a strong competitor to the LARS algorithm (FRIEDMAN *et al.*, 2007). The idea of the algorithm is to fix the penalty parameter λ in the Lagrangian form (2.5) and optimize successively over each parameter, keeping the other parameters fixed at their actual values. For more details, see for example Hastie, Tibshirani and Friedman (2009).

2.2.2 Elastic net

The elastic net approach was proposed by Zou and Hastie (2005). According to the authors, the method is similar to the lasso, in view of being a variable selection and continuous shrinkage. Therefore, this method selects groups of correlate variables.

Thus, the elastic net combines the ridge (HOERL; KENNARD, 1970) and lasso penalties to solve the following convex problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left[\frac{1}{2} (1 - \alpha) \|\boldsymbol{\beta}\|_2 + \alpha \|\boldsymbol{\beta}\|_1 \right] \right),$$

where $\alpha \in [0, 1]$ is an elastic net tuning parameter that controls the mixing between the l_1 and l_2 penalties. There are many alternatives of algorithms to solve the elastic net problem. Within them, the coordinate descent is efficient due to the fact that the updates will be a simple extension of lasso (HASTIE; TIBSHIRANI; WAINWRIGHT, 2015).

2.2.3 Spike-and-Slab Lasso

Under a Bayesian perspective, Rocková and George (2018) proposed the SSL for high-dimensional normal linear models, and showed that it has important properties.

The general form of the penalized likelihood approach estimates $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta} \in \mathbb{R}^D} \left\{ -\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \operatorname{pen}_\lambda(\boldsymbol{\beta}) \right\},$$

where $\operatorname{pen}_\lambda(\boldsymbol{\beta})$ is a penalty function that prioritizes suitable solutions.

The SSL involves placing a mixture prior on the regression coefficients $\boldsymbol{\beta}$, where each β_j , $j = 1, \dots, D$ is assumed a priori to be drawn from either a Laplacian ‘‘Spike’’ concentrated

around zero (and hence is considered negligible), or a diffuse Laplacian ‘‘Slab’’ (and hence may be large). Thus, the hierarchical prior over $\boldsymbol{\beta}$ and the latent indicator variables $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_D)$ is given by

$$\begin{aligned}\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) &= \prod_{j=1}^D [\gamma_j \psi_1(\beta_j) + (1 - \gamma_j) \psi_0(\beta_j)], \quad \boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma}), \\ \pi(\boldsymbol{\gamma}|\boldsymbol{\theta}) &= \prod_{j=1}^D \theta^{\gamma_j} (1 - \theta)^{1 - \gamma_j}, \quad \text{and} \quad \boldsymbol{\theta} \sim \text{Beta}(a, b),\end{aligned}\tag{2.7}$$

where $\psi_1(\beta_j) = (\lambda_1/2)e^{-|\beta_j|\lambda_1}$ is the Slab distribution, $\psi_0(\beta_j) = (\lambda_0/2)e^{-|\beta_j|\lambda_0}$ is the Spike distribution ($\lambda_1 \ll \lambda_0$) and the beta-binomial prior has been used for the latent indicators. The Figure 2 illustrates the spike and slab distribution for different values of λ_0 . In this thesis, we

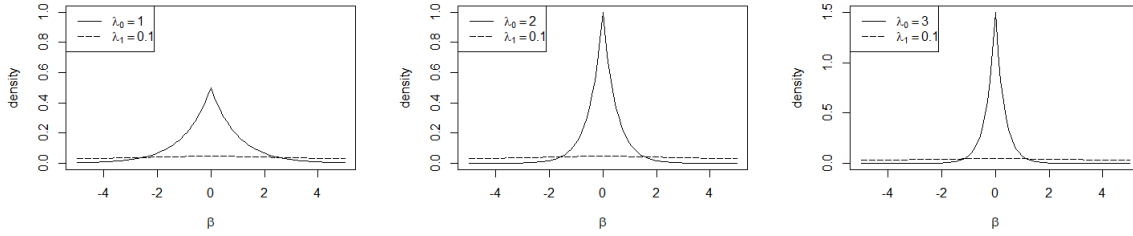


Figure 2 – Spike and Slab distributions for $\lambda_0 = 1, 2, 3$ and $\lambda_1 = 0.1$.

applied two types of SSL penalties studied in [Rocková and George \(2018\)](#): separable SSL and non-separable SSL. The first one is the separable SSL penalty that arises from an independent product prior (2.7), assuming $\boldsymbol{\theta}$ known, that is, it is fixed. Its definition is given below.

$$\text{pen}_S(\boldsymbol{\beta}|\boldsymbol{\theta}) = \sum_{j=1}^D \rho(\beta_j|\boldsymbol{\theta}) = -\lambda_1 |\boldsymbol{\beta}| + \sum_{j=1}^D \log \left(\frac{p_{\boldsymbol{\theta}}^*(0)}{p_{\boldsymbol{\theta}}^*(\beta_j)} \right),$$

where

$$p_{\boldsymbol{\theta}}^*(\beta_j) = \frac{\theta \psi_1(\beta_j)}{\theta \psi_1(\beta_j) + (1 - \theta) \psi_0(\beta_j)}$$

and

$$p_{\boldsymbol{\theta}}^*(0) = \frac{\theta \psi_1(0)}{\theta \psi_1(0) + (1 - \theta) \psi_0(0)} = \frac{\theta \lambda_1}{\theta(\lambda_1 - \lambda_0) + \lambda_0}.$$

Another one is the non-separable SSL penalty with unknown variance proposed by [Moran, Rocková and George \(2018\)](#). This penalty treats the $\boldsymbol{\theta}$ as a random, avoiding the need for cross-validation over $\boldsymbol{\theta}$. Thus, the non-separable SSL penalty with $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ is defined by

$$\text{pen}_{NS}(\boldsymbol{\beta}) = \log \left[\frac{\pi(\boldsymbol{\beta})}{\pi(\mathbf{0}_p)} \right] = -\lambda_1 |\boldsymbol{\beta}| + \log \left[\frac{\int \frac{\theta^p}{\prod_{j=1}^p p_{\boldsymbol{\theta}}^*(\beta_j)} d\pi(\boldsymbol{\theta})}{\int \frac{\theta^p}{\prod_{j=1}^p p_{\boldsymbol{\theta}}^*(0)} d\pi(\boldsymbol{\theta})} \right].$$

All the penalized methods presented in this section were implemented in the software R ([R Core Team, 2017](#)). Some examples of the routines for each penalized method used in this work were in Appendix A.

PENALIZED REGRESSION MODEL FOR COMPOSITIONAL RESPONSE VARIABLES

In this Chapter, we present the penalized regression model with restrictions in the response variable, that is, in the vector \mathbf{y} .

The model into a multivariate regression problem with compositional response is defined as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (3.1)$$

where \mathbf{y} is a vector ($D \times 1$) of compositional response variables, \mathbf{X} is a matrix ($D \times p$) of p covariates, where D is the number of the components, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is a vector ($p \times 1$) of unknown parameters and $\boldsymbol{\varepsilon}$ is the noise vector with distribution $N_D(\mathbf{0}, I_D)$, with a known variance $\sigma_2 = 1$. The intercept of the model is not included, since the response and predictor variables can be centered.

Based on the principle of working in coordinates, we can rewrite the model (3.1) as

$$\begin{aligned} \mathbf{y} &= \mathbf{X} \odot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \text{ilr}(\mathbf{y}) &= \mathbf{X}\boldsymbol{\beta} + \text{ilr}(\boldsymbol{\varepsilon}), \end{aligned} \quad (3.2)$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}_{D-1}, \Sigma_{\text{ilr}})$.

Here, we assume the following estimators for $\boldsymbol{\beta}$ of the regression models with compositional responses focused on regularization methods presented in Section 2. We considered the Lasso, elastic net, SSL with separable and non-separable penalty approaches, respectively, for the model (3.2) as follows.

1. Lasso:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} (\|\text{ilr}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}\|_2^2/n + \lambda \|\boldsymbol{\beta}\|_1) \quad (3.3)$$

where $\|\text{ilr}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^n (\text{ilr}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta})^2$ and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$.

2. Elastic net:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\frac{1}{n} \|\operatorname{ilr}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left[\frac{1}{2} (1 - \alpha) \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right] \right). \quad (3.4)$$

3. SSL with separable penalty (known variance):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{D-1}}{\operatorname{argmax}} \left\{ -\frac{1}{2} \|\operatorname{ilr}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \left[-\lambda_1 |\boldsymbol{\beta}| + \sum_{j=1}^p \log \left(\frac{p_{\theta}^*(0)}{p_{\theta}^*(\beta_j)} \right) \right] \right\}. \quad (3.5)$$

4. SSL with non-separable penalty (unknown variance):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{D-1}}{\operatorname{argmax}} \left\{ -\frac{1}{2} \|\operatorname{ilr}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \left(-\lambda_1 |\boldsymbol{\beta}| + \log \left[\frac{\int \frac{\theta^p}{\prod_{j=1}^p p_{\theta}^*(\beta_j)} d\pi(\theta)}{\int \frac{\theta^p}{\prod_{j=1}^p p_{\theta}^*(0)} d\pi(\theta)} \right] \right) \right\}. \quad (3.6)$$

For the estimation of the $\boldsymbol{\beta}$'s, we implemented the estimators (3.3) and (3.4) through by R package `glmnet` (FRIEDMAN; HASTIE; TIBSHIRANI, 2010). The algorithm used to find the minimum was cyclical coordinate descent. Such algorithm computes a grid of possible value of λ and a sequence of models related to the loss function is provided as output. One advantage of this algorithm is that it can implemented for generalized linear model. The estimators (3.5) and (3.6) were obtained through by R package `SSLASSO` (MORAN; ROCKOVÁ; GEORGE, 2018). The coordinate descent algorithm is used to fit the sequence of models indexed by the regularization parameter λ_0 .

3.1 Simulation Analysis

Here, we provided the simulation studies to investigate the efficacy of the penalized methods for the regression model with compositional responses variables. We replicated the simulation 1000 times and the results were summarized based on these replicates (Table 3 for $\operatorname{ilr}(y_1)$ and Table 4 for $\operatorname{ilr}(y_2)$). We generated a data matrix \mathbf{X} from a normal distribution with mean 0 and $\sigma = 2$ for each element of matrix \mathbf{X} . The compositional response variable is generated according to model (3.1), from a logistic normal distribution with mean $\mathbf{0}_D$ and covariance matrix $\Sigma = (\rho^{|i-j|})$, with $\rho = 0.2$ and $\rho = 0.5$, for $j = 1, \dots, D$. We assume $D = 3$, that is, we have 3 components (y_1, y_2, y_3) of a composition. The fixed values for the parameters $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*)$ were $\beta_1^* = (-2, -1.5, -1, 0, 1, 1.5, 2, 0, \dots, 0)^\top$ and $\beta_2^* = (2, -1, -2.5, 0, 1, -1, 0.5, 0, \dots, 0)^\top$ to $q = 6$ random directions (non-zero coefficients).

We assumed three scenarios with a different number of sample sizes and covariates: $(n, p) = (50, 30), (100, 200)$ and $(100, 1000)$. The adopted performance measures for our comparisons were the Mean Square Error (MSE) given by $\operatorname{MSE}(\hat{\boldsymbol{\beta}}^*) = \operatorname{Var}(\hat{\boldsymbol{\beta}}^*) + (\operatorname{Bias}(\hat{\boldsymbol{\beta}}^*))$, where $\operatorname{Var}(\hat{\boldsymbol{\beta}}^*)$ is the variance of the estimates of $\boldsymbol{\beta}^*$ and $\operatorname{Bias}(\hat{\boldsymbol{\beta}}^*) = \hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*$; the number of false positive (FP), the number of false negative (FN), where positive and negative refer to nonzero

and zero coefficients, respectively; and the Hamming (HAM) distance between the support of the estimated β and the true β^* , that is, suppose two vectors $x = (1, 0, 0)$ and $w = (0, 1, 0)$, the HAM distance $d(x, w)$ (number of different elements) between this two vectors, being that in this case $d(x, w) = 2$ because the first and second elements of these vectors are different from each other. In this way, lower values of HAM indicate better performance of the method. The Tables 3 and 4 report the averages of these performance measures for the five penalized methods adopted in this work.

According to the results in Tables 3 and 4, we can see that, in general, the SSL Separable ($SSL(1, 6/p)$ with $\sigma = 1$ fixed) performs better than other methods in all settings based on the HAM distance (lower values), for $p = 30, 200$ and 1000 covariates. Therefore, this method tends to select fewer FP compared with other penalized methods. Among the approaches studied, the elastic net has a worse performance in almost all settings.

Table 3 – Averages of some performance measures for penalized methods with compositional response variable ($\text{ilr}(y_1)$).

(n, p)	Method	MSE	FP	FN	HAM
$\rho=0.2$					
SSL (λ_1, θ)					
(50, 30)	SSL (1, 0.8) with $\sigma=1$ fixed	0.4845	0.1850	5.9550	6.2020
	SSL (1, 6/30) with $\sigma=1$ fixed	0.4834	0.0020	6.0000	6.0020
	SSL (1, 6/30) with unknown σ	0.4834	0.0030	5.0190	6.0030
	Lasso	0.4854	4.1180	5.7020	10.3090
	Elastic Net	0.4850	6.3010	5.4880	12.6020
(100, 200)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0733	1.9700	5.9540	7.9780
	SSL (1, 6/200) with $\sigma = 1$ fixed	0.0725	0.0020	6.0000	6.0020
	SSL (1, 6/200) with unknown σ	0.0769	31.9060	6.0000	38.0670
	Lasso	0.0726	9.9130	5.7020	15.9760
	Elastic Net	0.0726	16.4320	5.4880	22.5360
(100, 1000)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0149	4.7620	5.9750	10.6770
	SSL (1, 6/1000) with $\sigma = 1$ fixed	0.0145	0.0000	6.0000	6.0010
	SSL (1, 6/1000) with unknown σ	0.0153	0.0160	6.0000	6.0160
	Lasso	0.0145	12.2770	5.9280	18.2910
	Elastic Net	0.0145	23.3260	5.8590	32.9320
$\rho=0.5$					
SSL (λ_1, θ)					
(50, 30)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.4645	1.6550	5.5600	7.7260
	SSL (1, 6/30) with $\sigma = 1$ fixed	0.4523	0.1620	5.9530	6.1680
	SSL (1, 6/30) with unknown σ	0.4500	0.0140	5.9960	6.0140
	Lasso	0.4849	3.8260	5.0460	9.9790
	Elastic Net	0.4551	7.4290	4.1100	13.7340
(100, 200)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0742	4.0100	5.8640	10.0330
	SSL (1, 6/200) with $\sigma = 1$ fixed	0.0725	0.0090	6.0000	6.0090
	SSL (1, 6/200) with unknown σ	0.0725	0.0090	6.0000	6.0090
	Lasso	0.0726	9.0780	5.7120	15.1410
	Elastic Net	0.0726	15.1490	5.5680	21.2220
(100, 1000)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0152	7.2320	5.9640	13.2410
	SSL (1, 6/1000) with $\sigma = 1$ fixed	0.0145	0.0010	6.0000	6.0010
	SSL (1, 6/1000) with unknown σ	0.0145	0.0030	6.0000	6.0030
	Lasso	0.0145	11.1680	5.9370	17.1780
	Elastic Net	0.0145	21.5300	5.8740	27.5530

3.2 Real data application - ICMS dataset

The following dataset was made available by the *Secretaria da Fazenda* of Sao Paulo State. The dataset consists of ICMS (*Imposto sobre Circulação de Mercadorias e Prestação de Serviços*), which is the main revenue source for the Brazilian states.

The challenge with this dataset is the development of models which are disaggregated in three economic sectors: industry (y_1), commerce (y_2) and administered prices (y_3). These sectors

Table 4 – Averages of some performance measures for penalized methods with compositional response variable ($\text{ilr}(y_2)$).

(n, p)	Method	MSE	FP	FN	HAM
$\rho=0.2$					
SSL (λ_1, θ)					
(50, 30)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.4645	1.6550	5.5600	7.7260
	SSL (1, 6/30) with $\sigma = 1$ fixed	0.4523	0.1620	5.9530	6.1680
	SSL (1, 6/30) with unknown σ	0.4500	0.0140	5.9960	6.0140
	Lasso	0.4547	5.5720	4.5270	11.8300
	Elastic Net	0.4551	7.4290	4.1100	13.7340
(100, 200)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0749	11.9850	5.6220	18.0480
	SSL (1, 6/200) with $\sigma = 1$ fixed	0.0677	0.1440	5.9950	6.1450
	SSL (1, 6/200) with unknown σ	0.0723	8.3520	5.7350	14.3940
	Lasso	0.0679	11.5980	5.6290	17.6650
	Elastic Net	0.0678	19.5060	5.3710	25.6010
(100, 1000)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0150	12.3520	5.9200	18.3610
	SSL (1, 6/1000) with $\sigma = 1$ fixed	0.0135	0.0840	6.0000	6.0850
	SSL (1, 6/1000) with unknown σ	0.0135	0.0000	6.0000	6.0000
	Lasso	0.0136	14.8520	5.9120	20.8660
	Elastic Net	0.0135	26.9050	5.8120	32.9320
$\rho=0.5$					
SSL (λ_1, θ)					
(50, 30)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.4525	0.4250	5.8870	6.4460
	SSL (1, 6/30) with $\sigma = 1$ fixed	0.4502	0.0240	5.9960	6.0260
	SSL (1, 6/30) with unknown σ	0.4501	0.0060	5.9980	6.0060
	Lasso	0.4530	3.8230	4.9620	9.9770
	Elastic Net	0.4541	8.4840	3.7240	14.8560
(100, 200)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0694	4.1730	5.8720	10.1910
	SSL (1, 6/200) with $\sigma = 1$ fixed	0.0675	0.0040	5.9990	6.0040
	SSL (1, 6/200) with unknown σ	0.0675	0.0050	5.9990	6.0050
	Lasso	0.0678	13.9800	5.5570	20.0600
	Elastic Net	0.0677	23.4400	5.3080	29.5410
(100, 1000)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0142	7.1470	5.9520	13.1520
	SSL (1, 6/1000) with $\sigma = 1$ fixed	0.0135	0.0000	6.0000	6.0000
	SSL (1, 6/1000) with unknown σ	0.0135	0.0010	6.0000	6.0010
	Lasso	0.0136	17.7410	5.8700	23.7620
	Elastic Net	0.0135	33.1620	5.7680	39.2000

represent a linear combination, that is, they are defined as compositional data. The importance of disaggregating the sectors is to allow the government to forecast potential decreases in tax collection and plan efficient actions. The data were extracted from August 2007 to April 2018, that is, the sample size is $n = 128$ months. The Figure 3 presents the evolution of the ICMS series disaggregated in these three economic sectors: industry, commerce and administered price.

