# AVERTISSEMENT



# LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n° 92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992

http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

# THÈSE

**En vue de l'obtention du**

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :** *l'Université Toulouse 1 Capitole (UT1 Capitole)*

---

**Présentée et soutenue le** *Date de soutenance* **par :**

Thi Huong An NGUYEN

---

**Contribution to the statistical analysis of compositional data with an application to political economy**

---

**JURY**

| | | |
|---|---|---|
| Josep Antoni MARTIN-FERNANDEZ | Professeur d'Université | Rapporteur |
| Peter FILZMOSER | Professeur d'Université | Rapporteur |
| Anne RUIZ-GAZEN | Professeur d'Université | Directrice de thèse |
| Christine THOMAS-AGNAN | Professeur d'Université | Directrice de thèse |

---

**École doctorale et spécialité :**
*MITT : Domaine Mathématiques : Mathématiques appliquées*

**Unité de Recherche :**
*École d'Économie de Toulouse (TSE-R)*

**Directeur(s) de Thèse :**
*Christine THOMAS-AGNAN* et *Anne RUIZ-GAZEN*

**Rapporteurs :**
*Josep Antoni MARTIN-FERNANDEZ* et *Peter FILZMOSER*

# Acknowledgments

This thesis has been sponsored by the Vietnamese Government for three years.

This is the moment when I write some important things in my life.

First of all, I would like to thank my directors, Professor Christine Thomas-Agnan and Professor Anne Ruiz-Gazen. You spent a lot of time taking care of me, encouraging me, sharing with me all your ideas on my subject. You advised me not only on the scientific world but also in French culture and especially in my personal life. I am very proud of having the chance to work with both of you.

Secondly, I also want to thank the two referees Josep Antoni Martin-Fernandez and Peter Filzmoser. It is a great honor for me to have my thesis evaluated by admirable experts in compositional data analysis and in statistics. My gratitude also goes to all members of my jury.

Thirdly, thanks a lot to Thibault Laurent, who shared with me his experience with the software R. It has been a real pleasure to work with you. I hope that I have enriched my programming skills with you.

Fourthly, I would like to thank Cécile Ha Minh Tu, who transmits, motivates and gives me the chance to work in an industrial environment, and helped me improving my French.

Fifthly, I would like to thank my "adopted parents" Jean and Marie Perraud who treat me as their daughter. There is a lot to tell about this relationship. Studying in France has been the occasion for me to meet my "second parents". Jean, my second father, motivated me when I did not want to continue my research. Both of you helped me a lot in academic life as well as in personal life. Marie, my second mother, has taken care of me and gave me more courage to try my best. You also helped me to take care of my son Vinh so that I could have time to complete this thesis. And there are many, many things that I will keep in my heart.

Last but not least, thank you to my husband and my son who motivated me to complete my research. Thanks a lot to my mother, my brothers, my nephew Thuong Nguyen. Thank you very much Thien Trang Bui, my "sister" in Toulouse, professor Nguyen Tien Dung and the Vietnamese student group in Toulouse.

# Contents

**Introduction (English version)** **5**

**Introduction (version française)** **13**

**1 General compositional data analysis** **21**
1.1 Concepts of compositional data analysis . . . . . 22
1.1.1 Composition, subcomposition . . . . . 22
1.1.2 Principles of compositional data analysis . . . . . 22
1.1.3 Operations on the simplex . . . . . 24
1.1.4 The log-ratio transformations . . . . . 25
1.2 Distributions for random compositions . . . . . 28
1.2.1 The normal distribution on the simplex . . . . . 29
1.3 The CODA regression models . . . . . 29
1.3.1 Notations . . . . . 29
1.3.2 The CODA regression models: expression in the simplex space and in the coordinates space. . . . . 29

**2 Analyzing the impacts of socio-economic factors on French departmental elections with CODA methods** **31**
2.1 Introduction . . . . . 36
2.2 Data . . . . . 37
2.3 Compositional data analysis approach . . . . . 38
2.3.1 Principles of compositional data analysis . . . . . 38
2.3.2 Compositional regression models . . . . . 41
2.3.3 Impact of compositional and classical explanatory variables . . . . 42
2.4 Conclusion . . . . . 46

**3 Multivariate Student versus Multivariate Gaussian Regression Models with Application to Finance** **49**
3.1 Introduction . . . . . 53
3.2 Multivariate Regression Models . . . . . 55
3.2.1 Literature Review . . . . . 55
3.2.2 Univariate Regression Case Reminder . . . . . 57

3.2.3 The Multivariate Regression Model . . . 58
3.2.4 Multivariate Normal Error Vector . . . 59
3.2.5 Uncorrelated Multivariate Student (UT) Error Vector . . . 59
3.2.6 Independent Multivariate Student (IT) Error Vector . . . 60
3.3 Simulation Study . . . 62
3.3.1 Design . . . 62
3.3.2 Estimators of the $\beta$ Parameters . . . 63
3.3.3 Estimators of the Variance Parameters . . . 64
3.4 Selection between the Gaussian and IT Models . . . 67
3.4.1 Distributions of Mahalanobis Distances . . . 67
3.4.2 Examples . . . 68
3.5 Conclusions . . . 72

**4 CODA methods and the multivariate Student distribution with an application to political economy 75**
4.1 Introduction . . . 79
4.2 Data . . . 81
4.3 The multivariate regression models . . . 84
4.3.1 Multivariate Normal error vector . . . 84
4.3.2 Multivariate Independent Student error vector . . . 84
4.4 Compositional regression models . . . 85
4.4.1 Principles of compositional data analysis . . . 85
4.4.2 Logistic Student regression models . . . 87
4.4.3 Application to political economy . . . 88
4.5 Model selection . . . 89
4.6 Vote shares predictions . . . 91
4.7 Conclusion . . . 93

**5 A spatial autoregressive model for compositional data 95**
5.1 Introduction . . . 99
5.2 Definitions and notations in compositional data analysis . . . 100
5.3 Multivariate LAG regression model . . . 103
5.3.1 Model in coordinate space . . . 103
5.3.2 Writing the LAG regression model in the simplex space . . . 105
5.4 Simulation . . . 105
5.5 Application to political economics . . . 107
5.6 Conclusion . . . 109
5.7 Acknowledgements . . . 110

**Conclusion (English version) 111**

**Conclusion (version française) 113**

**Appendix A Appendix for Chapter 3 115**

**Appendix B Appendix for Chapter 5** **119**

# Introduction(English version)

Election is one of the basic principles of democracy where people get to choose their leaders. Election is the pillar of any democratic country, since democracy is defined as the government by the people, for the people which denotes that people are the source of democracy and the absolute sovereignty belongs to them. In this sense, elections are held so that people can choose their representatives who will represent different interests, policies, etc, ... formulated for the welfare of the people. Another way to sum it up, the important reason for an election is to make sure that people have the opportunity to participate in political affair indirectly.