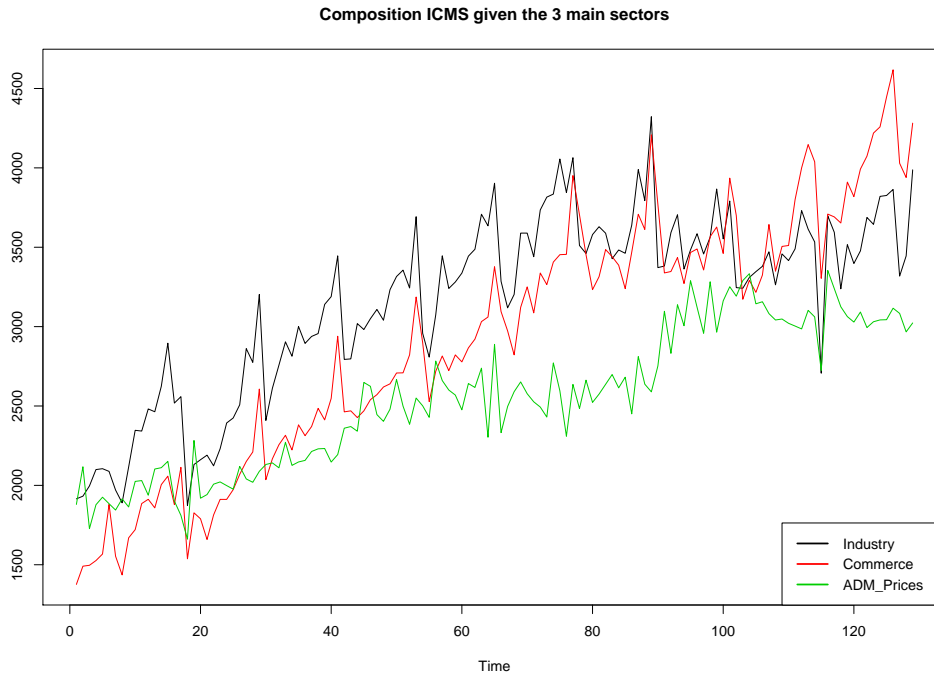


Figure 3 – ICMS series disaggregated in three economic sectors.

The exogenous variables or covariates provided by research institutes are:

- log of Monthly Industrial Survey (IBGE) - X_1 ;
- log of Monthly Trade Survey - PMC/IBGE - X_2 ;
- log of Monthly energy consumption in Sao Paulo State - X_3 ;
- Index of Economic Activity of the Central Bank - X_4 ;
- IGP-DI/FGV – General Price Index - X_5 .

Besides the covariates mentioned above, we also considered the lagged covariate in the period of 12 months of the proportion of ICMS in the industry (X_6) and commerce (X_7) and 6 months of the proportion of ICMC in the industry (X_8) and commerce (X_9). The total of covariates in the model is $p = 9$. Figures 4 and 5 present the solution path by the SSL (non-adaptative choice (separable), fixed θ ; non-adaptative oracle choice (separable); adaptative choice, $\theta \sim B(1, p)$ (non-separable)), Lasso and elastic net methods for modeling the ICMS disaggregated in 3 parts: industry, commerce and administered prices. These sectors represent compositional data, once they are dependent on each other. Thereby, the results showed the same performance for $\text{ilr}(y_1)$ and $\text{ilr}(y_2)$ when the SSL with separable penalties (Figures 4A, 4B, 5A and 5B), that is, these methods did not select any significant covariate for the model. On the other hand, SSL non-separable presented three significant covariates (X_6 , X_7 and X_8). The optimal λ calculated by the 10-fold cross-validation were 0.0036 ($\text{ilr}(y_1)$) and 0.0033 ($\text{ilr}(y_2)$) for the

Lasso method and 0.0076 ($\text{ilr}(y_1)$) and 0.0078 ($\text{ilr}(y_2)$) for the elastic net method (vertical line in Figures 4D, 4E, 5D and 5E). Moreover, Lasso method selected the covariates X_1 , X_2 and X_5 and elastic net method besides these covariates also selected X_3 considering the response variable $\text{ilr}(y_1)$. These results present the significant exogenous variables when the approached methods are applied considering compositional restriction on the response variable.

The models for each applied method is given by

1. SSL non-separable:

$$\begin{aligned} y_1 &= 0.268 * ICMSindustry12months + 0.280 * ICMScommerce12months \\ &\quad + 0.500 * ICMSindustry6months \\ y_2 &= -0.242 * \%ICMSindustry12months + 0.118 * \%ICMScommerce6months \end{aligned}$$

2. Lasso:

$$\begin{aligned} y_1 &= 0.002 * MonthlyIndustrialSurvey + 0.001 * MonthlyTradeSurvey - 0.002 * IGP - DI \\ y_2 &= 0.001 * MonthlyIndustrialSurvey + 0.004 * MonthlyTradeSurvey \\ &\quad + 0.003 * Monthlyenergyconsumption + 0.002 * IGP - DI \end{aligned}$$

3. Elastic net:

$$\begin{aligned} y_1 &= 0.002 * MonthlyIndustrialSurvey + 0.002 * MonthlyTradeSurvey \\ &\quad + 0.003 * Monthlyenergyconsumption - 0.002 * IGP - DI \\ y_2 &= 0.003 * MonthlyTradeSurvey + 0.004 * Monthlyenergyconsumption + 0.003 * IGP - DI \end{aligned}$$

3.3 Discussion

In this chapter, we presented a compositional regression model with restriction in the response variables under five penalties methods. We applied the ilr coordinates on the response variables to remove the dependence among the components.

A simulation study for the proposed model (3.2) showed that the model with SSL non-adaptative oracle choice (separable) performs better in terms of estimation if compared with the other penalized methods. It is noteworthy that this situation occurs in moderate dimensionality as in high-dimensionality.

In the case of application, the real data set involves the ICMC tax. As this data set is considered with moderate dimensionality, the SSL non-separable showed a better performance in relation to the other SSL penalties. The Lasso and elastic net estimators presented similar results, with only a little difference between the optimal λ . Based on these results, the SSL non-separable method considered the lagged covariates X_6 , X_7 and X_8 significant, that is, the

proportion of ICMS in the industry, commerce in the period of 12 months and the proportion of ICMS in the industry in the period of 6 months are relevant to explain the response variable $\text{ilr}(y_1)$ (proportion of the ICMS in the industry). On the other hand, the Lasso method considered only exogenous covariates significant, which are Monthly Industrial Survey, Monthly Trade Survey and IGP-DI/FGV General Price Index and for the elastic net method, besides these covariates mentioned above, including also the covariate monthly energy consumption in Sao Paulo State.

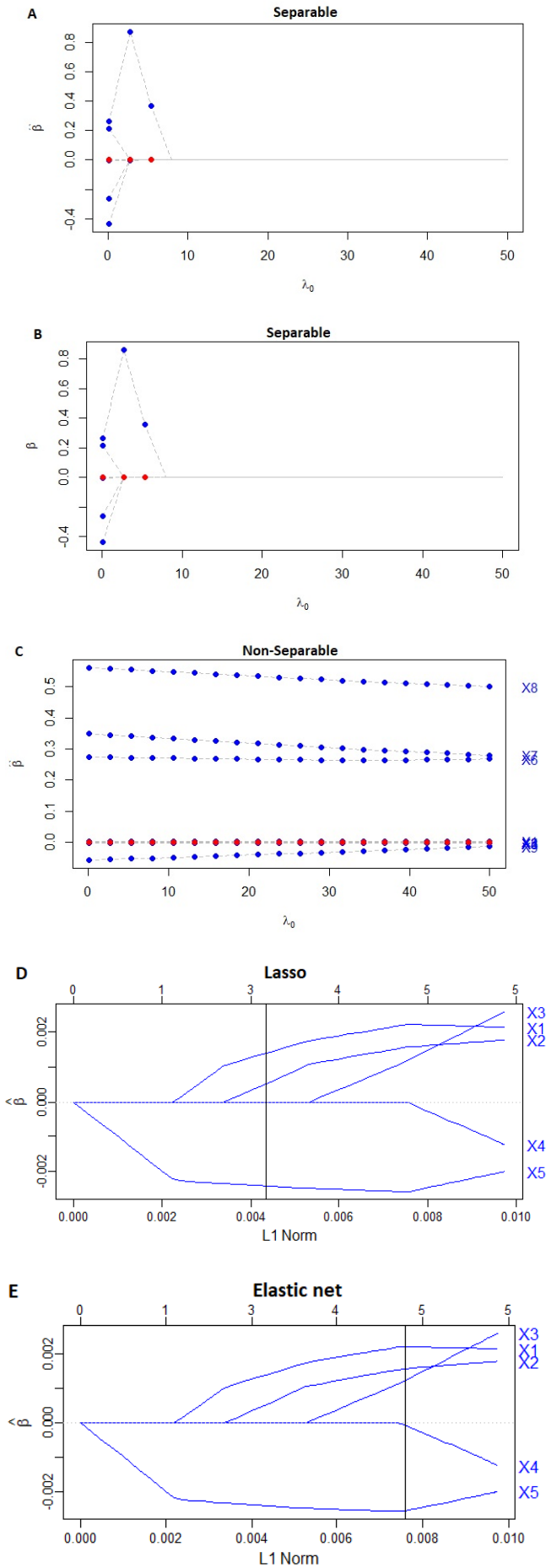


Figure 4 – The solution path SSL (A, B, C), Lasso (D) and elastic net (E) for $ilr(y_1)$. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model Lasso and elastic net (cross-validation), respectively.

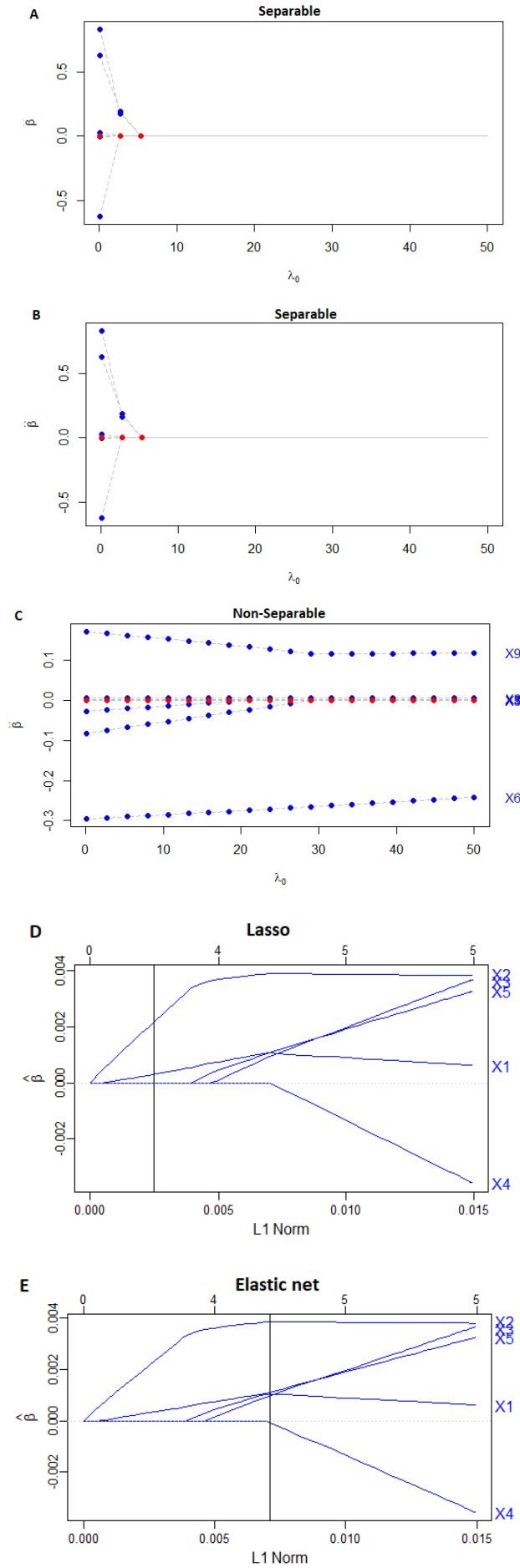


Figure 5 – The solution path SSL (A, B, C), Lasso (D) and elastic net (E) for $\text{ilr}(y_2)$. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model Lasso and elastic net (cross-validation), respectively.

PENALIZED REGRESSION MODEL FOR COMPOSITIONAL COVARIATES

In this section, we present the penalized regression model with compositional constraints in the covariates. The regression model based on methodology of compositional data is given by

$$\mathbf{y} = \text{ilr}(\mathbf{X})\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (4.1)$$

where \mathbf{y} is a vector ($l \times 1$) of response variables, \mathbf{X} is a matrix ($l \times D$) of D compositional covariates, where $l = 1, \dots, L$ and D is the number of the components, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_D)^\top$ is a vector ($D \times 1$) of unknown parameters and $\boldsymbol{\varepsilon}$ is the noise vector with distribution $N_l(\mathbf{0}, I_p)$, with a known variance $\sigma^2 = 1$. The intercept of the model is not included, equal to model 3.1.

Based on the principle of working in coordinates, we can rewrite the model (4.1) as

$$\begin{aligned} y &= \langle \boldsymbol{\beta}, \mathbf{X} \rangle_A + \boldsymbol{\varepsilon} \\ &= (\text{ilr}(\boldsymbol{\beta}), \text{ilr}(\mathbf{X})) + \boldsymbol{\varepsilon} \\ &= \sum_{k=1}^{D-1} \text{ilr}(\beta)_k \text{ilr}_k(X) + \boldsymbol{\varepsilon} \\ &= \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X) + \boldsymbol{\varepsilon}, \end{aligned} \quad (4.2)$$

with a vector of parameters $\boldsymbol{\beta} = \beta_k$ that afterwards might be mapped back to a composition through the inverse ilr transformation.

Now, we considered the following estimators for $\boldsymbol{\beta}$ of the regression models with compositional covariates focused on regularization methods presented in Section 2. We considered the lasso, elastic net, SSL with separable and non-separable penalty approaches, respectively, for the model (4.2) as follows.

1. Lasso:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\|\mathbf{y} - \sum_{k=1}^{D-1} \beta_k \mathbf{i}r_k(X)\|_2^2/n + \lambda \|\boldsymbol{\beta}\|_1 \right), \quad (4.3)$$

where $\|\mathbf{y} - \sum_{k=1}^{D-1} \beta_k \mathbf{i}r_k(X_i)\|_2^2 = \sum_{i=1}^n (y_i - \sum_{k=1}^{D-1} \beta_k \mathbf{i}r_k(X_i))^2$ and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{D-1} |\beta_j|$.

2. Elastic Net:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\frac{1}{n} \|\mathbf{y} - \sum_{k=1}^{D-1} \beta_k \mathbf{i}r_k(X)\|_2^2 + \lambda \left[\frac{1}{2} (1 - \alpha) \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right] \right). \quad (4.4)$$

3. SSL with separable penalty (known variance):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{D-1}}{\operatorname{argmax}} \left\{ -\frac{1}{2} \|\mathbf{y} - \sum_{k=1}^{D-1} \beta_k \mathbf{i}r_k(X)\|_2^2 + \left[-\lambda_1 \|\boldsymbol{\beta}\| + \sum_{j=1}^{D-1} \log \left(\frac{p_{\theta}^*(0)}{p_{\theta}^*(\beta_j)} \right) \right] \right\}. \quad (4.5)$$

4. SSL with non-separable penalty (unknown variance):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{D-1}}{\operatorname{argmax}} \left\{ -\frac{1}{2} \|\mathbf{y} - \sum_{k=1}^{D-1} \beta_k \mathbf{i}r_k(X)\|_2^2 + \left(-\lambda_1 \|\boldsymbol{\beta}\| + \log \left[\frac{\int \frac{\theta^{D-1}}{\prod_{j=1}^{D-1} p_{\theta}^*(\beta_j)} d\pi(\theta)}{\int \frac{\theta^{D-1}}{\prod_{j=1}^{D-1} p_{\theta}^*(0)} d\pi(\theta)} \right] \right) \right\}. \quad (4.6)$$

For the estimation of the $\boldsymbol{\beta}$'s, we implemented the estimators (4.3) and (4.4) through by R package `glmnet` (FRIEDMAN; HASTIE; TIBSHIRANI, 2010). The estimators (4.5) and (4.6) were obtained through by R package `SSLASSO` (MORAN; ROCKOVÁ; GEORGE, 2018).

4.1 Simulation Analysis

We provided the simulation studies to investigate the efficacy of the penalized methods for a regression model with compositional covariates. We replicated the simulation 1000 times and the results were summarized based on these replicates (Table 5). We generated a data compositional matrix \mathbf{X} of covariates from a logistic normal distribution with mean $\mathbf{0}_D$ and $\Sigma = (\rho^{|i-j|})$ with $\rho = 0.2$ and $\rho = 0.5$. The response is generated according to model (4.1) with $\boldsymbol{\beta}^* = \frac{1}{\sqrt{3}}(-2, -1.5, -1, 0, 1, 1.5, 2, 0, \dots, 0)^\top$ to $q = 6$ random directions (non-zeros coefficients).

We assumed three scenarios with a different number of sample size and covariates: $(n, p) = (50, 30), (100, 200)$ and $(100, 1000)$. The adopted performance measures for our comparisons were the MSE, the number of FP, the number of FN, and the HAM measure between the support of the estimated $\boldsymbol{\beta}$ and the true $\boldsymbol{\beta}^*$. Table 5 reports the averages of these performance measures for the five methods adopted.

As can be seen in Table 5, the SSL Separable Oracle $(1, 6/p)$ for $p = 30, 200, 1000$, performs better than other methods in all settings based on the mean squared error and false positives. The SSL Separable tends to select fewer false negatives in high dimensions, which is

acceptable because the omission of important variables is more relevant than the inclusion of shrunk variables.

Figures 6 and 7 present the estimates of coefficients over 1,000 replicates and the red circle represents the true value of coefficients. For both settings, Figures 6B and 7B show better accurate estimations.

Table 5 – Averages of some performance measures for penalized methods with compositional covariates.

(n, p)	Method	MSE	FP	FN	HAM
$\rho=0.2$					
SSL (λ_1, θ)					
(50, 30)	SSL Separable (1, 0.8)	0.0115	0.6190	0.0030	6.5620
	SSL Separable Oracle (1, 6/30)	0.0070	0.0290	0.0240	6.0250
	SSL (1, 6/30) with unknown σ	0.0079	0.1260	0.0210	6.1290
	Lasso	0.0258	4.1020	0.0030	10.1020
	Elastic Net	0.0262	6.4000	0.0000	12.3990
(100, 200)	SSL Separable (1, 0.8)	0.0032	4.7320	0.0000	10.7320
	SSL Separable Oracle (1, 6/200)	0.0004	0.0030	0.0000	6.0030
	SSL (1, 6/200) with unknown σ	0.0150	0.0180	1.7050	6.0180
	Lasso	0.0030	7.0350	0.0000	13.0350
	Elastic Net	0.0030	11.5890	0.0000	17.5890
(100, 1000)	SSL Separable (1, 0.8)	0.0010	7.4650	0.0000	13.4650
	SSL Separable Oracle (1, 6/1000)	0.0001	0.0000	0.0030	6.0000
	SSL (1, 6/1000) with unknown σ	0.0096	0.0000	4.4750	6.0000
	Lasso	0.0010	11.0290	0.0020	17.0290
	Elastic Net	0.0010	16.8400	0.0020	22.8390
$\rho=0.5$					
SSL (λ_1, θ)					
(50, 30)	SSL Separable (1, 0.8)	0.0184	0.5920	0.0430	6.5920
	SSL Separable Oracle (1, 6/30)	0.0165	0.0440	0.1840	6.0440
	SSL (1, 6/30) with unknown σ	0.0192	0.1430	0.2130	6.1430
	Lasso	0.0420	3.7790	0.0400	9.7790
	Elastic Net	0.0407	6.0690	0.0200	12.0690
(100, 200)	SSL Separable (1, 0.8)	0.0050	4.6700	0.0000	10.6700
	SSL Separable Oracle (1, 6/200)	0.0007	0.0030	0.0100	6.0030
	SSL (1, 6/200) with unknown σ	0.0092	0.0400	0.8090	6.0400
	Lasso	0.0049	6.9680	0.0010	12.9680
	Elastic Net	0.0049	11.9650	0.0000	17.9650
(100, 1000)	SSL Separable (1, 0.8)	0.0017	7.5630	0.0130	13.5630
	SSL Separable Oracle (1, 6/1000)	0.0002	0.0010	0.0510	6.0010
	SSL (1, 6/1000) with unknown σ	0.0078	0.0000	3.6610	6.0000
	Lasso	0.0015	10.6440	0.0250	16.6440
	Elastic Net	0.0016	16.6880	0.0140	22.6880

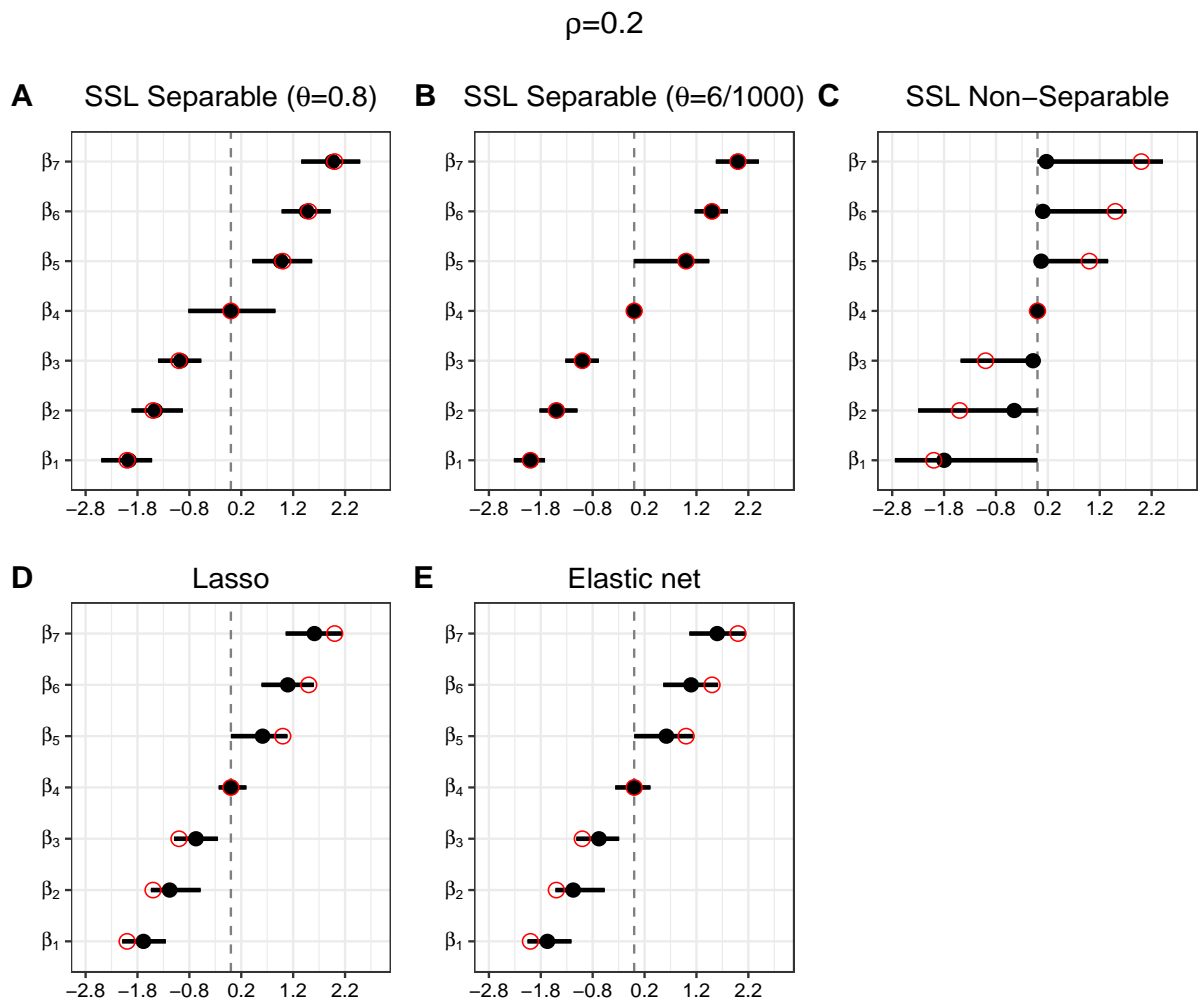


Figure 6 – The parameter estimation averaged over 1000 replicates assuming $\rho = 0.2$ for the covariance matrix ($n = 100$, $p = 1000$).

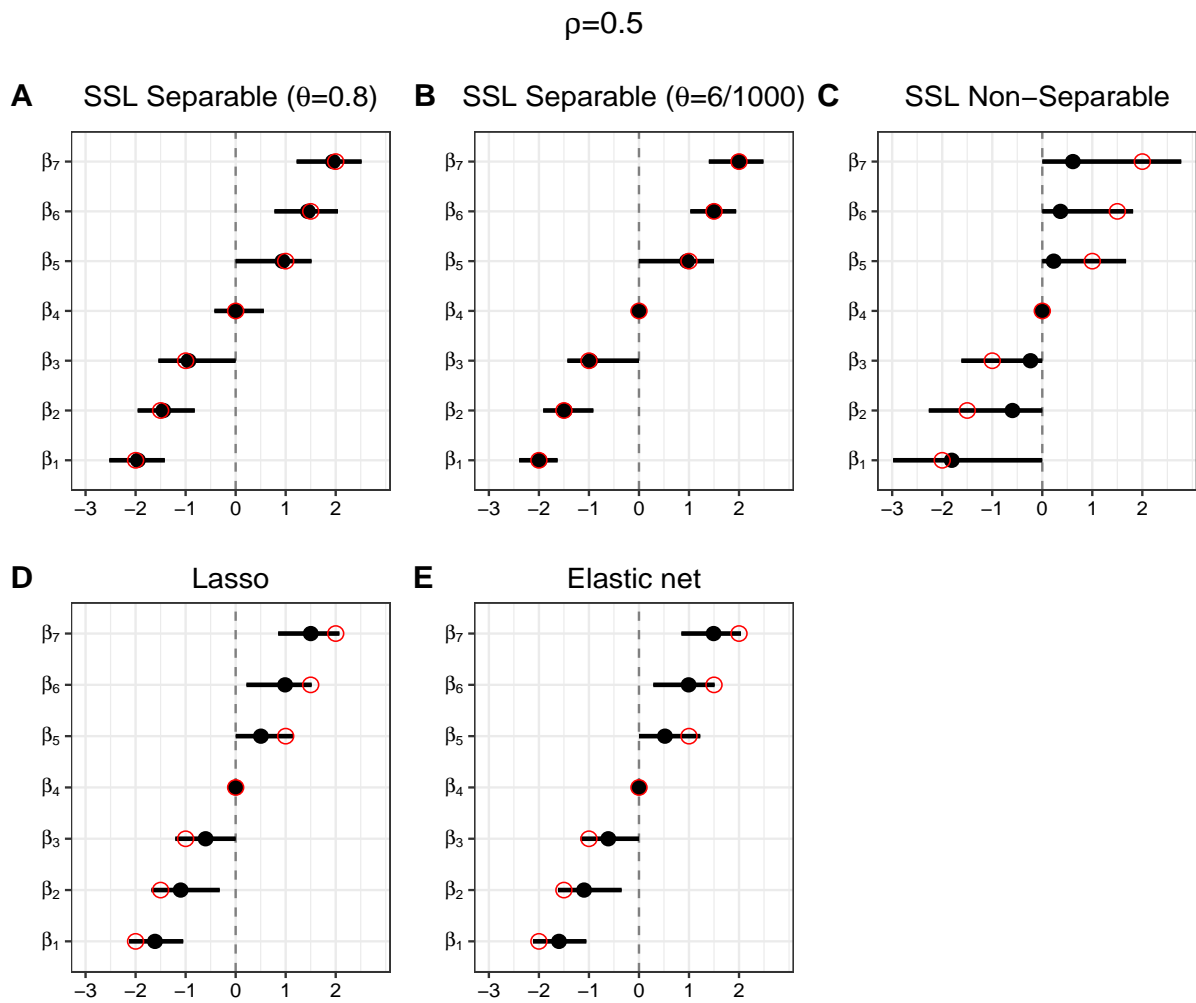


Figure 7 – The parameter estimation averaged over 1000 replicates assuming $\rho = 0.5$ for the covariance matrix ($n = 100, p = 1000$).

4.2 Artificial Data

A way of illustrating the proposed model (4.2) to compare the SSL, Lasso and elastic net penalties, we considered the following example. For $n = 100$ individuals with $D = 1000$ compositional covariates, we considered one generated sample of the simulation study presented in the last subsection.

We compared the SSL with a fixed variance with three settings: (i) separable choice $\theta = 0.8$, to verify the over-estimating of the true non-zero fraction $6/1000$, (ii) separable oracle choice $\theta = 6/1000$ and (iii) non-adaptative choice $\theta \sim B(1, D)$. The slab parameter was set to $\lambda_1 = 0.1$ and we used a ladder $\lambda_0 \in I = \{1, 2, \dots, 50\}$ for the spike parameter. In addition, we applied the generated data to the lasso and elastic net penalties implemented in the R package `glmnet` (FRIEDMAN; HASTIE; TIBSHIRANI, 2010). For this approach, an optimal value of λ was selected by 10-fold cross-validation. According to Figures 8 and 9, we can see the solution

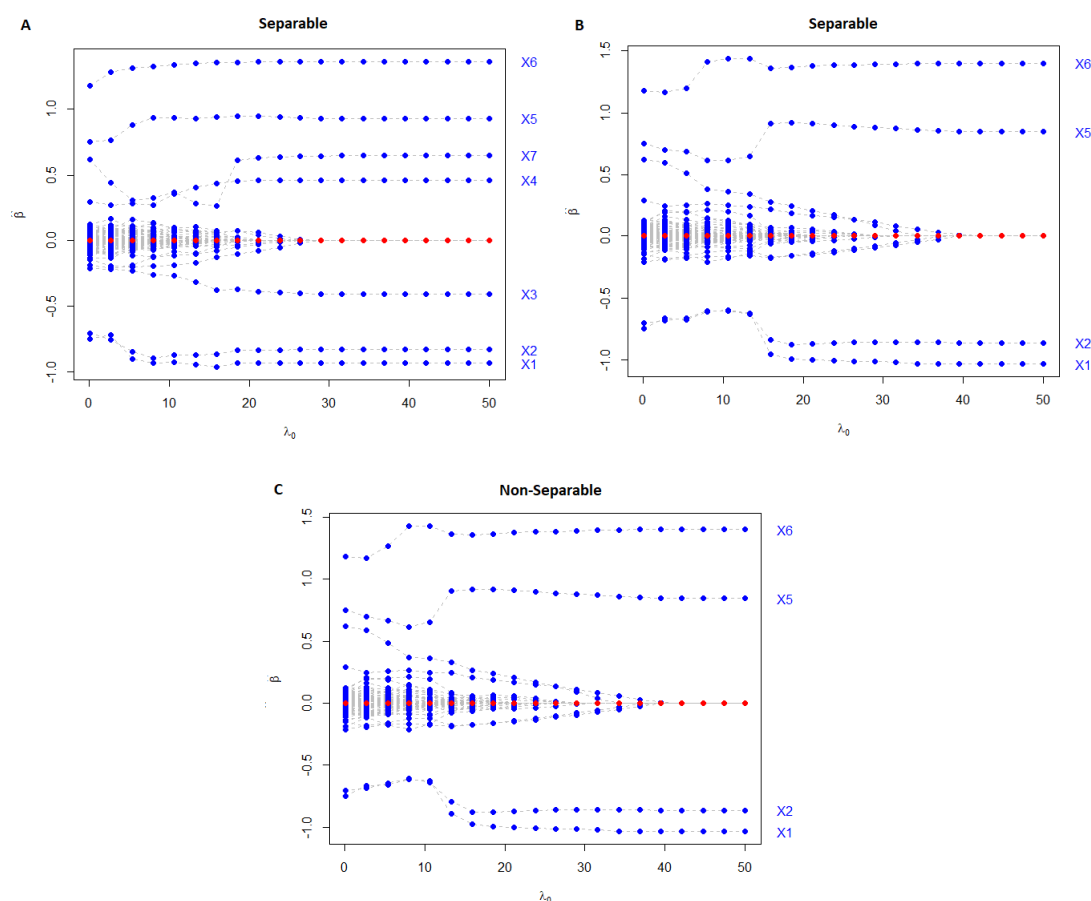


Figure 8 – The SSL solution paths (A, B, C).

paths for the five settings when we have the structure of compositional covariates. Each line represents a single regression coefficient and the horizontal dotted lines corresponds to the levels of true coefficients. The true coefficients are in blue and zero coefficients are in red. We can

observe that when θ is too large, there are some false positives (Figure 8A). Comparing oracle choice and when θ with distribution $Binomial(1, p)$, the solution path is similar between them. Moreover, the Spike-and-Slab lasso will keep the larger coefficients in the models. On the other hand, irrelevant coefficients are removed. Compared to Figure 9D, the lasso model included 6 nonzero coefficients (false negatives) when the optimal λ was 0.285. For the elastic net penalty for the model (4.2), we can see that it presents the same behaviour as the lasso penalty.

4.3 Real data application - Brazilian children malnutrition dataset

The SSL approach for the regression model with compositional covariates was applied to analyze the nutritional status of children treated at a tertiary university hospital. Our focus is to verify children with some types of pathology, including the following: osteogenesis imperfecta, cardiopathy, cystic fibrosis, respiratory disease and tumor who were hospitalized at the University Hospital Medical School in Ribeirão Preto/SP, Brazil. Basically, the study is based on knowledge of the prevalence of child malnutrition, where some information about the hospitalized children, such as sex, age, weight, height, gestational age, body mass index (BMI), bioelectrical impedance, among others. The BMI can be classified by cutoff points to age (BMI/A) that are determined according to the Z-score of the World Health Organization (WHO) table of parameters, where +2 means overweight and -2 means undernutrition. Through some measures, the phase angle based on resistance and reactance values was also obtained. The phase angle (PA) in children is a useful tool for evaluating nutritional assessment of body cell mass in stable pediatric patients and an important alternative method for predicting malnutrition (low PA value) (PILEGGI *et al.*, 2016)). More information about these measures can be found in Pileggi *et al.* (2016).

The motivation of this study has been to compare the prevalence of malnutrition in the pediatric wards based on the average of phase angle of patients with some specific diseases (University hospital). The evaluation of the children was between February 2008 and February 2009. We were able to use the data from 93 pediatric ward patients. However, our sample size is $n = 12$ months of search.