There are three main party systems: one-party, two-party and multiparty system[1]. A one-party system is a type of electoral system in which one political party takes over power and does not allow other parties to run their candidates for election. Vietnam is an example of nations in which one party controls the government. A two-party system has only two political parties that have the possibility of winning an election. The United States is a classical example with a two-party system. However, a multiparty system is also popular. In such a political system, multiple political parties compete for national election, and all have the capacity to gain control of government offices, separately or in coalition. France is one of the nations with a multiparty system with at least fifteen political parties. In 2019, the dominating political parties in France are "En Marche" (centrist and liberal), the "Rassemblement National" (right-wing populist/nationalist), "La France Insoumise" (left-wing and relatively socialist-leaning), "Les Républicains" (conservative), and the Socialist Party (left-wing).

Recently, a lot of authors in political economy concentrate on building models and understanding the drivers of the outcome of a two-party electoral system (Beauguitte and Colange (2013), Ansolabehere and Leblanc (2008)). Besides, multiparty system have also attracted attention. France is an example of a nation that has used a multiparty system effectively in its democracy. There are at least fifteen political parties in France. Based on result of the French departmental election in 2015, we aggregate these political parties into three main political parties: Left, Right and Extreme Right. In this multiparty system, the outcomes of the election consists of vectors whose components are the percentages or proportions of votes per party. Their sum is therefore constrained to be constant, equal to 1 for proportions, 100 for percentages. This type of data is called

[1]https://en.wikipedia.org/wiki/Multi-party_system

compositional data (CODA).

The outcome of an election can be influenced by the campaign strategies of candidates, demographic factors such as age, domain of activity, rate of unemployment, and so on. In an interview with Time magazine, a group of Obama senior campaign advisers revealed an enormous data effort to support fundraising, micro-targeting TV ads and modeling of swing-state voters. Therefore, it is interesting to investigate in the French multiparty system. In this thesis, we study the French departmental election in 2015 with the following questions. Firstly, we would like to understand the drivers of the outcome of an election or how to predict the outcome of an election. The election cost is usually very expensive. In terms of decision-making, the outcome of an election involves uncertainty. Forecasts of election results can cut down risks for decision-makers and thus smoother decision-making. Forecasting the outcome of an event can be of use for experts in several different areas, for example financial strategists, political strategists, policy makers and so on. In order to predict the outcome of election, the most important thing is to define a model with some core factors which have an impact on the outcome. Secondly, some papers reveal that electoral data exibit heavy tail behavior (see Nguyen et al. (2019)). The question is how to treat these data and which model will better fit these data. Thirdly, the electoral data in French departmental election are observed at some spatial scale. Thus, it is interesting to study how the decisions in voting are influenced by their neighbors decisions.

This thesis concentrates on applications of compositional data analysis regression models which led me to generalize these models in several directions. At first, we were interested in exploring the 2015 French departmental election. Our attention focuses on the behavior of the electors and the relation between votes and demographic and social factors such as age, education levels, domain of activities, unemployment rate and so on. Besides, the outcome of the election in France is observed at the departmental as well as the canton levels. To go beyond the application of classical compositional regression models, we consider the fact, documented in the literature, that electoral data may exhibit heavy tail behaviors and also spatial autocorrelation. Therefore, it is complicated to analyze these data with classical methods due to the constraints and the spatial correlation issue. At the beginning of my manuscript, I present the principles of compositional data analysis. I build a regression model that could be considered to explain the outcome of an election and to clarify its relations with the socio-economic factors. Furthermore, this model allows to forecast the outcome of an election and the winning party. In order to eliminate the heavy tail problem, a proposal found in the literature is to replace the Gaussian distribution by the Student distribution. However, since there is not a unique way of using the multivariate Student distribution in a multivariate regression model, we first need to study the properties of two competing models: the uncorrelated Student (UT) and the Independent Student (IT) models. We also provide some supplemental material for the R implementation of the estimation of these models. For the departmental election data, we compare the results between the Gaussian and the Student (IT) models. In order to choose one model, we carry out a test based on the Mahalanobis distance. The Kolmogorov–Smirnov test tells us that we

do not reject the IT model and do reject the Normal model. Thereby, we use this IT model to predict the vote share to see the impact of socio-economic factors. Finally, we take into account possible spatial autocorrelation in compositional data. We show how to use multivariate versions of the classical spatial autoregressive model for areal data used in spatial econometrics in the coordinate space after a proper transformation of our vectors of proportions.

## Description of issues

My thesis will address the four following issues:

**Issue 1**: Our CODA regression model contains variables which are both of the classical and of the compositional type. One of the challenges in this type of model is to provide sensible interpretations of the results. An interpretation in the log-ratio space is mathematically easy but not natural to answer practical questions. On the other hand, an interpretation in the simplex space is complex because of its constrained nature. We look at how CODA regression models can improve our practical understanding of the electoral results.

**Issue 2**: Distributions with heavy tails are distributions which allow extreme values more often. Because of its inability to model heavy tails, the classical multivariate Gaussian model is limited. This is a reason for turning attention to a more flexible family of distributions as the multivariate Student distribution. In one dimension, the generalized Student distribution contains the Gaussian models as a limit when the shape parameter is large. However, the difficulty in higher dimensions is that, there is not a unique way of using the multivariate Student distribution (see Johnson and Kotz (1972), Kotz and Nadarajah (2004)). Two options are described in Kelejian and Prucha (1985) for the case of univariate regression when using the multivariate Student distribution to define a univariate regression model. Furthermore, the property of equivalence between independence and uncorrelatedness for components of a Gaussian vector are not satisfied anymore for a multivariate Student vector. This motivates us to investigate further the multivariate Student regression models.

**Issue 3**: It is mentioned in the literature that electoral data often exhibit heavy tails behavior. In order to verify which model fits better our electoral data, we propose a test based on the Mahalanobis distance. The merit of this approach is that the Mahalanobis distance is a one-dimensional variable while the original observations are multidimensional variables. Moreover, its approximate distribution is known under the Gaussian model and under the Student model. Therefore, an approach is to test whether the squared Mahalanibis distances follow the Chi-square distribution for testing the normality and follow the Fisher distribution for testing the Student distribution.

**Issue 4**: A problem using conventional statistical methods in electoral analysis is that the assumption of statistical independence across statistical units may be questionable for units which are indeed geographical areas. Such type of data usually exhibit spatial autocorrelation. Since electoral data also have the compositional nature, we need to

study a spatial model which can adapted to this particular feature.

### Contributions

Among papers concentrating on the relationship between socio-economic variables and election results, Beauguitte and Colange (2013) study a linear regression at three levels of aggregation (polling stations, cities and electoral districts) in France and show that the socio-economic variables are significant in France. Kavanagh et al. (2006) use geographically weighted regression, which produces parameter estimates for each data point, i.e. for each electoral division. On the other hand in the statistical literature, people have developed CODA regression models where the dependent and independent variables may be compositional variables (see Mert et al. (2018) for a review). Morais (2017) studies the impact of media investments on brand's market shares with a CODA regression model. Trinh and Morais (2017) use a CODA regression model to highlight the nutrition transition and to explain it according to household characteristics. Honaker et al. (2002), Katz and King (1999) use a statistical model for multiparty electoral data assuming that the territorial units yield independent observations. We first propose a statistical model for studying the multiparty system using compositional data analysis (CODA) with departmental level data. The dependent vector is the vector of vote shares for the French departmental election in 2015. The explanatory variables include some compositional and classical socio-economic variables such as proportions in age groups, diploma groups, domain of activity, rate of unemployment, and so on. We present several exploratory plots to study the impact of explanatory variables of a classical type and compositional type and show the relationships between coefficients in the simplex space and in the coordinate space.

For modeling multivariate possibly heavy tailed data, at first we recall the multivariate Normal distribution and present a nice complete summary for two options of the multivariate Student distribution. We then compare the multivariate Normal model (N) with two versions of the multivariate Student model: the independent multivariate Student (IT) and the uncorrelated (UT) multivariate Student. The UT regression model is first introduced by Zellner (1976) for the case of univariate regression. Other references in that case include Singh (1988) and Kelejian and Prucha (1985). The IT univariate regression model is studied in Fraser (1979) and Kelejian and Prucha (1985). In the multivariate case, the UT model is studied in Sutradhar and Ali (1986), with a method of moments estimation approach. The multivariate IT model is found for example in Prucha and Kelejian (1984) and Katz and King (1999). We then prove that the maximum likelihood estimator of the variance-covariance matrix in the multivariate UT model is asymptotically biased and propose an unbiased version. For the IT model, we propose an iterative reweighted least squares algorithm to compute the maximum likelihood estimators. We present a simulation study to compare the bias and root mean square error of the ensuing estimators of the regression coefficients and variance-covariance matrix under several scenarios of the potential data generating process, mispecified or not. We propose a graphical tool based on the Mahalanobis distance to guide the choice between

the competing models. We illustrate an application in finance about a series of daily close share prices of IBM and MSFT in the period 2007-2018.

After clarifying the multivariate Student distribution, we present two applications inspired by the following papers. In the political economy literature, for the case of fully contested elections, Katz and King (1999) combine log-ratio transformations classically used in compositional data analysis with an IT type of multivariate Student regression model to define the additive logistic Student distribution and find that it is superior to previous models for multiparty voting data. In the finance literature, several authors Platen and Rendek (2008), Fung and Seneta (2010)) advocate the use of the Student distribution to model log returns of financial assets. In our case, we compare the Gaussian and the Student IT model for the French departmental election data for canton level.

Spatial models have been considered for studying election outcomes. Kelejian and Prucha (2004) introduce a series-type instrumental variable (IV) estimator of the parameters of a spatial first order autoregressive model with first order autoregressive disturbances. Then via Monte Carlo techniques, Kelejian et al. (2004) give small sample results relating to their suggested estimator, the maximum likelihood (ML) estimator, and other IV estimators for univariate model. Sutter (2005) examines the spatial autoregressive relationship between county-level voting outcomes in the 2000 Presidential election in the US. Mansley and Demšar (2015) explores geographic variability in relationships between the turnout at the London mayoral election and socio-demographic variables at a detailed spatial level. His analysis is approached through geographically weighted regression (GWR), which enables the investigation of local variations in voting patterns. On the other hand, some spatial model for compositional data have been developped. Tjelmeland and Lund (2003) consider a Bayesian framework for CODA regression, discuss appropriate prior distributions and define efficient Markov chain Monte Carlo algorithms. Pirzamanbein (2015) constructs a hierarchical model for spatial compositional data using a Gaussian Markov Random Field (GMRF) with Dirichlet observations. Pawlowsky-Glahn et al. (2015) uses the additive-logratio transformation of generalized compositions to deal with the spatial covariance structure in a geostatistical fashion. We study whether one can introduce spatial autocorrelation in IT models from a spatial econometrics point of view, i.e. using simultaneous spatial autoregressive (LAG) model.

## Data

The electoral data used in my thesis are collected from the Cartelec website[2] and the INSEE website[3]. For more detail, we collect from the Cartelec website vote shares data of the 2015 French departmental election for 95 departments in France and for 207 cantons of Occitanie region in France and we download corresponding socio-economic data (for 2014) from the INSEE website. Moreover, two databases have been used in my thesis: one for the departmental level and one for the canton level. Both databases contain:

[2]https://www.data.gouv.fr/fr/datasets/elections-departementales-2015-resultats-par-bureaux-de-vote/

[3]https://www.insee.fr/fr/statistiques

1. **Vote shares**: the outcome of the French departmental election where the number of vote per party are aggregated into three big parties which are **Left (L)**, **Right (R)**, and **Extreme Right (XR)**.