The predictors analyzed in the model were the number of patients: with Z-score BMI (divided into 3 classes: underweight, normal weight and overweight, has a compositional structure) (Z_1, Z_2, Z_3), where these covariates became ilr coordinates (X_1 and X_2); birth normal weight (X_3); gestational age less than 37 weeks (X_4); male (X_5); with age less than 5 years (X_6); cesarean birth (X_7). The PA was used for the response of the model. Figures 10 and 11 present the solution path by the SSL model (non-adaptative choice (separable), fixed θ ; non-adaptative oracle choice (separable); adaptative choice, $\theta \sim Binomial(1, p)$ (non-separable)), Lasso and elastic net penalties for modeling healthy patients and patients with some disease, respectively. For the group of healthy children, the three settings of SSL (Figure 10 A, B and C) obtained

similar results, that is, the predictor number of patients who were born with normal weight (X_3) and number of patients who had gestational age less than 37 weeks (X_4) were significant. This result showed that the healthy children who did not have malnutrition are those who have a normal weight at birth and gestational age less than 37 weeks (coefficients with positive values). On the other hand, the Lasso and elastic net methods included X_1 and X_5 covariates in the model, where it presented optimal $\lambda = 0.824$ by a 10-fold cross-validation (Figure 10D), that is, the healthy children tend to have malnutrition when there are underweight at birth and are female (negative estimative of X_5).

For the group of children with some diseases, the solution paths for $\theta = 0.5$ when it is fixed (Figure 11A) and θ is set to the separable penalty choice 2/7 (Figure 11B), the predictor is the number of patients who had cesarean births, which was included in the model (coefficient with negative value). However, the SSL with the non-separable (adaptative) choice (Figure 11C) did not include no coefficient in the model. The Lasso and elastic net methods included the number of patients who had cesarean births (X_7) in the model, assuming the optimal $\lambda = 1.017$ calculated by the 10-fold cross-validation (Figure 11D). If we observe the Figures (Figure 11A, 11B, 11D and 11E), the children with diseases who were not born by cesarean have prevalence to a malnutrition based on the PA measure. This is an important fact to analyze the remarkable question of malnutrition between healthy children and who have some type of pathology. The same result can be seen in Figures 11D and 11E.

The models for each applied method (healthy children group) is given by

1. SSL separable (Figure 10A):

$$y = 10.040 * birthnormalweight + 6.479 * gestationalagelessthan37weeks$$

2. SSL separable (Figure 10B):

$$y = 10.940 * birthnormalweight + 6.479 * gestationalagelessthan37weeks$$

3. SSL non-separable (Figure 10C):

$$y = 10.940 * birthnormalweight + 6.479 * gestationalagelessthan37weeks$$

4. Lasso:

$$y = 0.589 * ZscoreBMIunderweight - 0.197 * male$$

5. Elastic net:

$$y = 0.275 * ZscoreBMIunderweight - 1.084 * male$$

The models for each applied method (children group with some pathologies) is given by

1. SSL separable (Figure 10A):

$$y = -6.725 * cesareanbirth$$

2. SSL separable (Figure 10B):

$$y = -6.725 * cesareanbirth$$

3. Lasso:

$$y = -1.190 * cesareanbirth$$

4. Elastic net:

$$y = -2.158 * cesareanbirth$$

4.4 Discussion

In this chapter, we applied a new methodology for regression model with compositional covariates for child malnutrition data. The SSL, Lasso and elastic net penalties were applied in the model with constraint covariates assuming dependence among them. Such a modelling approach had a motivation based on a real data set that focused on the nutrition status of children with some pathologies and a control group of healthy children by some measures defined by the WHO.

The main key is to apply such penalties in the regression model with compositional constraints when $n \ll p$. This methodology yields good solutions by the fact of removing irrelevant predictors and keeping the larger coefficients, thus obtaining accuracy of coefficient estimation. We compared the SSL method with Lasso and elastic net, which is similar when we do not have the slab component, and the performance of Lasso and elastic net were different from the SSL method for the healthy children, showing that this approach moves more coefficients toward zero, even if we adopt a strong penalty. On the other hand, for the group of children with some pathologies, the SSL methods (except SSL non-separable), Lasso and elastic net incorporated the same significant covariate (X_7) in the model, that is, children with some pathologies tend to have malnutrition when they were not born by cesarean (negative value of estimative).

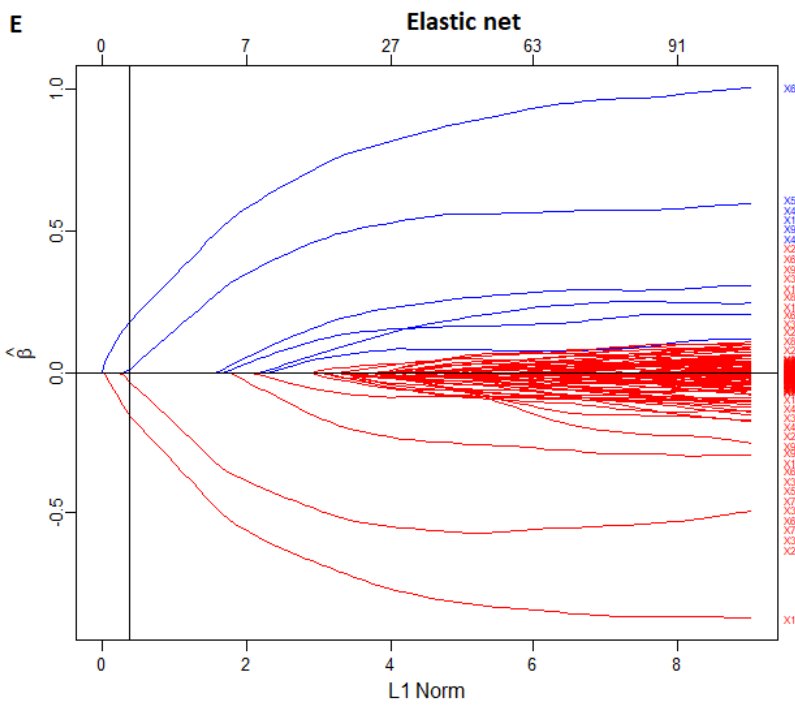
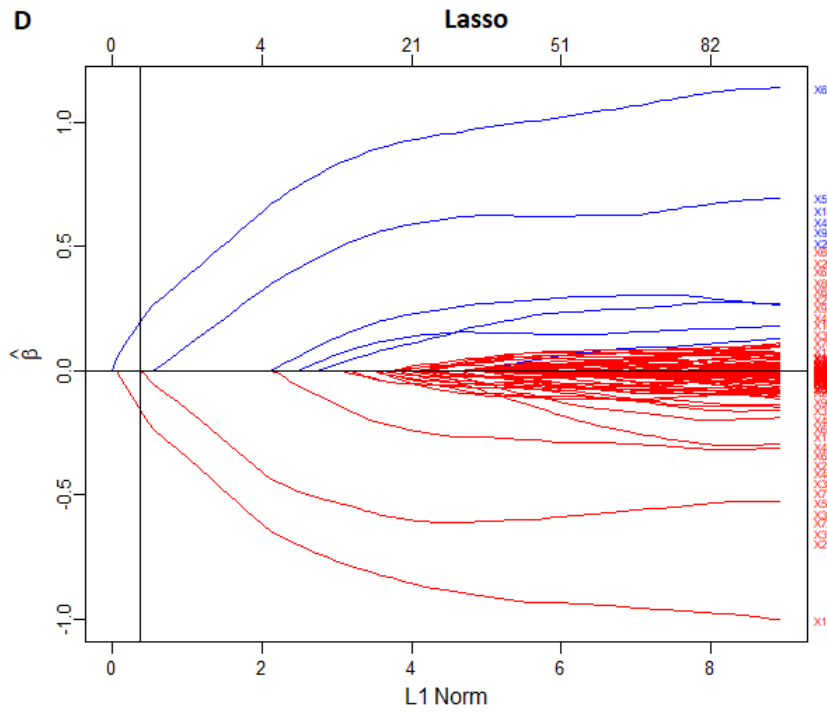


Figure 9 – The Lasso solution path (D) and elastic net path (E).

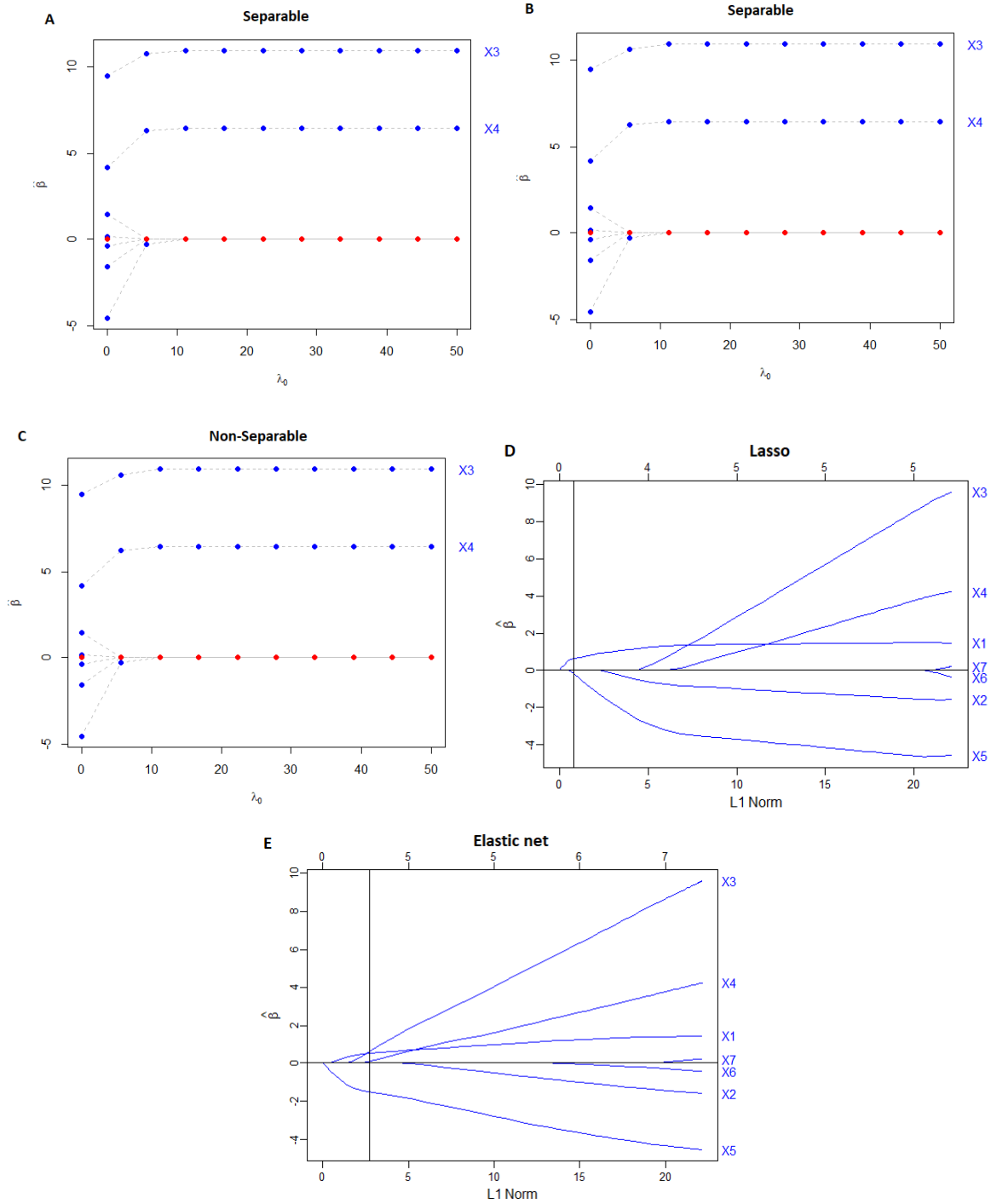


Figure 10 – The solution path SSL (A, B, C), lasso (D) and elastic net (E) for healthy patients. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model lasso and elastic net (cross-validation), respectively.

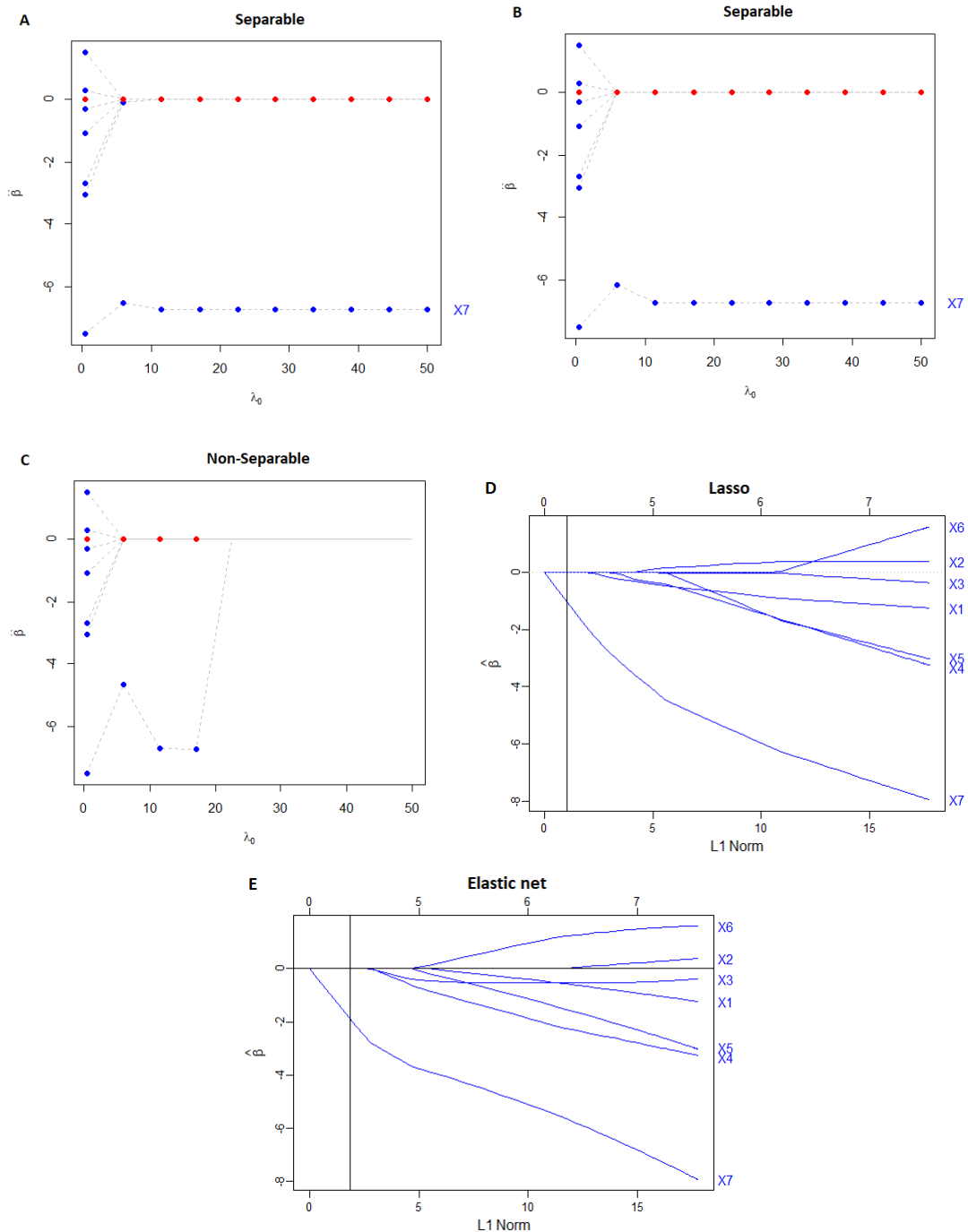


Figure 11 – The solution path SSL (A, B, C), lasso (D) and elastic net (E) for patients with pathologies. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model lasso and elastic net (cross-validation), respectively.

PENALIZED REGRESSION MODEL FOR COMPOSITIONAL RESPONSE VARIABLES AND COVARIATES

In this section, we present the penalized regression model with compositional response and covariates. The regression model based on the methodology of compositional data is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (5.1)$$

where \mathbf{y} is a vector ($D \times 1$) of compositional response variables, \mathbf{X} is a matrix ($D \times D$) of D compositional covariates, where D is the number of components, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_D)^\top$ is a vector ($D \times 1$) unknown parameters and $\boldsymbol{\varepsilon}$ is the noise vector with distribution $N_D(\mathbf{0}, I_p)$, with a known variance $\sigma^2 = 1$. The intercept of the model is not included, equal to models (3.1) and (4.1).

Based on the principle of working in coordinates, we can rewrite the model (5.1) as

$$\begin{aligned} \text{ilr}(\mathbf{y}) &= \langle \boldsymbol{\beta}, \mathbf{X} \rangle_A + \boldsymbol{\varepsilon}, \\ &= \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X) + \boldsymbol{\varepsilon}, \end{aligned} \quad (5.2)$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}_{D-1}, \Sigma_{\text{ilr}})$ and with a vector of parameters $\boldsymbol{\beta} = (\beta_k)$ that afterwards might be mapped back to a composition through the inverse ilr transformation.

Considering the same scheme of Chapter 3 and Chapter 4, we have the following estimators for $\boldsymbol{\beta}$ of the regression models with compositional responses and covariates focused on regularization methods presented in Section 2. We considered the Lasso, elastic net, SSL with separable and non-separable penalty approaches, respectively, for model (5.2) as follows.

1. Lasso:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left(\|\text{ilr}(\mathbf{y}) - \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X)\|_2^2/n + \lambda \|\boldsymbol{\beta}\|_1 \right), \quad (5.3)$$

where $\|\text{ilr}(\mathbf{y}) - \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X_i)\|_2^2 = \sum_{i=1}^n (\text{ilr}(\mathbf{y}) - \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X))^2$ and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^D |\beta_j|$.

2. Elastic Net:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\frac{1}{n} \|\text{ilr}(\mathbf{y}) - \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X)\|_2^2 + \lambda \left[\frac{1}{2} (1 - \alpha) \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right] \right). \quad (5.4)$$

3. SSL with separable penalty (known variance):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{D-1}}{\operatorname{argmax}} \left\{ -\frac{1}{2} \|\text{ilr}(\mathbf{y}) - \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X)\|^2 + \left[-\lambda_1 \|\boldsymbol{\beta}\| + \sum_{j=1}^D \log \left(\frac{p_{\boldsymbol{\theta}}^*(0)}{p_{\boldsymbol{\theta}}^*(\beta_j)} \right) \right] \right\}. \quad (5.5)$$

4. SSL with non-separable penalty (unknown variance):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{D-1}}{\operatorname{argmax}} \left\{ -\frac{1}{2} \|\text{ilr}(\mathbf{y}) - \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X)\|^2 + \left(-\lambda_1 \|\boldsymbol{\beta}\| + \log \left[\frac{\int \frac{\theta^D}{\prod_{j=1}^D p_{\boldsymbol{\theta}}^*(\beta_j)} d\pi(\boldsymbol{\theta})}{\int \frac{\theta^D}{\prod_{j=1}^D p_{\boldsymbol{\theta}}^*(0)} d\pi(\boldsymbol{\theta})} \right] \right) \right\}. \quad (5.6)$$

For the estimation of the $\boldsymbol{\beta}$'s, we implemented the estimators (5.3) and (5.4) through by R package `glmnet` (FRIEDMAN; HASTIE; TIBSHIRANI, 2010). The estimators (5.5) and (5.6) were obtained through by R package `SSLASSO` (MORAN; ROCKOVÁ; GEORGE, 2018).