2. **Age**: the age of voters. It has three components **Age_**1840 for people from 18 to 40 years old, **Age_**4064 for people from 40 to 65 years old, and **Age_**65 for elderly.

3. **Diploma**: the educational level of voters. It also has three components: <**BAC** for people with at most some secondary education, **BAC** for people with at least some secondary education and at most a high school diploma, and **SUP** for people with a university diploma.

4. **Employment** with five categories: **AZ** (agriculture, fisheries), **BE** (manufacturing industry, mining industry and others), **FZ** (construction), **GU** (business, transport and services) and **OQ** (public administration, teaching, human health).

5. **Unemployment rate**: the unemployment rate (**unemp**).

6. **Employment evolution**: the mean annual growth rate of employment (**emp_evol**).

7. **Owner**: the proportion of people who own assets (**owner**).

8. **Income**: the proportion of people who pay income tax (**income**).

9. **Foreign**: the proportion of foreigners (**foreign**).

### Structure of the thesis

Most of this thesis is written in English, except for the introduction, the summaries of the four chapters and the conclusion which are in French. This thesis is partitioned into five chapters.

**Chapter 1** is a general presentation of compositional data analysis. It presents the definitions of composition and subcomposition, the principles of compositional data analysis, some operations such as perturbation and powering in the simplex. Because of constraints inherent to compositonal data, classical statistical methods cannot be used directly in the simplex. Therefore, some log-ratio transformations are usually applied. The chapter also introduces the Normal distribution for random compositions and the CODA regression model.

**Chapter 2** is an illustration of compositional data regression model with the Normal distribution in the simplex. This chapter presents the CODA regression models both in the simplex space and in the ilr coodinates space. It illustrates the relationship between parameters of these models in the two spaces. We analyze how the predicted values in these models vary with the predictors. We also propose new graphical tools to explore the impact of some socio-economic variables on election results with the departmental level data.

**Chapter 3** is about the comparison between the multivariate Normal distribution and the multivariate Student distribution and its application in financial assets returns data. We recall the multivariate Normal distribution and clarify two versions of the multivariate Student distribution. We construct a simulation to compare the bias, the root mean squared error and covariance matrix of estimators of the regression coefficients under different scenarios of the potential data-generating process (DGP). A test based on the Mahalanobis distance is proposed to select a model.

**Chapter 4** is about the multivariate Student distribution and its application in political economy. Some researchers show that the electoral data often exhibit the heavy tails behavior. Using Chapter 3, we propose to replace the Gaussian distribution by the Student distribution. Thus defining the Student distribution in the simplex. We apply the test developped in Chapter 3 to choose between the Normal and the Student models.

**Chapter 5** develops a spatial LAG regression model for compositional data then rewrite this model in the coordinate space which include both spatial correlation and correlations across equations. In order to estimate this model, we adopt instrumental variable (IV) estimator of the parameters of a spatial autoregressive model in a multivariate setting as in Kelejian and Prucha (2004). We present a simulation to compare the relative root mean square error(RRMSE) of parameters estimate under several data generating processes (DGP) between the spatial two-stages least square (S2SLS) and spatial three-stages least squre (S3SLS). An example in political economy are also illustrated.

# Introduction (version française)

L'élection est l'un des principes de base de la démocratie selon laquelle les citoyens doivent choisir leurs dirigeants. L'élection est le pilier de tout pays démocratique, puisque la démocratie est définie comme le gouvernement par le peuple, ce qui signifie que le peuple est la source de la démocratie et que la souveraineté absolue leur appartient. En ce sens, des élections sont organisées pour que les gens puissent choisir leurs représentants qui représenteront différents intérêts, politiques, etc., formulés pour le bien-être de la population. Autrement dit, la principale raison d'une élection est de s'assurer que les gens ont la possibilité de participer indirectement aux affaires politiques.

Il existe trois principaux systèmes de parti: le système à parti unique, à deux partis et multipartis [4]. Un système à parti unique est un type de système électoral dans lequel un parti politique prend le pouvoir et ne permet pas aux autres partis de présenter leurs candidats aux élections. Le Vietnam est un exemple de pays dans lequel un parti contrôle le gouvernement. Un système à deux partis ne compte que deux partis politiques susceptibles de remporter des élections. Les États-Unis sont un exemple classique avec un système à deux partis. Cependant, un système multipartite est également populaire. Dans un tel système politique, plusieurs partis politiques se disputent les élections nationales et ont tous la capacité de prendre le contrôle des bureaux du gouvernement, séparément ou en coalition. La France est l'un des pays du multipartisme qui a plus de quinze partis politiques. En 2019, les partis politiques dominants en France sont "En Marche" (centristes et libéraux), le "Rassemblement National" (populiste / nationaliste de droite), "La France Insoumise" (de gauche et relativement socialiste), "Les Républicains" (conservateur) et le Parti socialiste (de gauche).

Récemment, de nombreux auteurs en économie politique se sont concentrés sur la construction de modèles et sur la compréhension des facteurs déterminants du résultat d'un système électoral à deux partis (Beauguitte and Colange (2013), Ansolabehere and Leblanc (2008)). Par ailleurs, le système multipartite a également attiré l'attention. La France est un exemple de nation qui a utilisé efficacement le système multipartite dans sa démocratie. Il y a au moins quinze partis politiques en France. Sur la base des résultats des élections départementales françaises de 2015, nous regroupons ces partis politiques en trois principaux partis politiques: gauche, droite et extrême droite. Dans ce système multipartite, les résultats de l'élection se composent de vecteurs dont les composantes sont les pourcentages ou les proportions de votes par parti. Leur somme

[4] https://en.wikipedia.org/wiki/Multi-party_system

est donc contrainte à être constante, égale à 1 pour les proportions, à 100 pour les pourcentages. Ce type de données est appelé données de composition (CODA).

Le résultat d'une élection peut être influencé par les stratégies de campagne des candidats, des facteurs démographiques tels que l'âge, le domaine d'activité, le taux de chômage, etc. Dans un entretien avec le magazine Time, un groupe de conseillers de campagne d'Obama a révélé un effort considérable en matière de données pour soutenir la collecte de fonds, les annonces télévisées de micro-ciblage et la modélisation des électeurs des états pivots. Il est donc intéressant de s'intéresser au système multipartite français. Dans cette thèse, nous étudions l'élection départementale française en 2015 avec les questions suivantes. Premièrement, nous aimerions comprendre les facteurs qui déterminent le résultat d'une élection ou comment prédire le résultat d'une élection. Le coût des élections est généralement très coûteux. En terme de prise de décision, le résultat d'une élection implique une incertitude. Les prévisions des résultats des élections peuvent réduire les risques pour les décideurs et donc faciliter la prise de décision. La prévision des résultats d'un événement peut être utile pour les experts de plusieurs domaines, tels que les stratèges financiers, les stratèges politiques, les décideurs, etc. Afin de prédire le résultat de l'élection, le plus important est de définir un modèle avec certains facteurs essentiels ayant une incidence sur le résultat. Deuxièmement, certains documents révèlent que les données électorales révèlent un comportement extrêmement lourd (voir Nguyen et al. (2019)). La question est de savoir comment traiter ces données et quel modèle conviendra le mieux à ces données. Troisièmement, les données électorales des élections départementales françaises sont observées à une certaine échelle spatiale. Il est donc intéressant d'étudier comment les décisions de vote sont influencées par les décisions de leurs voisins. Cette thèse se concentre sur des généralisations des modèles de régression d'analyse de données compositionnelles dans plusieurs directions. Au début, nous nous sommes intéressés à l'exploration des élections départementales françaises de 2015. Notre attention se concentre sur le comportement des électeurs et la relation entre les votes et les facteurs démographiques et sociaux tels que l'âge, le niveau d'éducation, le domaine d'activité, le taux de chômage, etc. En outre, l'issue des élections en France s'observe aux niveaux départemental et cantonal. Pour aller au-delà de l'application des modèles classiques de régression compositionnelle, nous considérons le fait, documenté dans la littérature, que les données électorales peuvent présenter des comportements extrêmes ainsi qu'une autocorrélation spatiale. Il est donc compliqué d'analyser ces données avec des méthodes classiques en raison des contraintes et du problème de corrélation spatiale. Au début de mon manuscrit, je présente les principes de l'analyse des données compositionnelles. Je construis un modèle de régression qui pourrait être envisagé pour expliquer le résultat d'une élection et clarifier ses relations avec les facteurs socio-économiques. De plus, ce modèle permet de prévoir le résultat d'une élection et du parti vainqueur. Afin d'éliminer le problème de la queue lourde, une proposition trouvée dans la littérature est de remplacer la distribution gaussienne par la distribution de Student. Toutefois, comme il n'existe pas de manière unique d'utiliser la distribution multivariée de Student dans un modèle de régression multivariée, nous devons d'abord étudier les propriétés de deux modèles concurrents: les modèles de Student non corrélé (UT) et de Student in-

dépendant (IT). Nous fournissons également des éléments supplémentaires pour la mise en œuvre R de l'estimation de ces modèles. Pour les données électorales départementales, nous comparons les résultats entre les modèles Gaussien et Student (IT). Afin de choisir un modèle, nous effectuons un test basé sur la distance de Mahalanobis. Le test de Kolmogorov – Smirnov nous indique que nous ne rejetons pas le modèle IT et que nous rejetons le modèle Normal. Ainsi, nous utilisons ce modèle IT pour prédire la part de vote et voir l'impact des facteurs socio-économiques. Enfin, nous prenons en compte une possible autocorrélation spatiale dans les données de composition. Nous montrons comment utiliser des versions multivariées du modèle classique autorégressif spatial pour les données surfaciques utilisées en économétrie spatiale dans l'espace de coordonnées après une transformation appropriée de nos vecteurs de proportions.

## Description des problèmes

Ma thèse portera sur les quatre problèmes suivants:

**Problème 1**: Notre modèle de régression CODA contient des variables qui sont à la fois du type classique et du type compositionnel. L'un des défis de ce type de modèle est de fournir des interprétations judicieuses des résultats. Une interprétation dans l'espace log-ratio est mathématiquement facile mais pas naturel pour répondre à des questions pratiques. D'autre part, une interprétation dans l'espace du simplexe est complexe en raison de sa nature contrainte. Nous examinons comment les modèles de régression CODA peuvent améliorer notre compréhension pratique des résultats électoraux.

**Problème 2**: Les distributions avec des queues épaisses sont des distributions qui permettent plus souvent des valeurs extrêmes. En raison de son incapacité à modéliser des queues lourdes, le modèle gaussien classique à plusieurs variables est limité. C'est une raison pour attirer l'attention sur une famille de distributions plus flexible, comme la distribution multivariée de Student. En une dimension, la distribution de Student généralisée contient les modèles gaussiens en tant que limite lorsque le paramètre de forme est grand. Cependant, la difficulté dans les dimensions supérieures est qu'il n'existe pas de manière unique de définir la distribution multivariée de Student (voir Johnson and Kotz (1972), Kotz and Nadarajah (2004)). Deux options sont décrites dans Kelejian and Prucha (1985) dans le cas de la régression (à variable dépendante) univariée lors de l'utilisation de la distribution multivariée de Student. De plus, la propriété d'équivalence entre indépendance et décorrélation pour les composantes d'un vecteur Gaussien n'est plus satisfaite pour un vecteur de Student multivarié. Cela nous incite à approfondir nos recherches sur les modèles de régression multivariés de Student.

**Problème 3**: Il est mentionné dans la littérature que les données électorales montrent souvent un comportement à queue lourde. Afin de trouver le modèle qui correspond le mieux à nos données électorales, nous proposons un test basé sur la distance de Mahalanobis. Le mérite de cette approche est que la distance de Mahalanobis est une variable unidimensionnelle alors que les observations originales sont des variables multidimension-

nelles. De plus, la distribution de cette distance au carré est connue approximativement sous le modèle gaussien et sous le modèle de Student. Par conséquent, une approche consiste à tester si les distances carrées de Mahalanobis suivent la distribution du chi carré pour tester la normalité et suivent la distribution de Fisher pour tester la distribution de Student.