5.1 Simulation Analysis

We provided the simulation studies to investigate the efficacy of the penalized methods for a regression model with compositional response variable and covariates. We replicated the simulation 1000 times and the results were summarized based on these replicates (Tables 6 and 7). We generated compositional data matrix \mathbf{X} from a logistic normal distribution with mean $\mathbf{0}_D$ and covariance matrix $\Sigma = (\rho^{|i-j|})$ with $\rho = 0.2$ and $\rho = 0.5$, for $i, j = 1, \dots, D$. Moreover, the compositional response variable is generated according to model (5.2), from a logistic normal distribution with mean $\mathbf{0}_D$ and $\Sigma = (\rho^{|i-j|})$ with $\rho = 0.2$ and $\rho = 0.5$. We assume $D = 3$, that is, we have 3 components (y_1, y_2, y_3) of a composition. The fixed values for the parameters $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*)$ were $\beta_1^* = (-2, -1.5, -1, 0, 1, 1.5, 2, 0, \dots, 0)^\top$ and $\beta_2^* = (2, -1, -2.5, 0, 1, -1, 0.5, 0, \dots, 0)^\top$ to $q = 6$ random directions (non-zero coefficients).

We assumed three scenarios with different number of sample sizes and covariates: $(n, p) = (50, 30), (100, 200)$ and $(100, 1000)$. The adopted performance measures for our comparisons were the MSE, FP, FN, HAM measure. Tables 6 and 7 report the averages of these performance measures for the five regularization methods adopted. As can be seen in Tables 6 and 7, similar results are presented for all the settings. It is worth highlighting that the lasso and elastic net estimator perform slightly better than SSL penalties in high dimensions according to the HAM measure.

Table 6 – Averages of some performance measures for penalized methods with compositional dependent variables and covariates ($\text{ilr}(y_1)$).

(n, p)	Method	MSE	FP	FN	HAM
$\rho=0.2$					
SSL (λ_1, θ)					
(50, 30)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.5007	0.1820	5.9610	6.1820
	SSL (1, 6/30) with $\sigma = 1$ fixed	0.5001	0.0060	5.9990	6.0060
	SSL (1, 6/30) with unknown σ	0.5001	0.0070	5.9990	6.0070
	Lasso	0.5039	1.5170	5.3930	7.5170
	Elastic Net	0.5028	1.9220	5.2400	7.9220
(100, 200)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0733	1.7710	5.9540	7.7710
	SSL (1, 6/200) with $\sigma = 1$ fixed	0.0729	0.0010	6.0000	6.0010
	SSL (1, 6/200) with unknown σ	0.0729	0.0080	6.0000	6.0080
	Lasso	0.1465	1.1470	5.8790	7.1470
	Elastic Net	0.0730	3.8840	5.7570	9.8840
(100, 1000)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.1465	9.0870	5.4270	15.0870
	SSL (1, 6/1000) with $\sigma = 1$ fixed	0.1465	7.0500	5.5620	13.0500
	SSL (1, 6/1000) with unknown σ	0.1465	15.3680	5.0480	21.3680
	Lasso	0.1465	0.7290	5.9390	6.7290
	Elastic Net	0.1465	0.7270	5.9390	6.7270
$\rho=0.5$					
SSL (λ_1, θ)					
(50, 30)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.5017	0.4320	5.9050	6.4320
	SSL (1, 6/30) with $\sigma = 1$ fixed	0.5001	0.0200	5.9970	6.0200
	SSL (1, 6/30) with unknown σ	0.5000	0.0040	5.9980	6.0040
	Lasso	0.4675	1.5170	5.3590	7.5170
	Elastic Net	0.5027	1.6720	5.4050	7.6720
(100, 200)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0740	3.9720	5.8860	9.9720
	SSL (1, 6/200) with $\sigma = 1$ fixed	0.0729	0.0010	6.0000	6.0010
	SSL (1, 6/200) with unknown σ	0.0729	0.0050	6.0000	6.0050
	Lasso	0.0730	2.2810	5.8380	8.2810
	Elastic Net	0.0730	3.2900	5.7940	9.2900
(100, 1000)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.1466	12.7020	5.1950	18.7020
	SSL (1, 6/1000) with $\sigma = 1$ fixed	0.1465	10.3150	5.3290	16.3150
	SSL (1, 6/1000) with unknown σ	0.1465	11.0040	5.3050	17.0040
	Lasso	0.0145	2.8290	5.9610	8.8290
	Elastic Net	0.1465	0.6470	5.9470	6.6470

5.2 Toy example

A way of illustrating the proposed model (5.2) to compare the SSL, lasso and elastic net penalties, we considered the following example. For $n = 100$ individuals with $D = 1000$ compositional covariates, we considered one generated sample of the simulation study presented in the last section.

We compared the SSL with fixed variance $\sigma^2 = 1$ with three settings: (i) separable

Table 7 – Averages of some performance measures for penalized methods with compositional dependent variables and covariates ($\text{ilr}(y_2)$).

(n, p)	Method	MSE	FP	FN	HAM
$\rho=0.2$					
SSL (λ_1, θ)					
(50, 30)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.4747	1.5680	5.5900	7.5680
	SSL (1, 6/30) with $\sigma = 1$ fixed	0.4671	0.1890	5.9460	6.1890
	SSL (1, 6/30) with unknown σ	0.4656	0.0100	5.9970	6.0100
	Lasso	0.4675	1.5170	5.3590	7.5170
	Elastic Net	0.4673	1.8500	5.3250	7.8500
(100, 200)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0725	12.1800	5.6400	18.1800
	SSL (1, 6/200) with $\sigma = 1$ fixed	0.0680	0.1500	5.9960	6.1500
	SSL (1, 6/200) with unknown σ	0.0678	0.0040	6.0000	6.0040
	Lasso	0.1364	1.0600	5.8780	7.0600
	Elastic Net	0.0679	3.5070	5.8180	9.5070
(100, 1000)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.1366	21.7880	4.6050	27.7880
	SSL (1, 6/1000) with $\sigma = 1$ fixed	0.1364	18.9530	4.7720	24.9530
	SSL (1, 6/1000) with unknown σ	0.1364	4.4560	5.6820	10.4560
	Lasso	0.1364	0.5970	5.9460	6.5970
	Elastic Net	0.1364	0.5920	5.9460	6.5920
$\rho=0.5$					
SSL (λ_1, θ)					
(50, 30)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.4673	0.3940	5.8750	6.3940
	SSL (1, 6/30) with $\sigma = 1$ fixed	0.4657	0.0150	5.9960	6.0150
	SSL (1, 6/30) with unknown σ	0.4655	0.0030	5.9990	6.0030
	Lasso	0.4692	1.8110	5.2710	7.8110
	Elastic Net	0.4683	2.2020	5.1620	8.2020
(100, 200)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.0690	4.0420	5.8800	10.0420
	SSL (1, 6/200) with $\sigma = 1$ fixed	0.0679	0.0050	5.9980	6.0050
	SSL (1, 6/200) with unknown σ	0.0678	0.0080	5.9990	6.0080
	Lasso	0.0679	2.8000	5.7780	8.8000
	Elastic Net	0.0679	4.0810	5.7780	10.0810
(100, 1000)	SSL (1, 0.8) with $\sigma = 1$ fixed	0.1364	12.6170	5.1660	18.6170
	SSL (1, 6/1000) with $\sigma = 1$ fixed	0.1364	10.2980	5.3290	16.2980
	SSL (1, 6/1000) with unknown σ	0.1364	11.0540	5.2670	17.0540
	Lasso	0.0135	4.1300	5.9380	10.1300
	Elastic Net	0.0135	0.6440	5.9360	6.6440

choice $\theta = 0.8$, to verify the over-estimating of the true non-zero fraction 6/1000, (ii) separable oracle choice $\theta = 6/1000$ and (iii) non-separable (adaptative) choice $\theta \sim \text{Binomial}(1, D)$. The slab parameter was set to $\lambda_1 = 0.1$ and we used a ladder $\lambda_0 \in I = \{1, 2, \dots, 50\}$ for the spike parameter. In addition, we applied the generated data to the Lasso and elastic net penalties implemented in the R package `glmnet` (FRIEDMAN; HASTIE; TIBSHIRANI, 2010). For this approach, an optimal value of λ was selected by 10-fold cross-validation. According to Figures 12, 13, 14 and 15, we can see the solution paths for the five settings when we have the structure of

compositional response variables and covariates together. Each line represents a single regression coefficient and the horizontal dotted lines corresponds to the levels of true coefficients. The true coefficients are in blue and zero coefficients are in red. We can observe that when θ is too large, there are more false positives than false negatives (Figures 12A, 12B and 12C; 14A, 14B and 14C). In comparison with the Figures 13D, 13E, 15D and 15E, the lasso and elastic net presented more false negatives, that is, the model includes unimportant variables with shrunk coefficients.

5.3 Real data application - ICMS dataset

The description of the applied dataset is in Chapter 3. The focus on this Chapter is the restriction in the regression model with compositional response and covariates.

In this case, we considered the same three economic sectors: industry (y_1), commerce (y_2) and administered prices (y_3), defined as compositional data. Besides the covariates mentioned in Chapter 3, we add the lagged compositional covariate in the period of 12 months of the proportion of ICMC in the industry (X_6), commerce (X_7) and administered prices (X_8) and 6 months of the proportion of ICMC in the industry (X_9), commerce (X_{10}) and administered prices (X_{11}).

Figures 16 and 17 present the solution path by the SSL (non-adaptative choice (separable), fixed θ ; non-adaptative oracle choice (separable); adaptative choice, $\theta \sim \text{Binomial}(1, p)$ (non-separable)), lasso and elastic net methods for modeling the ICMS disaggregated in 3 parts: industry, commerce and administered prices considering compositional covariates (X_6 to X_{11}). Thereby, the results showed the same performance for $\text{ilr}(y_1)$ and $\text{ilr}(y_2)$ when the SSL with separable penalties (Figures 16A, 16B, 17A and 17B), that is, these methods did not select any significant covariate for the model. On the other hand, SSL non-separable presented three significant covariates ($\text{ilr}(X_6)$, $\text{ilr}(X_7)$ and $\text{ilr}(X_8)$). The optimal λ calculated by the 10-fold cross-validation were 0.0043 ($\text{ilr}(y_1)$) and 0.0020 ($\text{ilr}(y_2)$) for the lasso method and 0.0069 ($\text{ilr}(y_1)$) and 0.0054 ($\text{ilr}(y_2)$) for the elastic net method (vertical line in Figures 16D, 16E, 17D and 17E).

The models for each applied method is given by

1. SSL non-separable:

$$\begin{aligned}
y_1 &= 0.004 * \text{MonthlyIndustrialSurvey} + 0.001 * \text{MonthlyTradeSurvey} \\
&\quad - 0.002 * \text{IndexEconomicActivity} - 0.001 * \text{IGP} - \text{DI/FGV} + 0.080 * \text{ICMSindustry12months} \\
&\quad + 0.062 * \text{ICMScommerce12months} + 0.116 * \text{ICMSadm12months} \\
&\quad - 0.007 * \text{ICMScommerce12months} \\
y_2 &= -0.001 * \text{MonthlyIndustrialSurvey} + 0.005 * \text{MonthlyTradeSurvey} \\
&\quad + 0.001 * \text{Monthlyenergyconsumption} + 0.001 * \text{IGP} - \text{DI/FGV} \\
&\quad - 0.075 * \text{ICMSindustry12months} + 0.0357 * \text{ICMSindustry6months}
\end{aligned}$$

2. Lasso:

$$\begin{aligned}
y_1 &= 0.002 * \text{MonthlyIndustrialSurvey} + 0.002 * \text{MonthlyTradeSurvey} \\
&\quad + 0.003 * \text{Monthlyenergyconsumption} - 0.001 * \text{IndexEconomicActivity} - 0.002 * \text{IGP} - \text{DI/FGV} \\
y_2 &= 0.001 * \text{MonthlyIndustrialSurvey} + 0.003 * \text{MonthlyTradeSurvey} \\
&\quad + 0.004 * \text{Monthlyenergyconsumption} - 0.003 * \text{IndexEconomicActivity} + 0.003 * \text{IGP} - \text{DI/FGV}
\end{aligned}$$

3. Elastic net:

$$\begin{aligned}
y_1 &= 0.002 * \text{MonthlyIndustrialSurvey} + 0.001 * \text{MonthlyTradeSurvey} \\
&\quad + 0.003 * \text{Monthlyenergyconsumption} - 0.001 * \text{IndexEconomicActivity} - 0.002 * \text{IGP} - \text{DI/FGV} \\
y_2 &= 0.001 * \text{MonthlyIndustrialSurvey} + 0.004 * \text{MonthlyTradeSurvey} \\
&\quad + 0.004 * \text{Monthlyenergyconsumption} - 0.003 * \text{IndexEconomicActivity} + 0.003 * \text{IGP} - \text{DI/FGV}
\end{aligned}$$

5.4 Discussion

In this chapter, we presented a compositional regression model with restriction in the response variables and covariates under five regularization methods presented in Chapter 2. We applied the ilr coordinates on the response variables and covariates simultaneously to remove the dependence among the components.

A simulation study for the proposed model (5.2) showed that the model with lasso and elastic net estimators perform better in terms of estimation if comparable to the other penalized methods in high-dimensions.

In order to illustrate the methodology, a toy example was presented. When the lasso and elastic net estimators are applied, there are many more false negatives if compared with SSL estimators. Clearly, for the $\text{ilr}(y_2)$, the SSL estimators obtained a performance better than lasso and elastic net. Therefore, SSL non-separable (Figure 14C) has superior performance compared with the other SSL estimators (separable).

In the case of application, the real data set involves the ICMC tax as in Chapter 3. The SSL non-separable showed a better performance in relation to the other SSL penalties. The lasso and elastic net estimators presented similar results, with only a little difference between the optimal λ . Based on these results, the SSL non-separable method considered the lagged covariates administered prices $\text{ilr}(X_6)$, $\text{ilr}(X_7)$ and $\text{ilr}(X_8)$ significant, that is, the proportion of ICMS in the industry, commerce and administered prices in the period of 12 months are relevant to explain the response variable $\text{ilr}(y_1)$ (proportion of the ICMS in the industry). On the other hand, the lasso method considered only exogenous covariates significant, which are Monthly Industrial Survey, Monthly Trade Survey and IGP-DI/FGV General Price Index and for the elastic net method, besides these covariates mentioned above, including also the covariate Monthly energy consumption in Sao Paulo State. We observe that these results were similar with obtained in Chapter 3.

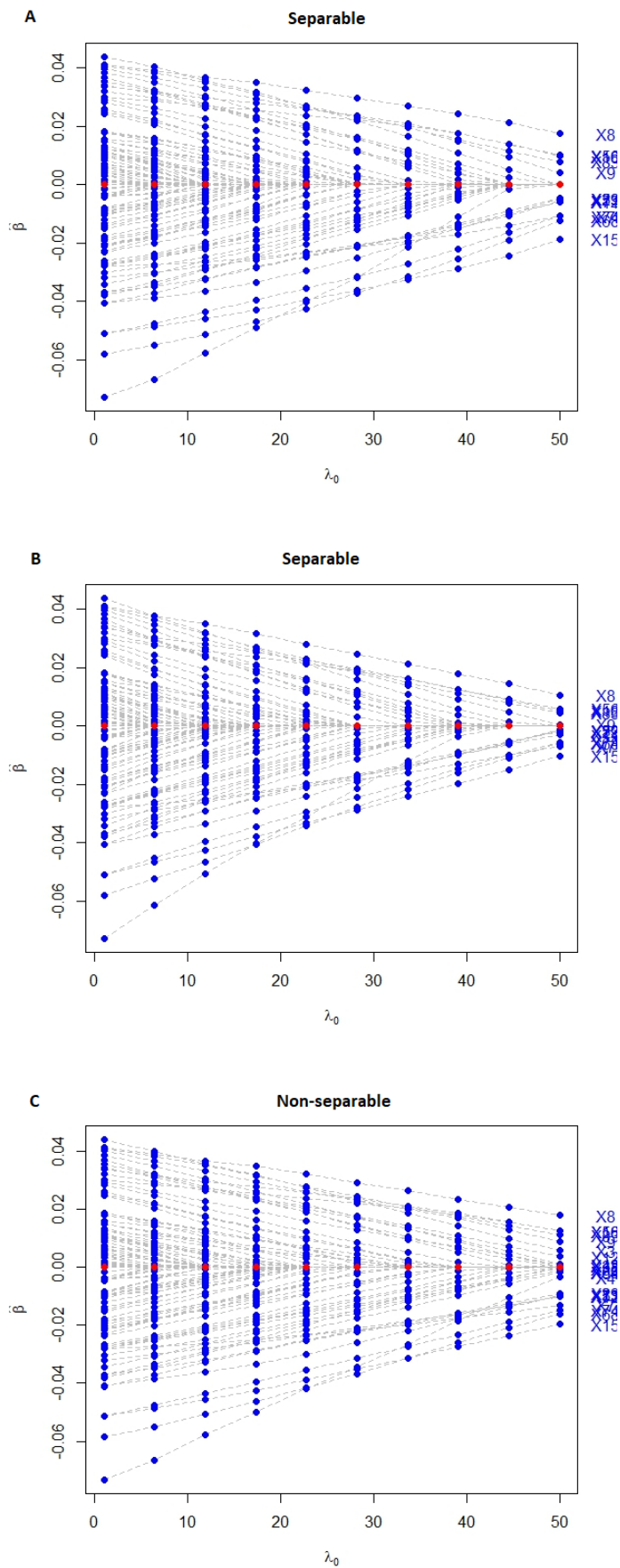
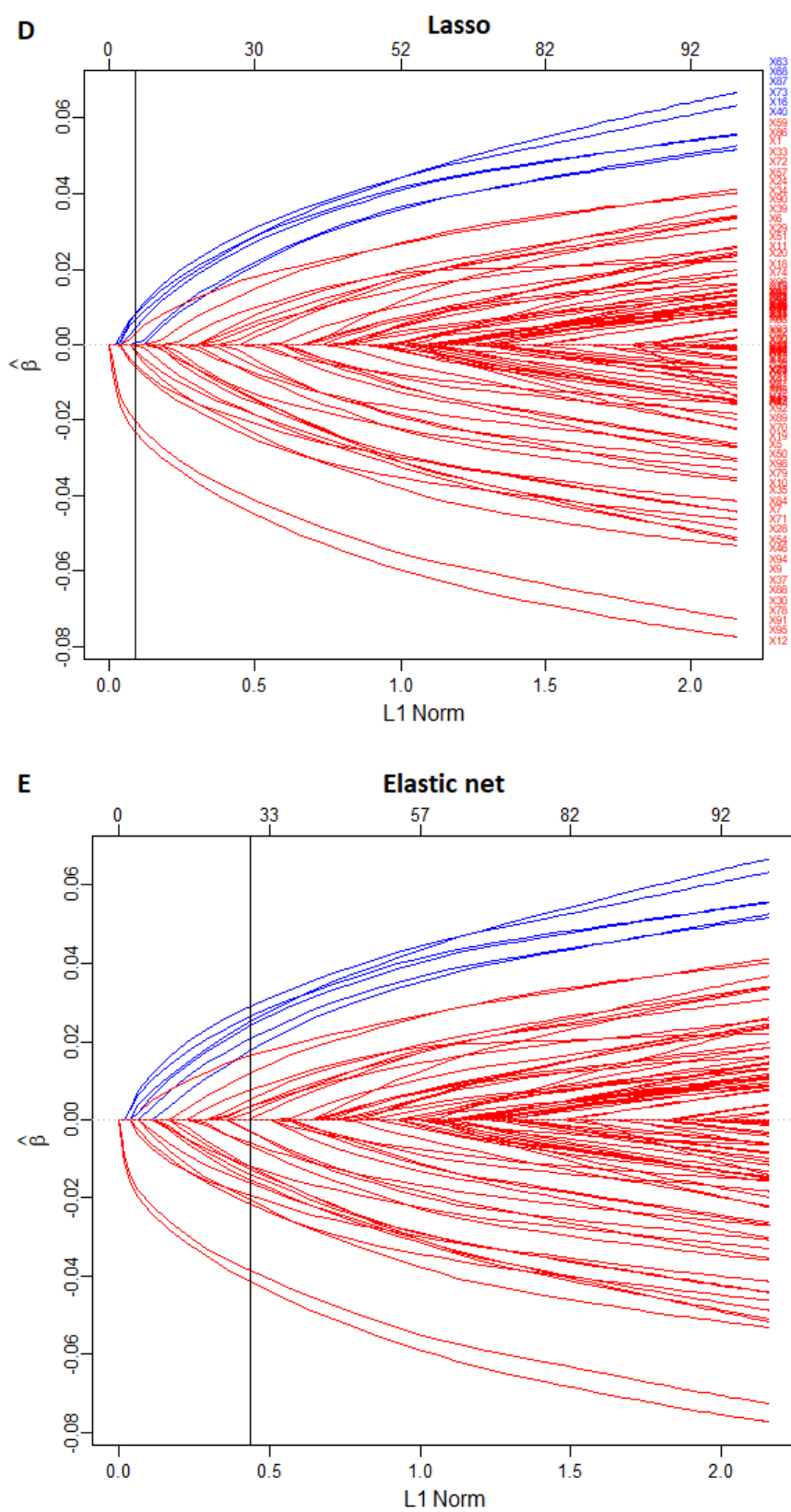
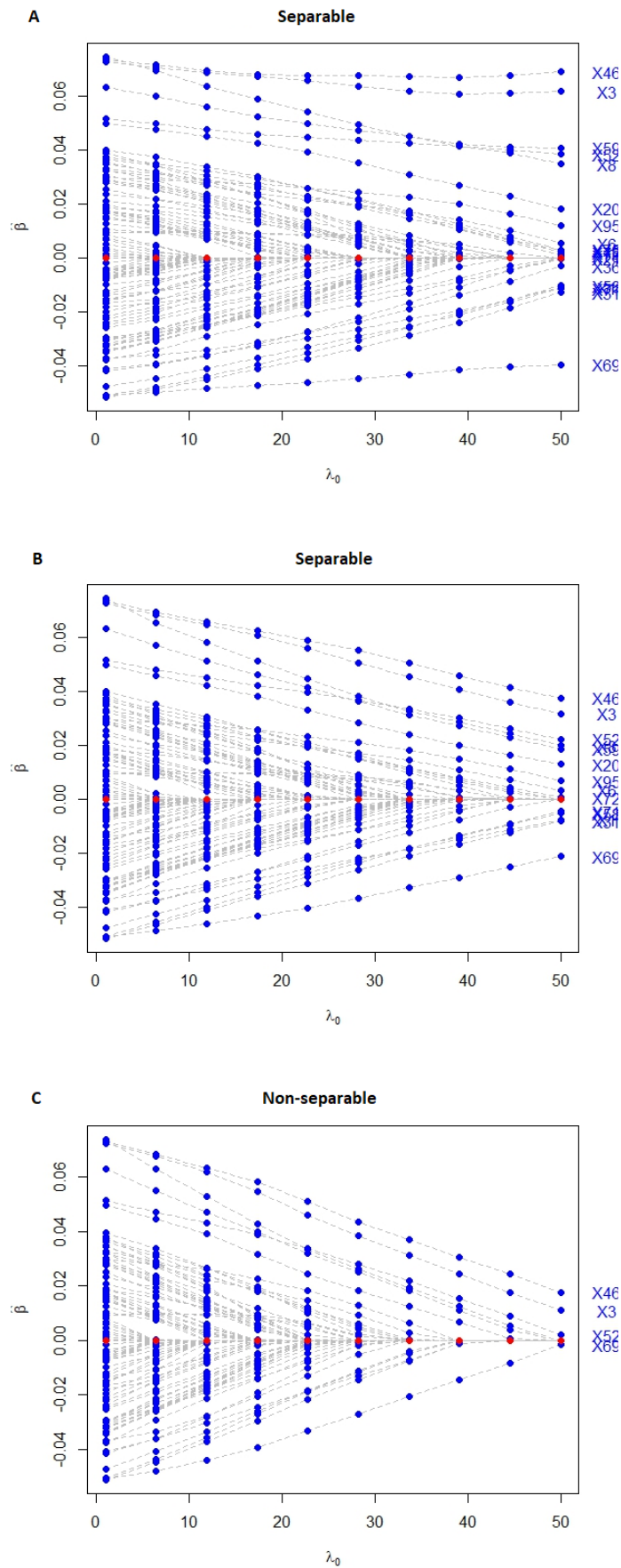


Figure 12 – The SSL solution paths (A, B, C) (for $\text{ilr}(y_1)$).

Figure 13 – The lasso solution path (D) and elastic net path (E) (for $\text{ilr}(y_1)$).

Figure 14 – The SSL solution paths (A, B, C) (for $\text{ilr}(y_2)$).

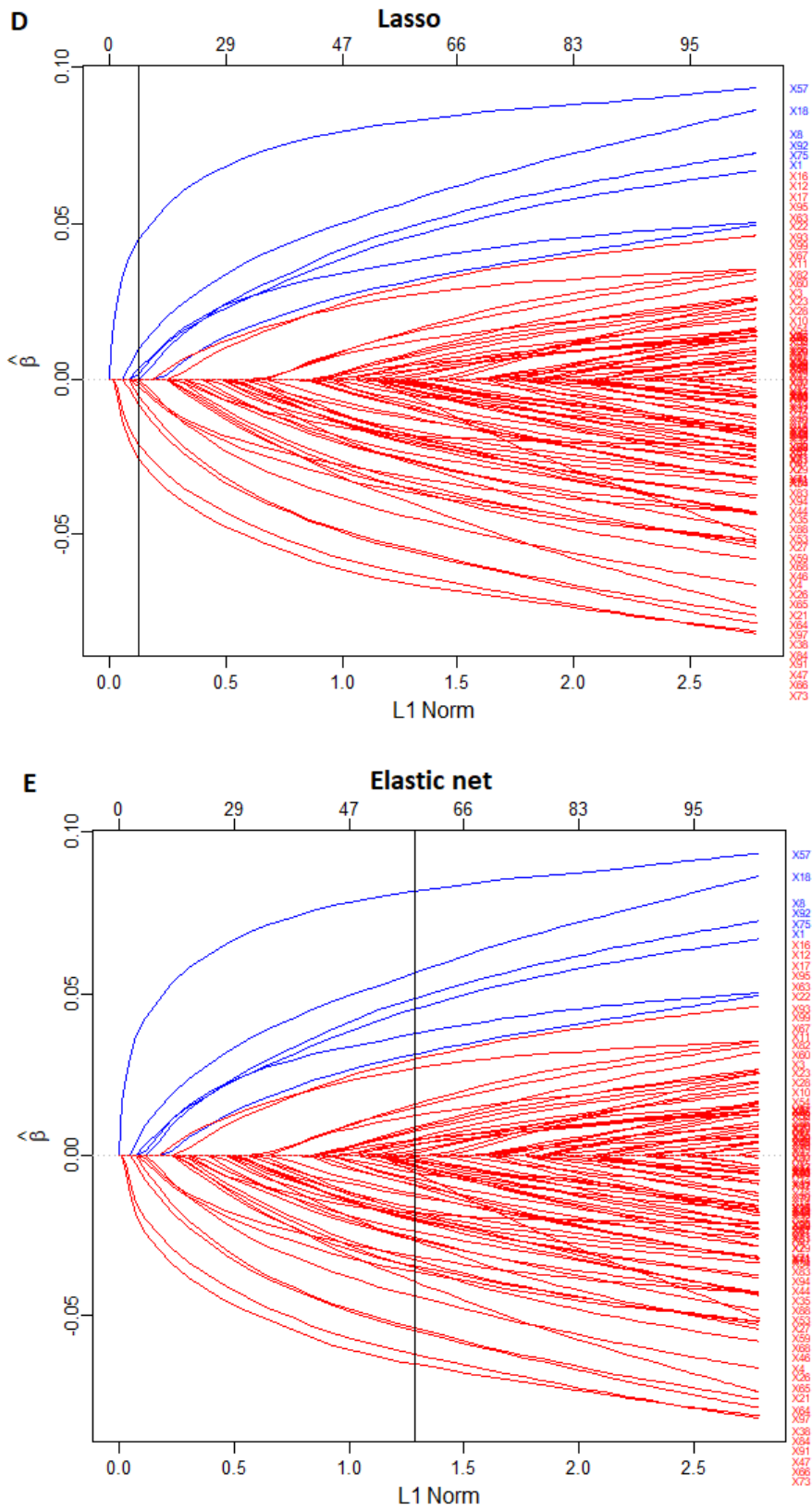


Figure 15 – The lasso solution path (D) and elastic net path (E) (for $\text{ilr}(y_2)$).

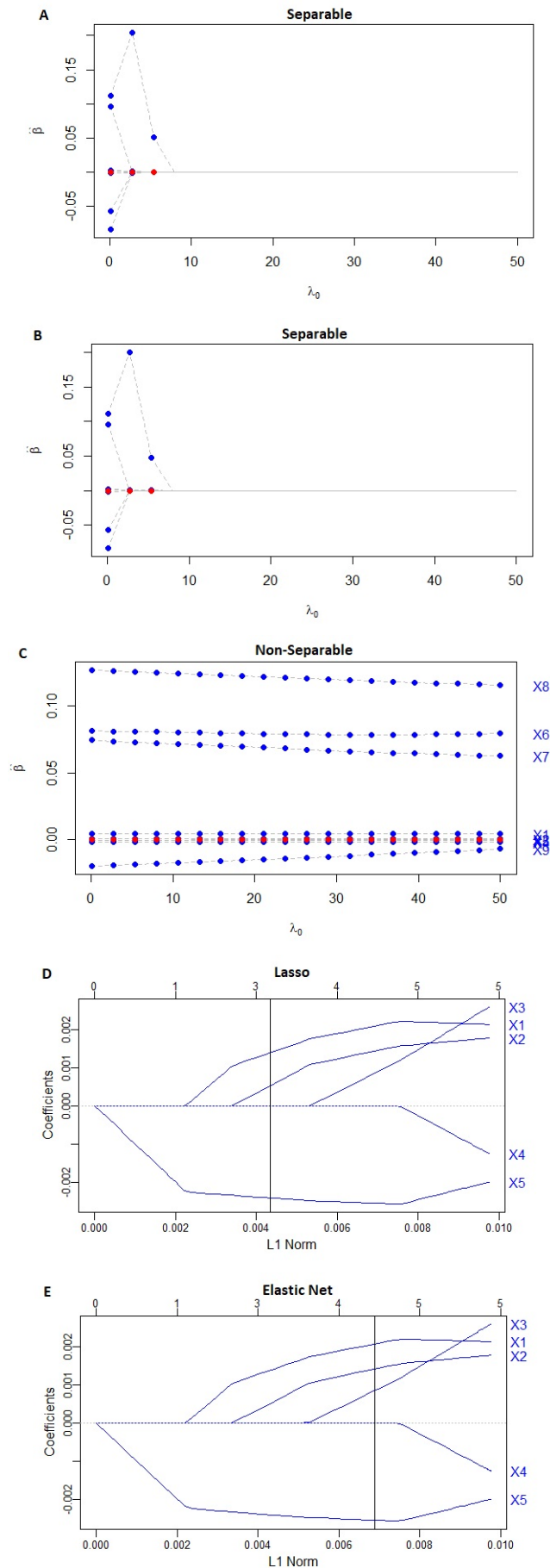


Figure 16 – The solution path SSL (A, B, C), lasso (D) and elastic net (E) for $\text{ilr}(y_1)$. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model lasso and elastic net (cross-validation), respectively.

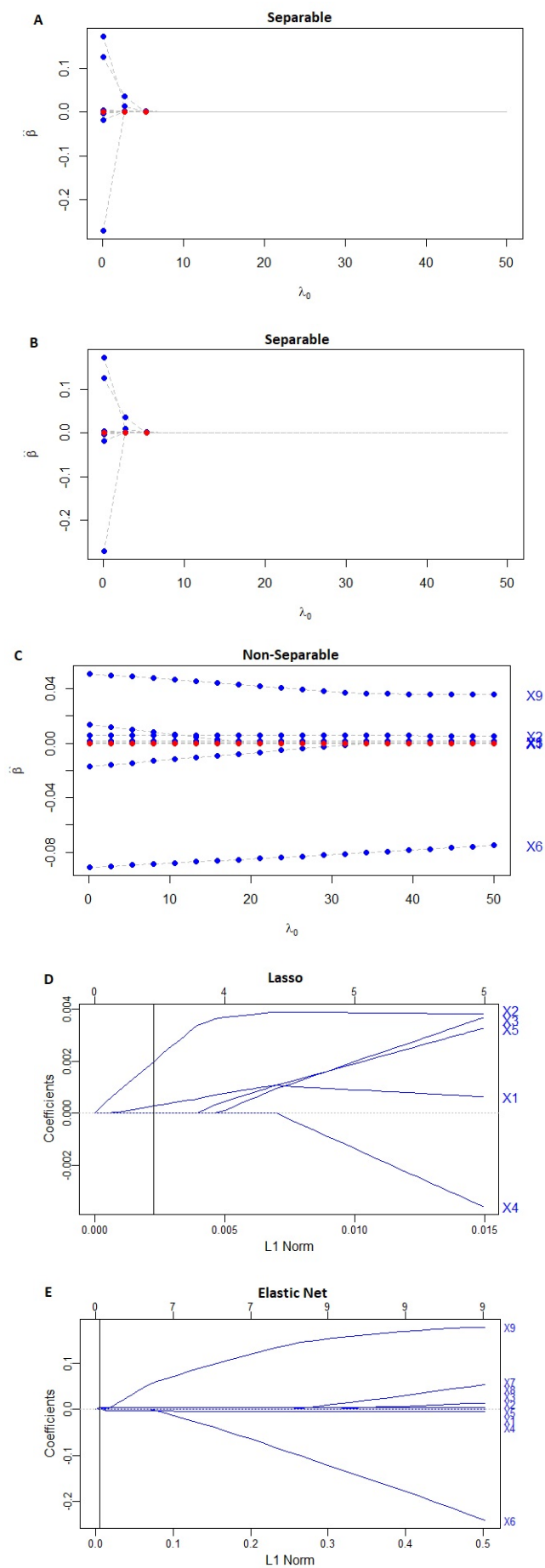


Figure 17 – The solution path SSL (A, B, C), lasso (D) and elastic net (E) for $\text{ilr}(y_2)$. The colored points on the solution path represent the estimated values of the coefficients. The vertical line (D) and (E) corresponds to the optimal model lasso and elastic net (cross-validation), respectively.

CONCLUSION

In this thesis, we considered penalized regression methods, in particular the Lasso (least absolute shrinkage and selection operator), elastic net and Spike-and-Slab lasso when there are compositional restrictions in the response variable, covariates or both of them.

One of the principal constraints of compositional data is the nature of dependence among the components, which cannot be ignored in order to obtain accurate inferences. Thus, the increase in large datasets, whose dimensionality is much larger than sample size, poses new challenges to the current methodology of compositional data.

For the context of regression models, we presented three novel models based on compositional data with an application of different regularization methods. We note that considering the penalized model with compositional response variables, the simulation studies and application in a real data set proved that the SSL estimators (oracle separable and non-separable) performed better than the other regularization methods.