**Problème 4**: L'utilisation des méthodes statistiques classiques en analyse électorale pose un problème: l'hypothèse d'une indépendance statistique entre les unités statistiques peut être remise en question puisque les unités sont des zones géographiques. Ce type de données présente généralement une autocorrélation spatiale. Étant donné que les données électorales ont également un caractère compositionnel, nous devons étudier un modèle spatial pouvant être adapté à cette caractéristique particulière.

## Contributions

Parmi les articles se concentrant sur la relation entre les variables socio-économiques et les résultats des élections, Beauguitte and Colange (2013) étudient une régression linéaire à trois niveaux d'agrégation (bureaux de vote, villes et districts électoraux) en France et montrent que les variables socio-économiques sont significatives. Kavanagh et al. (2006) utilisent une régression pondérée géographiquement, qui produit des estimations de paramètres pour chaque point de données, c'est-à-dire pour chaque circonscription électorale. D'autre part, dans la littérature statistique, les gens ont développé des modèles de régression CODA dans lesquels les variables dépendantes et indépendantes peuvent être des variables de composition (voir Mert et al. (2018) pour une revue). Morais (2017) étudie l'impact des investissements des médias sur les parts de marché de la marque à l'aide d'un modèle de régression CODA. Trinh and Morais (2017) utilisent un modèle de régression CODA pour mettre en évidence la transition nutritionnelle et l'expliquer en fonction des caractéristiques du ménage. Honaker et al. (2002), Katz and King (1999) utilisent un modèle statistique pour les données électorales multipartites en supposant que les unités territoriales produisent des observations indépendantes. Nous proposons d'abord un modèle statistique pour l'étude du système multipartite utilisant l'analyse de données compositionnelles (CODA) avec des données de niveau départemental. Les variables dépendantes sont les vecteurs des parts de vote pour l'élection départementale française de 2015. Les variables explicatives incluent certaines variables socio-économiques classiques et de composition telles que les proportions dans les groupes d'âge, les groupes de diplômes, le domaine d'activité, le taux de chômage, etc. Nous présentons plusieurs graphiques exploratoires pour étudier l'impact de variables explicatives de types classique et compositionnel et pour montrer les relations entre les coefficients dans l'espace du simplexe et dans l'espace des coordonnées.

Pour modéliser des données multivariées, nous rappelons tout d'abord la distribution multivariée Normale et présentons un résumé complet et synthétique pour deux options de la distribution multivariée de Student. Nous comparons ensuite le modèle multivarié Normal (N) avec deux versions du modèle de Student multivarié: le Student multivarié indépendant (IT) et le Student multivarié non corrélé (UT). Le modèle de régression

UT est introduit pour la première fois par Zellner (1976) dans le cas de la régression univariée. Singh (1988), Kelejian and Prucha (1985) sont d'autres références. Le modèle de régression univarié en IT est étudié dans Fraser (1979) et Kelejian and Prucha (1985). Dans le cas multivarié, le modèle UT est étudié dans Sutradhar and Ali (1986), avec une méthode d'estimation par moments. Le modèle multivarié IT se trouve par exemple dans Prucha and Kelejian (1984) et Katz and King (1999). Nous prouvons ensuite que l'estimateur du maximum de vraisemblance de la matrice de variance-covariance dans le modèle multivarié UT est asymptotiquement biaisé et proposons une version non biaisée. Pour le modèle informatique, nous proposons un algorithme des moindres carrés itératif repondéré pour calculer les estimateurs du maximum de vraisemblance. Nous présentons une étude de simulation visant à comparer le biais et l'erreur quadratique moyenne des estimateurs des coefficients de régression et de la matrice de variance-covariance selon plusieurs scénarios du processus de génération de données, bien spécifiés ou non. Nous proposons un outil graphique basé sur la distance de Mahalanobis pour guider le choix de l'utilisateur entre les modèles concurrents. Nous illustrons une application dans la finance sur une série de prix de clôture quotidiens des actions d'IBM et de MSFT au cours de la période 2007-2018. Nous illustrons une application dans la finance sur une série de prix de clôture quotidiens des actions d'IBM et de MSFT au cours de la période 2007-2018.

Après avoir clarifié la distribution multivariée de Student, nous présentons deux applications inspirées des articles suivants. Dans la littérature en économie politique, Katz and King (1999) combinent des transformations log-ratio classiquement utilisées dans l'analyse de composition avec un modèle de régression multivarié de type Student pour définir la distribution additive de Student. Ils montrent que ce modèle est supérieur aux modèles précédents pour les données de vote à plusieurs partis. Dans la littérature financière, plusieurs auteurs, Platen and Rendek (2008), Fung and Seneta (2010)), préconisent l'utilisation de la distribution de Student pour modéliser les rendements en journal des actifs financiers. Dans notre cas, nous comparons le modèle de régression gaussien et Student aux données électorales départementales françaises au niveau cantonal.

Des modèles spatiaux ont été envisagés pour étudier les résultats des élections. Kelejian and Prucha (2004) introduisent un estimateur de type variable instrumentale (IV) des paramètres d'un modèle spatial autorégressif du premier ordre avec perturbations autorégressives du premier ordre. Ensuite, via les techniques de Monte Carlo, Kelejian et al. (2004) donne de résultats de petits échantillonnage relatifs à l'estimateur suggéré, à l'estimateur de maximum de vraisemblance (ML) et à d'autres estimateurs IV pour le modèle univarié. Sutter (2005) examine la relation autorégressive spatiale entre les résultats du vote au niveau du comté lors de l'élection présidentielle de 2000. Mansley and Demšar (2015) explorent la variabilité géographique des relations entre le taux de participation à l'élection du maire de Londres et les variables sociodémographiques à un niveau spatial détaillé. Son analyse est abordée à travers une régression géographiquement pondérée (GWR), qui permet d'enquêter sur les variations locales des habitudes de vote. D'autre part, certains modèles spatiaux pour les données de composition ont été développés. Tjelmeland et Lund (2003) examinent un cadre Bayésien pour la ré-

gression de CODA, discutent des distributions antérieures appropriées et définissent des algorithmes efficaces de Monte Carlo utilisant une chaîne de Markov. Pirzamanbein (2015) construit un modèle hiérarchique pour les données de composition spatiale en utilisant un champ aléatoire markovien gaussien (GMRF) avec des observations de Dirichlet. Pawlowsky-Glahn et al. (2015) utilisent la transformation additive-logratio de compositions généralisées pour traiter la structure de covariance spatiale de manière géostatistique. Nous prévoyons d'étudier la possibilité d'introduire une autocorrélation spatiale dans les modèles de regression du point de vue de l'économétrie spatiale, c'est-à-dire en utilisant un modèle autorégressif spatial (LAG) simultané.

### Les données

Les données utilisées dans ma thèse proviennent du site Web de Cartelec et du site Web de l'Insee. Pour plus de détails, nous collectons sur le site Cartelec[5] les données de proportions de vote sur l'élection départementales françaises de 2015 concernant 95 départements en France et 207 cantons de la région Occitanie en France, et nous téléchargeons les données socio-économiques correspondantes (pour 2014) sur le site Web de l'INSEE[6]. De plus, deux bases de données ont été utilisées dans ma thèse: une au niveau départemental et une au niveau cantonal. Les deux bases de données contiennent:

1. Parts de vote: résultat des élections départementales françaises où le nombre de voix par parti est agrégé en trois grands partis, à gauche (L), à droite (R) et à l'extrême droite (XR).

2. Age: l'âge des électeurs. Il a trois composantes Age_1840 pour les personnes âgées de 18 à 40 ans, Age_4064 pour les personnes âgées de 40 à 65 ans et Age_65 pour les personnes âgées de plus de 65 ans.

3. Diplôme: le niveau d'éducation des électeurs. Il comporte également trois composantes: <BAC pour les personnes ayant au maximum une éducation secondaire, BAC pour les personnes ayant au moins une éducation secondaire et au plus un diplôme de lycée, et SUP pour les personnes ayant un diplôme universitaire.

4. Emploi en cinq catégories: AZ (agriculture, pêche), BE (industrie manufacturière, industries minières et autres), FZ (construction), GU (entreprises, transports et services) et OQ (administration publique, enseignement, santé humaine).

5. Taux de chômage: le taux de chômage (unemp).

6. Evolution de l'emploi: le taux de croissance annuel moyen de l'emploi (emp_evol).

7. Propriétaire: la proportion de personnes qui possèdent des actifs (propriétaire).

8. Revenu: la proportion de personnes qui paient l'impôt sur le revenu (revenu).

[5]https://www.data.gouv.fr/fr/datasets/elections-departementales-2015-resultats-par-bureaux-de-vote/

[6]https://www.insee.fr/fr/statistiques

9. Étranger: la proportion d'étrangers (étrangers).

### Structure de la thèse

La majeure partie de cette thèse est rédigée en anglais, excepté pour l'introduction, les résumés des quatre chapitres et la conclusion qui sont traduits en français. Cette thèse est divisée en cinq chapitres.

**Le chapitre 1** est une présentation générale de l'analyse des données de composition. Il présente les définitions de la composition et de la sous-composition, les principes de l'analyse des données de composition, certaines opérations telles que la perturbation et l'alimentation du simplexe. À cause de contraintes inhérentes aux données de composition, les méthodes statistiques classiques ne peuvent pas être utilisées directement dans le simplexe. Par conséquent, certaines transformations de log-ratio sont généralement appliquées. Le chapitre présente également les distributions normales pour les compositions aléatoires et le modèle de régression CODA.

**Le chapitre 2** illustre le modèle de régression de données compositionnelles avec la distribution normale dans le simplexe. Ce chapitre présente les modèles de régression CODA dans l'espace du simplexe et dans l'espace des coordonnées. Il illustre la relation entre les paramètres de ces modèles dans les deux espaces. Nous analysons comment les valeurs prédites dans ces modèles varient avec les prédicteurs. Nous proposons également de nouveaux outils graphiques pour explorer l'impact de certaines variables socio-économiques sur les résultats des élections avec les données au niveau départemental.

**Le chapitre 3** traite de la comparaison entre la distribution multivariée Normale et la distribution multivariée de Student et son application dans les données de retour des actifs financiers. Nous rappelons la distribution multivariée Normale et clarifions deux versions de la distribution multivariée de Student. Nous construisons une simulation pour comparer le biais, la matrice d'erreur quadratique moyenne et la covariance des estimateurs des coefficients de régression selon différents scénarios de processus de génération de données (DGP). Un test basé sur la distance de Mahalanobis est proposé pour sélectionner un modèle.

**Le chapitre 4** traite de la distribution multivariée de Student et de son application en économie politique. Certains chercheurs ont montré que les données électorales présentent souvent un comportement à queues lourdes. En utilisant le chapitre 3, nous proposons de remplacer la distribution gaussienne par la distribution de Student. Définissant ainsi la distribution de Student dans le simplexe. Nous appliquons le test développé au chapitre 3 pour choisir entre les modèles Normal et Student.

**Le chapitre 5** développe un modèle de régression spatiale des LAG dans le simplexe. Afin d'estimer ce modèle, nous adoptons un estimateur de variable instrumentale (IV) des paramètres d'un modèle autorégressif spatial dans un contexte multivarié, comme dans Kelejian and Prucha (2004). Nous présentons une simulation visant à comparer l'erreur quadratique moyenne relative (RRMSE) de l'estimation de paramètres dans le cadre de plusieurs processus de génération de données (DGP) entre les moindres carrés

à deux étages spatial (S2SLS) et les moindres carrés à trois étages spatial ( S3SLS). Un exemple d'économie politique est également illustré.

# Chapter 1

# General compositional data analysis


## Abstract

Compositional data are vectors with non negative components and whose sum is a constant. In practice, there are many areas which involve compositional data. In archaeology, the compositional analysis of raw materials (clays, lithic materials used to make stone tool, etc.) is used to understand the histories of trade and exchange in ancient economies. A challenge is how to identify or distinguish different places which are far from the original points. Besides, in sedimentology, specimens of sediments are separated into three components: sand, silt and clay. It is interesting to study the dependence of specimens of sediments on water depth. In household budget survey, people investigate how households spend their budget in housing, foodstuff, services and others. Is there any difference between men and women? In political economy, it is interesting to study vote shares of an electoral party. In France, the electoral system contains at least fifteen electoral parties. The above data are compositional data.