Considering the penalized regression model with compositional covariates, the analysis under this approach in the child malnutrition data is an important contribution to the present study. These data focus on the nutrition status of children with some pathologies with some measures defined by the WHO. For this model, the SSL estimators presented good solutions by the fact of removing irrelevant predictors and keeping the larger coefficients, thus obtaining accuracy of coefficient estimation.

Finally, the last penalized regression model considered restrictions for both the response variables and covariates. The lasso and elastic net estimators perform better if compared with the other penalized methods in high-dimensions in the simulation study.

For the further development, there are several extensions of this current work. In particular, we can consider longitudinal and spatio-temporal longitudinal models with compositional restriction under regularization methods, semiparametric or non-parametric approaches for regression model with log-contrast, where new methods of the regularized estimation could be

developed. Another extension could be to study appropriate models in the presence of zero in compositional data in a high-dimensional setting.

BIBLIOGRAPHY

AITCHISON, J. The statistical analysis of compositional data. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 44, n. 2, p. 139–177, 1982. Citations on pages 23, 24, and 28.

_____. The statistical analysis of compositional data. **Journal of the Royal Statistical Society. Series B (Methodological)**, p. 139–177, 1982. Citation on page 23.

_____. **The statistical analysis of compositional data**. [S.l.]: Chapman and Hall, 1986. Citations on pages 24 and 30.

AITCHISON, J.; EGOZCUE, J. J. Compositional data analysis: Where are we and where should we be heading? **Mathematical Geology**, v. 37, n. 7, p. 829–850, 2005. Citations on pages 24 and 28.

AITCHISON, J.; SHEN, S. M. Logistic-normal distributions: Some properties and uses. **Biometrika**, v. 67, n. 2, p. 261–272, 1980. Citations on pages 23 and 24.

BOOGAART, K. G. van den; TOLOSANA-DELGADO, R. **Analyzing Compositional Data with R**. [S.l.]: Springer Publishing Company, Incorporated, 2013. ISBN 3642368085, 9783642368080. Citations on pages 24 and 32.

BUHLMANN, P.; GEER, S. van de. **Statistics for High-Dimensional Data: Methods, Theory and Applications**. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2011. ISBN 3642201911, 9783642201912. Citation on page 34.

CHEN, J.; ZHANG, X.; LI, S. Multiple linear regression with compositional response and covariates. **Journal of Applied Statistics**, v. 44, n. 12, p. 2270–2285, 2017. Citations on pages 24, 31, and 32.

CHIPMAN, H. Bayesian variable selection with related predictions. **Canadian Journal of Statistics**, v. 24, n. 1, p. 17–36, 1996. Citation on page 26.

EFRON, B.; HASTIE, T.; JOHNSTONE, I.; TIBSHIRANI, R. Least angle regression. **The Annals of Statistics**, v. 32, n. 2, p. 407–499, 2004. Citations on pages 25 and 35.

EGOZCUE, J. J.; PAWLOWSKY-GLAHN, V.; MATEU-FIGUERAS, G.; BARCELÓ-VIDAL, C. Isometric logratio transformations for compositional data analysis. **Mathematical Geology**, v. 35, n. 3, p. 279–300, 2003. Citation on page 24.

FAN, J.; LI, R. Variable selection via nonconcave penalized likelihood and its oracle properties. **Journal of the American Statistical Association**, Taylor & Francis, v. 96, n. 456, p. 1348–1360, 2001. Citation on page 25.

FRIEDMAN, J.; HASTIE, T.; HÖFLING, H.; TIBSHIRANI, R. Pathwise coordinate optimization. *The Institute of Mathematical Statistics*, v. 1, n. 2, p. 302–332, 12 2007. Available: <<https://doi.org/10.1214/07-AOAS131>>. Citation on page 35.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1–22, 2010. Citations on pages [25](#), [38](#), [48](#), [53](#), [62](#), and [64](#).

FU, W. J. Penalized regressions: The bridge versus the lasso. **Journal of Computational and Graphical Statistics**, v. 7, n. 3, p. 397–416, 1998. Citation on page [35](#).

GEORGE, E. I.; MCCULLOCH, R. E. Variable selection via gibbs sampling. **Journal of the American Statistical Association**, v. 88, n. 423, p. 881–889, 1993. Citation on page [26](#).

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference and prediction**. 2. ed. [S.l.]: Springer, 2009. Citations on pages [34](#) and [35](#).

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. **Statistical Learning with Sparsity: The Lasso and Generalizations**. [S.l.]: Chapman & Hall/CRC, 2015. ISBN 1498712169, 9781498712163. Citations on pages [33](#), [34](#), and [35](#).

HIJAZI, R. An em-algorithm based method to deal with rounded zeros in compositional data under dirichlet models. In: PROCEEDINGS OF THE 1TH INTERNATIONAL WORKSHOP ON COMPOSITIONAL DATA ANALYSIS, 4. Giron, 2011. Citation on page [24](#).

HIJAZI, R.; JERNIGAN, R. W. Modelling compositional data using dirichlet regression models. **Journal of Applied Probability & Statistics**, v. 4, n. 1, p. 77–91, 2009. Citation on page [24](#).

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, v. 12, p. 55–67, 1970. Citation on page [35](#).

HRON, K.; FILZMOSER, P.; THOMPSON, K. Linear regression with compositional explanatory variables. **Journal of Applied Statistics**, v. 39, n. 5, p. 1115–1128, 2012. Citations on pages [24](#), [31](#), and [32](#).

ISHWARAN, H.; RAO, J. S. Spike and slab gene selection for multigroup microarray data. **Journal of the American Statistical Association**, v. 100, n. 471, p. 764–780, 2005. Citation on page [26](#).

JOHNSON, R.; WICHERN, D. **Applied multivariate statistical analysis**. [S.l.]: New Jersey: Prentice Hall, 1998. Citation on page [24](#).

LIN, W.; SHI, P.; FENG, R.; LI, H. Variable selection in regression with compositional covariates. **Biometrika**, v. 101, n. 4, p. 1–13, 2014. Citations on pages [25](#) and [33](#).

MARTÍN-FERNÁNDEZ, J.; BARCELÓ-VIDAL, C.; PAWLOWSKY-GLAHN, V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. **Mathematical Geology**, v. 35, n. 3, p. 253–278, 2003. Citation on page [24](#).

MORAN, G. E.; ROCKOVÁ, V.; GEORGE, E. I. On variance estimation for Bayesian variable selection. **ArXiv e-prints**, Jan. 2018. Citations on pages [36](#), [38](#), [48](#), and [62](#).

PAWLOWSKY-GLAHN, V.; BUCCIANTI, A. **Compositional data analysis: Theory and applications**. [S.l.]: John Wiley & Sons, 2011. Citation on page [24](#).

PAWLOWSKY-GLAHN, V.; EGOZCUE, J. Geometric approach to statistical analysis on the simplex. **Stochastic Environmental Research and Risk Assessment**, v. 15, p. 384–398, 2001. Citation on page [29](#).

PAWLOWSKY-GLAHN, V.; EGOZCUE, J. J.; TOLOSANA-DELGADO, R. **Modeling and analysis of compositional data**. [S.l.]: John Wiley & Sons, 2015. Citations on pages 24, 27, 28, and 30.

PILEGGI, V. N.; MONTEIRO, J. P.; MARGUTTI, A. V. B.; JR., J. S. C. Prevalence of child malnutrition at a university hospital using the world health organization criteria and bioelectrical impedance data. **Brazilian Journal of Medical and Biological Research**, v. 49, n. 3, 2016. Citation on page 54.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017. Available: <<https://www.R-project.org/>>. Citation on page 36.

ROCKOVÁ, V.; GEORGE, E. I. Emvs: The em approach to bayesian variable selection. **Journal of the American Statistical Association**, v. 109, n. 506, p. 828–846, 2014. Citation on page 26.

_____. The spike-and-slab lasso. **Journal of the American Statistical Association**, v. 113, n. 521, p. 431–444, 2018. Citations on pages 25, 26, 35, and 36.

SHELTON, J. A.; SHEIKH, A. S.; BORNSCHEIN, J.; STERNE, P.; LUCKE, J. Nonlinear spike-and-slab sparse coding for interpretable image encoding. **PLoS One**, v. 10, p. e0124088, 2015. Citation on page 26.

TANG, Z.; SHEN, Y.; ZHANG, X.; YI, N. The spike-and-slab lasso cox model for survival prediction and associated genes detection. **Bioinformatics**, v. 33, n. 18, p. 2799–2807, 2017. Citation on page 26.

_____. The spike-and-slab lasso generalized linear models for prediction and associated genes detection. **Genetics**, v. 205, n. 1, p. 77–88, 2017. Citation on page 26.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society B**, v. 58, p. 267–288, 1996. Citations on pages 25, 33, and 35.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society, Series B**, v. 67, p. 301–320, 2005. Citations on pages 25 and 35.

Computational Routines for estimation of parameters - *software R*

```

1 #####
2 #      Simulation Chapter 3
3 #####
4 rm(list=ls())
5
6 #generate compositional data
7 rcompos <- function(n,l,p,beta,sigma){
8 y = matrix(NA, nrow=n, ncol=l)
9 Z = matrix(NA, nrow=n, ncol=p)
10 Y = matrix(NA, nrow=n, ncol=l)
11
12 for (j in 1:p){
13   Z[,j] <- rnorm(n,2)
14 }
15
16 mu = Z%*%beta
17 mean = apply(mu,2,mean)
18 y = rcompnorm(n,mean,sigma,type="alr")
19
20 return(list(y,Z))
21 }
22
23 output1.1 = list()
24 output1.2 = list()
25 output1.3 = list()

```

```
26 output2.1 = list ()
27 output2.2 = list ()
28 output2.3 = list ()
29 output3.1 = list ()
30 output3.2 = list ()
31 output3.3 = list ()
32 est1.1 = matrix (0, nrow=p, ncol=S-1)
33 est1.2 = matrix (0, nrow=p, ncol=S-1)
34 est1.3 = matrix (0, nrow=p, ncol=S-1)
35 est2.1 = matrix (0, nrow=p, ncol=S-1)
36 est2.2 = matrix (0, nrow=p, ncol=S-1)
37 est2.3 = matrix (0, nrow=p, ncol=S-1)
38 est3.1 = matrix (0, nrow=p, ncol=S-1)
39 est3.2 = matrix (0, nrow=p, ncol=S-1)
40 est3.3 = matrix (0, nrow=p, ncol=S-1)
41 est.lasso1 = matrix (0, nrow=p, ncol=S-1)
42 est.lasso2 = matrix (0, nrow=p, ncol=S-1)
43 est.lasso3 = matrix (0, nrow=p, ncol=S-1)
44 est.elastic1 = matrix (0, nrow=p, ncol=S-1)
45 est.elastic2 = matrix (0, nrow=p, ncol=S-1)
46 est.elastic3 = matrix (0, nrow=p, ncol=S-1)
47 ham.ss11.1 = c ()
48 ham.ss11.2 = c ()
49 ham.ss11.3 = c ()
50 ham.ss12.1 = c ()
51 ham.ss12.2 = c ()
52 ham.ss12.3 = c ()
53 ham.ss13.1 = c ()
54 ham.ss13.2 = c ()
55 ham.ss13.3 = c ()
56 ham.lasso1 = c ()
57 ham.lasso2 = c ()
58 ham.lasso3 = c ()
59 ham.elastic1 = c ()
60 ham.elastic2 = c ()
61 ham.elastic3 = c ()
62
63 ##### GENERATE COMPOSITIONAL DATA MATRIX ###
64 set.seed(2018)
```

```

65 while(j < S){
66 Y <- rcompos(n,l,p,beta ,sigma)
67 y <- as.matrix(pivotCoord(Y[[1]]))
68 Z <- as.matrix(Y[[2]])
69 #####
70 ##SSL
71 #####
72 lambda1 <- 1
73 lambda0 <- seq(lambda1,50, length=10)
74 L <- length(lambda0)
75
76 # Oracle SSLASSO with known variance (Separable)
77 result1.1 <- SSLASSO(Z, y[,1], penalty = "separable", variance
      = "known",
78           lambda1 = lambda1, lambda0 = lambda0,
79           theta = 0.8)
80
81 # Oracle SSLASSO with known variance (Separable Oracle)
82 result1.2 <- SSLASSO(Z, y[,1], penalty = "separable", variance
      = "known",
83           lambda1 = lambda1, lambda0 = lambda0,
84           theta = 6/p)
85
86 # Oracle SSLASSO with unknown variance (Non-Separable)
87 result1.3 <- SSLASSO(Z, y[,1], penalty = "adaptive", variance =
      "unknown",
88           lambda1 = lambda1, lambda0 = lambda0,
89           theta = 6/p)
90
91 #####
92 ## Lasso
93 #####
94 result1.4 = glmnet(Z,y[,1],family="gaussian",alpha=1,
      standardize=TRUE,intercept=FALSE)
95 cv.lasso.modtotal1.4 = cv.glmnet(Z,y[,1],family="gaussian",
      alpha=1)
96 bestlam.lasso1.4=cv.lasso.modtotal1.4$lambda.1se
97

```

```

98 result1.4.opt = glmnet(Z,y[,1],alpha=1,standardize=TRUE,lambda=
    bestlam.lasso1.4,intercept=FALSE)
99 #print(coef(result1.opt), digit=3)
100 coef.lasso1 = round(matrix(coef(result1.4.opt)[-1,],nrow=p),4)
101
102 #####
103 ## Elastic Net
104 #####
105 a <- seq(0.1, 0.9, 0.05)
106 search <- foreach(i = a, .combine = rbind) %dopar% {
107   cv <- cv.glmnet(Z, y[,1], family = "gaussian", nfold = 10,
    type.measure = "deviance", paralle = TRUE, alpha = i)
108   data.frame(cvm = cv$cvm[cv$lambda == cv$lambda.1se], lambda.1
    se = cv$lambda.1se, alpha = i)
109 }
110 cv1 <- search[search$cvm == min(search$cvm), ]
111
112 result1.5 = glmnet(Z,y[,1],family="gaussian",alpha=cv1$alpha,
    lambda=cv1$lambda.1se,intercept=FALSE,standardize=TRUE)
113
114 coef.elastic1 = round(matrix(coef(result1.5)[-1,],nrow=p),4)

```

```

1 #####
2 #      Simulation Chapter 4
3 #####
4 ##### GENERATE COMPOSITIONAL DATA MATRIX
5 seed=2018
6 set.seed(2018)
7
8 while(j < S){
9   mean <- c(rep(0,p-1)) #means for each component
10  sigma <- matrix(0.2,nrow=p-1,ncol=p-1)
11  diag(sigma) <- 1
12
13  X <- rcompnorm(n,m=mean,s=sigma,type="alr")
14  X.mean = apply(X,2,mean)
15  ind <- order(X.mean,decreasing=T)[1:3]
16  X.new <- (cbind(X[,ind],X[-ind]))
17

```

```
18 #####
19 ## SSL
20 #####
21 ### GENERATE RESPONSE VECTOR Y
22 beta = c(-2, -1.5, -1, 0, 1, 1.5, 2, rep(0, p-8))
23
24 y = Z[,1]*beta[1]+Z[,2]*beta[2]+Z[,3]*beta[3]+Z[,4]*beta[4]+Z
      [,5]*beta[5]+Z[,6]*beta[6]+Z[,7]*beta[7]+rnorm(n)
25
26
27 # Oracle SSLASSO with known variance (Separable)
28 result1 <- SSLASSO(Z, y, penalty = "separable", variance = "
      known",
29                   lambda1 = lambda1, lambda0 = lambda0,
30                   theta = 0.8)
31
32 # Oracle SSLASSO with known variance (Separable Oracle)
33 result2 <- SSLASSO(Z, y, penalty = "separable", variance = "
      known",
34                   lambda1 = lambda1, lambda0 = lambda0,
35                   theta = 6/p)
36
37 # Oracle SSLASSO with unknown variance (Non-Separable)
38 result3 <- SSLASSO(Z, y, penalty = "adaptive", variance = "
      unknown",
39                   lambda1 = lambda1, lambda0 = lambda0,
40                   theta = 6/p)
41
42 output1 [[j]] = result1$beta[,10]
43 output2 [[j]] = result2$beta[,10]
44 output3 [[j]] = result3$beta[,10]
45
46 est1[,j] = as.matrix(output1 [[j]])
47 est2[,j] = as.matrix(output2 [[j]])
48 est3[,j] = as.matrix(output3 [[j]])
49
50 medias1 = apply(est1, 1, mean)
51 medias2 = apply(est2, 1, mean)
52 medias3 = apply(est3, 1, mean)
```

```

53
54 var1 = apply(est1 ,1 , var)
55 var2 = apply(est2 ,1 , var)
56 var3 = apply(est3 ,1 , var)
57
58 sd1 = apply(est1 ,1 , sd)
59 sd2 = apply(est2 ,1 , sd)
60 sd3 = apply(est3 ,1 , sd)
61
62 ##### diagnostics statistics
63 bias1 = medias1 - beta
64 bias2 = medias2 - beta
65 bias3 = medias3 - beta
66
67 mse1 = mean(var1 + (bias1 ^2))
68 mse2 = mean(var2 + (bias2 ^2))
69 mse3 = mean(var3 + (bias3 ^2))
70
71 ## error prediction
72 pe1 = sum((y - Z %*% est1 [,j]) ^2 / n)
73 pe2 = sum((y - Z %*% est2 [,j]) ^2 / n)
74 pe3 = sum((y - Z %*% est3 [,j]) ^2 / n)
75
76 j=j+1
77 cat(j, " ", iter+1, "\n")
78 iter<-iter+1
79 }
80
81 ## FP = false positive number
82 fp1 = sum(est1 [4,] != 0, est1 [8:(p-1),] != 0) / (S-1)
83 fp2 = sum(est2 [4,] != 0, est2 [8:(p-1),] != 0) / (S-1)
84 fp3 = sum(est3 [4,] != 0, est3 [8:(p-1),] != 0) / (S-1)
85
86 ## FN = false negative number
87 fn1 = sum(est1 [1:3,] == 0, est1 [5:7,] == 0) / (S-1)
88 fn2 = sum(est2 [1:3,] == 0, est2 [5:7,] == 0) / (S-1)
89 fn3 = sum(est3 [1:3,] == 0, est3 [5:7,] == 0) / (S-1)
90
91 ## Hamming distance

```

```

92 ham1 = sum(est1 [1:(p-1),] != beta [1:(p-1)]) / (S-1)
93 ham2 = sum(est2 [1:(p-1),] != beta [1:(p-1)]) / (S-1)
94 ham3 = sum(est3 [1:(p-1),] != beta [1:(p-1)]) / (S-1)
95
96 diagn = round(rbind(mse1, mse2, mse3, ham1, ham2,
97                   ham3, fp1, fp2, fp3, fn1, fn2, fn3, pe1, pe2, pe3)
98               ,5)

```

```

1
2 #####
3 #      Simulation Chapter 5
4 #####
5 rm(list=ls())
6 #generate compositional data
7 rcompos <- function(n,l,p, beta , sigma){
8   y = matrix(NA, nrow=n, ncol=1)
9   Z = matrix(NA, nrow=n, ncol=p)
10  Y = matrix(NA, nrow=n, ncol=1)
11
12  mean1 <- c(rep(0,p-1)) #means for each component
13  sigma <- matrix(0.4, nrow=p-1, ncol=p-1)
14  diag(sigma) <- 2
15
16  X <- rcompnorm(n,m=mean1, s=sigma, type="alr")
17  X.mean = apply(X,2, mean)
18  ind <- order(X.mean, decreasing=T)[1:3]  ## Extraindo os 5
19  maiores componentes
20
21  X.new <- (cbind(X[, ind], X[,-ind]))
22
23  Z <- as.matrix(pivotCoord(X.new))
24
25  mu = Z%*%beta
26  mean2 = apply(mu,2, mean)
27  y = rcompnorm(n, mean2, sigmac, type="alr")
28
29  return(list(y,Z))
30 }

```

```
30 ##### Y1 #####
31 #####
32 ## SSL
33 #####
34 lambda1 <- 1
35 lambda0 <- seq(lambda1,50, length=10)
36 L <- length(lambda0)
37
38 output1.1 = list()
39 output1.2 = list()
40 output1.3 = list()
41 output2.1 = list()
42 output2.2 = list()
43 output2.3 = list()
44 output3.1 = list()
45 output3.2 = list()
46 output3.3 = list()
47 est1.1 = matrix(0,nrow=p-1,ncol=S-1)
48 est1.2 = matrix(0,nrow=p-1,ncol=S-1)
49 est1.3 = matrix(0,nrow=p-1,ncol=S-1)
50 est2.1 = matrix(0,nrow=p-1,ncol=S-1)
51 est2.2 = matrix(0,nrow=p-1,ncol=S-1)
52 est2.3 = matrix(0,nrow=p-1,ncol=S-1)
53 est3.1 = matrix(0,nrow=p-1,ncol=S-1)
54 est3.2 = matrix(0,nrow=p-1,ncol=S-1)
55 est3.3 = matrix(0,nrow=p-1,ncol=S-1)
56 est.lasso1 = matrix(0,nrow=p-1,ncol=S-1)
57 est.lasso2 = matrix(0,nrow=p-1,ncol=S-1)
58 est.lasso3 = matrix(0,nrow=p-1,ncol=S-1)
59 est.elastic1 = matrix(0,nrow=p-1,ncol=S-1)
60 est.elastic2 = matrix(0,nrow=p-1,ncol=S-1)
61 est.elastic3 = matrix(0,nrow=p-1,ncol=S-1)
62 ham.ssl1.1 = c()
63 ham.ssl1.2 = c()
64 ham.ssl1.3 = c()
65 ham.ssl2.1 = c()
66 ham.ssl2.2 = c()
67 ham.ssl2.3 = c()
68 ham.ssl3.1 = c()
```

```
69 ham.ssl3.2 = c()
70 ham.ssl3.3 = c()
71 ham.lasso1 = c()
72 ham.lasso2 = c()
73 ham.lasso3 = c()
74 ham.elastic1 = c()
75 ham.elastic2 = c()
76 ham.elastic3 = c()
77
78 ##### GENERATE COMPOSITIONAL DATA MATRIX ###
79 while(j < S){
80   jj=jj+1
81   set.seed(jj)
82
83   Y <- rcompos(n,l,p,beta,sigma)
84   y <- as.matrix(pivotCoord(Y[[1]]))
85   Z <- as.matrix(Y[[2]])
86
87   # Oracle SSLASSO with known variance (Separable)
88   result1.1 <- SSLASSO(Z, y[,1], penalty = "separable",
89     variance = "known",
90     lambda1 = lambda1, lambda0 = lambda0,
91     theta = 0.8)
92
93   # Oracle SSLASSO with known variance (Separable Oracle)
94   result1.2 <- SSLASSO(Z, y[,1], penalty = "separable",
95     variance = "known",
96     lambda1 = lambda1, lambda0 = lambda0,
97     theta = 6/p)
98
99   # Oracle SSLASSO with unknown variance (Non-Separable)
100  result1.3 <- SSLASSO(Z, y[,1], penalty = "adaptive", variance
101    = "unknown",
102    lambda1 = lambda1, lambda0 = lambda0,
103    theta = 6/p)
104 #####
105 ## Lasso
106 #####
```

```

105 result1.4 = glmnet(Z,y[,1],family="gaussian",alpha=1,
    standardize=TRUE,intercept=FALSE)
106 cv.lasso.modtotal1.4 = cv.glmnet(Z,y[,1],family="gaussian",
    alpha=1)
107 bestlam.lasso1.4=cv.lasso.modtotal1.4$lambda.1se
108
109 result1.4.opt = glmnet(Z,y[,1],alpha=1,standardize=TRUE,
    lambda=bestlam.lasso1.4,intercept=FALSE)
110 #print(coef(result1.opt), digit=3)
111 coef.lasso1 = round(matrix(coef(result1.4.opt)[-1,],nrow=p-1)
    ,4)
112
113 #####
114 ## Elastic net
115 #####
116 a <- seq(0.1, 0.9, 0.05)
117 search <- foreach(i = a, .combine = rbind) %dopar% {
118   cv <- cv.glmnet(Z, y[,1], family = "gaussian", nfold = 10,
    type.measure = "deviance", parallel = TRUE, alpha = i)
119   data.frame(cvm = cv$cvm[cv$lambda == cv$lambda.1se], lambda
    .1se = cv$lambda.1se, alpha = i)
120 }
121 cv1 <- search[search$cvm == min(search$cvm), ]
122
123 result1.5 = glmnet(Z,y[,1],family="gaussian",alpha=cv1$alpha,
    lambda=cv1$lambda.1se,intercept=FALSE,standardize=TRUE)
124
125 coef.elastic1 = round(matrix(coef(result1.5)[-1,],nrow=p-1)
    ,4)
126
127 ##### Y2 #####
128 #####
129 ## SSL
130 #####
131 # Oracle SSLASSO with known variance (Separable)
132 result2.1 <- SSLASSO(Z, y[,2], penalty = "separable",
    variance = "known",
133                 lambda1 = lambda1, lambda0 = lambda0,
134                 theta = 0.8)

```

```

135
136 # Oracle SSLASSO with known variance (Separable Oracle)
137 result2.2 <- SSLASSO(Z, y[,2], penalty = "separable",
138                   variance = "known",
139                   lambda1 = lambda1, lambda0 = lambda0,
140                   theta = 6/p)
141 # Oracle SSLASSO with unknown variance (Non-Separable)
142 result2.3 <- SSLASSO(Z, y[,2], penalty = "adaptive", variance
143                   = "unknown",
144                   lambda1 = lambda1, lambda0 = lambda0,
145                   theta = 6/p)
146 #####
147 ## Lasso
148 #####
149 result2.4 = glmnet(Z, y[,2], family="gaussian", alpha=1,
150                 standardize=TRUE, intercept=FALSE)
151 cv.lasso.modtotal2.4 = cv.glmnet(Z, y[,2], family="gaussian",
152                               alpha=1)
153 bestlam.lasso2.4= cv.lasso.modtotal2.4$lambda.1se
154 result2.4.opt = glmnet(Z, y[,2], alpha=1, standardize=TRUE,
155                       lambda=bestlam.lasso2.4, intercept=FALSE)
156 #print(coef(result1.opt), digit=3)
157 coef.lasso2 = round(matrix(coef(result2.4.opt)[-1,], nrow=p-1)
158                   ,4)
159 #####
160 ## Elastic net
161 #####
162 a <- seq(0.1, 0.9, 0.05)
163 search <- foreach(i = a, .combine = rbind) %dopar% {
164   cv <- cv.glmnet(Z, y[,2], family = "gaussian", nfold = 10,
165                 type.measure = "deviance", par = TRUE, alpha = i)
166   data.frame(cvm = cv$cvm[cv$lambda == cv$lambda.1se], lambda
167             .1se = cv$lambda.1se, alpha = i)
168 }
169 cv2 <- search[search$cvm == min(search$cvm), ]

```

```
166
167 result2.5 = glmnet(Z,y[,2],family="gaussian",alpha=cv2$alpha,
168 lambda=cv2$lambda.1se,intercept=FALSE,standardize=TRUE)
169 coef.elastic2 = round(matrix(coef(result2.5)[-1,],nrow=p-1)
170 ,4)
171 ##### OUTPUTS #####
172 output1.1[[j]] = result1.1$beta[,10]
173 output1.2[[j]] = result1.2$beta[,10]
174 output1.3[[j]] = result1.3$beta[,10]
175 output2.1[[j]] = result2.1$beta[,10]
176 output2.2[[j]] = result2.2$beta[,10]
177 output2.3[[j]] = result2.3$beta[,10]
178
179 est1.1[,j] = as.matrix(output1.1[[j]])
180 est1.2[,j] = as.matrix(output1.2[[j]])
181 est1.3[,j] = as.matrix(output1.3[[j]])
182 est2.1[,j] = as.matrix(output2.1[[j]])
183 est2.2[,j] = as.matrix(output2.2[[j]])
184 est2.3[,j] = as.matrix(output2.3[[j]])
185 est.lasso1[,j] = coef.lasso1
186 est.lasso2[,j] = coef.lasso1
187 est.elastic1[,j] = coef.elastic1
188 est.elastic2[,j] = coef.elastic2
189
190 medias1.1 = apply(est1.1,1,mean)
191 medias1.2 = apply(est1.2,1,mean)
192 medias1.3 = apply(est1.3,1,mean)
193 medias2.1 = apply(est2.1,1,mean)
194 medias2.2 = apply(est2.2,1,mean)
195 medias2.3 = apply(est2.3,1,mean)
196 medias.lasso1 = apply(est.lasso1,1,mean)
197 medias.lasso2 = apply(est.lasso2,1,mean)
198 medias.elastic1 = apply(est.elastic1,1,mean)
199 medias.elastic2 = apply(est.elastic2,1,mean)
200
201 var1.1 = apply(est1.1,1,var)
202 var1.2 = apply(est1.2,1,var)
```

```
203 var1.3 = apply(est1.3,1,var)
204 var2.1 = apply(est2.1,1,var)
205 var2.2 = apply(est2.2,1,var)
206 var2.3 = apply(est2.3,1,var)
207 var.lasso1 = apply(est.lasso1,1,var)
208 var.lasso2 = apply(est.lasso2,1,var)
209 var.elastic1 = apply(est.elastic1,1,var)
210 var.elastic2 = apply(est.elastic2,1,var)
211
212 sd1.1 = apply(est1.1,1,sd)
213 sd1.2 = apply(est1.2,1,sd)
214 sd1.3 = apply(est1.3,1,sd)
215 sd2.1 = apply(est2.1,1,sd)
216 sd2.2 = apply(est2.2,1,sd)
217 sd2.3 = apply(est2.3,1,sd)
218 sd.lasso1 = apply(est.lasso1,1,sd)
219 sd.lasso2 = apply(est.lasso2,1,sd)
220 sd.elastic1 = apply(est.elastic1,1,sd)
221 sd.elastic2 = apply(est.elastic2,1,sd)
222
223 ##### diagnostics statistics
224 bias1.1 = medias1.1 - beta1
225 bias1.2 = medias1.2 - beta1
226 bias1.3 = medias1.3 - beta1
227 bias2.1 = medias2.1 - beta2
228 bias2.2 = medias2.2 - beta2
229 bias2.3 = medias2.3 - beta2
230 bias.lasso1 = medias.lasso1 - beta1
231 bias.lasso2 = medias.lasso2 - beta2
232 bias.elastic1 = medias.elastic1 - beta1
233 bias.elastic2 = medias.elastic2 - beta2
234
235 mse1.1 = mean(var1.1 + (bias1.1^2))
236 mse1.2 = mean(var1.2 + (bias1.2^2))
237 mse1.3 = mean(var1.3 + (bias1.3^2))
238 mse2.1 = mean(var2.1 + (bias2.1^2))
239 mse2.2 = mean(var2.2 + (bias2.2^2))
240 mse2.3 = mean(var2.3 + (bias2.3^2))
241 mse.lasso1 = mean(var.lasso1 + (bias.lasso1^2))
```

```

242 mse.lasso2 = mean(var.lasso2 + (bias.lasso2^2))
243 mse.elastic1 = mean(var.elastic1 + (bias.elastic1^2))
244 mse.elastic2 = mean(var.elastic2 + (bias.elastic2^2))
245
246 ## error prediction
247 pe1.1 = sum(y[,1]-Z%*%est1.1[,j])^2/n
248 pe1.2 = sum(y[,1]-Z%*%est1.2[,j])^2/n
249 pe1.3 = sum(y[,1]-Z%*%est1.3[,j])^2/n
250 pe2.1 = sum(y[,2]-Z%*%est2.1[,j])^2/n
251 pe2.2 = sum(y[,2]-Z%*%est2.2[,j])^2/n
252 pe2.3 = sum(y[,2]-Z%*%est2.3[,j])^2/n
253 pe.lasso1 = sum(y[,1]-Z%*%est.lasso1[,j])^2/n
254 pe.lasso2 = sum(y[,2]-Z%*%est.lasso2[,j])^2/n
255 pe.elastic1 = sum(y[,1]-Z%*%est.elastic1[,j])^2/n
256 pe.elastic2 = sum(y[,2]-Z%*%est.elastic2[,j])^2/n
257
258 j=j+1
259 cat(j," ",iter+1,"\n")
260 iter<-iter+1
261 }
262
263 ## FP = false positive number
264 fp1.1 = sum(est1.1[4,] != 0, est1.1[8:(p-1),] != 0)/(S-1)
265 fp1.2 = sum(est1.2[4,] != 0, est1.2[8:(p-1),] != 0)/(S-1)
266 fp1.3 = sum(est1.3[4,] != 0, est1.3[8:(p-1),] != 0)/(S-1)
267 fp2.1 = sum(est2.1[4,] != 0, est2.1[8:(p-1),] != 0)/(S-1)
268 fp2.2 = sum(est2.2[4,] != 0, est2.2[8:(p-1),] != 0)/(S-1)
269 fp2.3 = sum(est2.3[4,] != 0, est2.3[8:(p-1),] != 0)/(S-1)
270 fp.lasso1 = sum(est.lasso1[4,] != 0, est.lasso1[8:(p-1),] != 0)
      /(S-1)
271 fp.lasso2 = sum(est.lasso2[4,] != 0, est.lasso2[8:(p-1),] != 0)
      /(S-1)
272 fp.elastic1 = sum(est.elastic1[4,] != 0, est.elastic1[8:(p-1),]
      != 0)/(S-1)
273 fp.elastic2 = sum(est.elastic2[4,] != 0, est.elastic2[8:(p-1),]
      != 0)/(S-1)
274
275 ## FN = false negative number
276 fn1.1 = sum(est1.1[1:3,] == 0, est1.1[5:7,] == 0)/(S-1)

```

```
277 fn1.2 = sum(est1.2[1:3,] == 0, est1.2[5:7,] == 0)/(S-1)
278 fn1.3 = sum(est1.3[1:3,] == 0, est1.3[5:7,] == 0)/(S-1)
279 fn2.1 = sum(est2.1[1:3,] == 0, est2.1[5:7,] == 0)/(S-1)
280 fn2.2 = sum(est2.2[1:3,] == 0, est2.2[5:7,] == 0)/(S-1)
281 fn2.3 = sum(est2.3[1:3,] == 0, est2.3[5:7,] == 0)/(S-1)
282 fn.lasso1 = sum(est.lasso1[1:3,] == 0, est.lasso1[5:7,] == 0)/(
      S-1)
283 fn.lasso2 = sum(est.lasso2[1:3,] == 0, est.lasso2[5:7,] == 0)/(
      S-1)
284 fn.elastic1 = sum(est.elastic1[1:3,] == 0, est.elastic1[5:7,]
      == 0)/(S-1)
285 fn.elastic2 = sum(est.elastic2[1:3,] == 0, est.elastic2[5:7,]
      == 0)/(S-1)
286
287 ## Hamming distance
288 ham.ssl1.1 = sum(est1.1[1:(p-1),] != beta1[1:(p-1)])/(S-1)
289 ham.ssl1.2 = sum(est1.2[1:(p-1),] != beta1[1:(p-1)])/(S-1)
290 ham.ssl1.3 = sum(est1.3[1:(p-1),] != beta1[1:(p-1)])/(S-1)
291 ham.ssl2.1 = sum(est2.1[1:(p-1),] != beta2[1:(p-1)])/(S-1)
292 ham.ssl2.2 = sum(est2.2[1:(p-1),] != beta2[1:(p-1)])/(S-1)
293 ham.ssl2.3 = sum(est2.3[1:(p-1),] != beta2[1:(p-1)])/(S-1)
294 ham.lasso1 = sum(est.lasso1[1:(p-1),] != beta1[1:(p-1)])/(S-1)
295 ham.lasso2 = sum(est.lasso2[1:(p-1),] != beta2[1:(p-1)])/(S-1)
296 ham.elastic1 = sum(est.elastic1[1:(p-1),] != beta1[1:(p-1)])/(S
      -1)
297 ham.elastic2 = sum(est.elastic2[1:(p-1),] != beta2[1:(p-1)])/(S
      -1)
```