This chapter introduces some concepts, principles and operations of compositional data analysis. Because the classical statistical methods cannot be used directly for compositional data, log-ratio transformations are used and we present them here. We also recall the possible distributions on the simplex: Dirichlet and Normal distributions. Finally, we present the CODA general regression model, which may contain both classical explanatory variables and compositional explanatory variables.



## Résumé

Les données de composition sont des vecteurs de composantes non négatives et dont la somme est une constante. En pratique, de nombreux domaines impliquent des données de composition. En archéologie, l'analyse de la composition des matières premières (argiles, matériaux lithiques utilisés pour fabriquer des outils en pierre, etc.) est util-

isée pour comprendre l'histoire du commerce et des échanges dans les économies anciennes. Un défi consiste à identifier ou à distinguer différents endroits éloignés des points d'origine. Par ailleurs, en sédimentologie, les échantillons de sédiments sont séparés en trois composants: le sable, le limon et l'argile. Il est intéressant d'étudier la dépendance des échantillons de sédiments sur la profondeur de l'eau. Dans l'enquête sur le budget des ménages, les gens étudient comment leur budget est utilisé pour le logement, l'alimentation, les services, etc. Y a-t-il une différence entre les hommes et les femmes? En économie politique, il est intéressant d'étudier les parts de vote d'un parti électoral. En France, le système électoral comprend au moins quinze partis électoraux. Les données ci-dessus sont des données de composition.

Ce chapitre présente quelques concepts, principes et opérations utiles pour l'analyse de données compositionnelles. En outre, les transformations log-ratio sont présentées car les méthodes statistiques classiques ne peuvent pas être utilisées directement pour les données de composition. On introduit également les distributions possibles sur le simplexe: loi de Dirichlet et Logistique Normale. Enfin, nous présentons le modèle de régression général CODA, qui peut contenir à la fois des variables explicatives classiques et des variables explicatives de composition.

## 1.1 Concepts of compositional data analysis

### 1.1.1 Composition, subcomposition

A composition $\mathbf{x}$ is a vector of $D$ non-negative parts of some whole which carries relative information. A composition only representing a subset of the possible components is called a subcomposition.
Each composition $\mathbf{x}$ has a unique representer in the so-called simplex space $\mathbf{S}^D$ defined by:

$$\mathbf{S}^D = \{\mathbf{x} = (x_1, ..., x_D)' : x_d > 0, d = 1, ..., D; \sum_{d=1}^{D} x_d = 1\}$$

The simplex of $D = 3$ parts can be presented with a ternary diagram, where the three components are projected in barycentric coordinates. The simplex of $D = 4$ parts can be represented by a tetrahedron, where each possible 3-part subcomposition is represented on one side of the tetrahedron.

### 1.1.2 Principles of compositional data analysis

The statistical methods in compositional data analysis must satisfy the following principles (see Aitchison (2011), Pawlowsky-Glahn et al. (2015)):

#### Scale invariance

Any meaningful function $f$ of a compositional vector $x$ must be homogeneous of degree 0 (scale invariant). It means that if a composition is scaled by a constant, e.g. changing

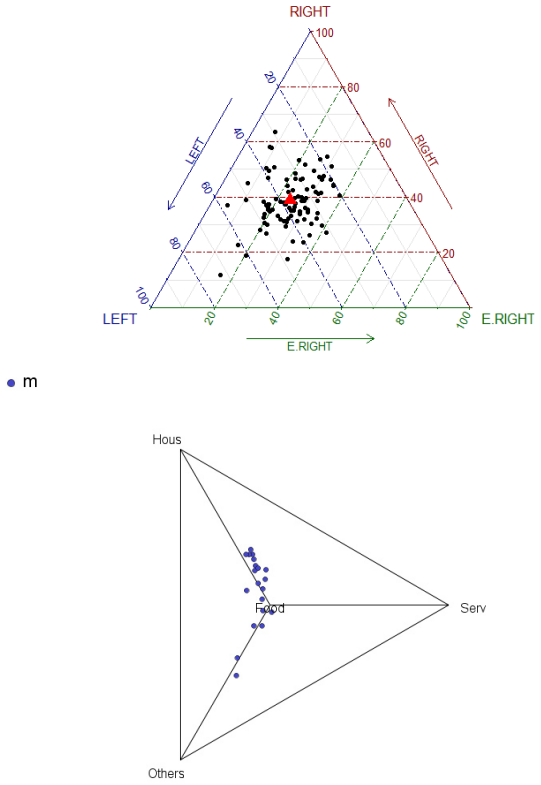


Figure 1.1: Vote shares (Left, Right, Extreme Right) in the French 95 departments (black points) with the geometric mean of vote shares as the red triangle on the left hand side. The tetrahedron describes the household expenditures on House, Food, Service and Others on the right hand side.

from parts per unit to percentages, the information carried is completely equivalent.

$$f(\lambda x) = f(x)$$

The traditional way to select a representative of the equivalence class is to normalize the vector in such a way that the components sum to a given constant $k$ by using the closure operation. For $\mathbf{x} = (x_1, \ldots, x_D)$, a vector of $D$ components, its closure is defined as

$$\mathcal{C}(\mathbf{x}) = \left( \frac{kx_1}{\sum_{d=1}^{D} x_d}, \ldots, \frac{kx_D}{\sum_{d=1}^{D} x_d} \right)$$

For example, the vectors $a = [2, 5, 10]$, $b = [2/7, 5/7, 10/7]$, and $c = [2/17, 5/17, 10/17]$ represent the same composition, the ratios between their components are the same. Therefore, any meaningful function of a compositional vector can be expressed in terms

of ratios of its components (any group invariant function can be expressed as a function of any maximal invariant)

**Subcompositional coherence**

This principle states that inferences about subcompositions should be coherent whether they are based on the subcomposition or the full composition. Analyses concerning a subset of parts must not depend on other non-involved parts. Subcompositional coherence contains two main things:
- The ratios between any parts in the subcomposition are equal to the corresponding ratios in the original composition.
- Distances between two compositions must be higher than, or equal to, the distance between their subcompositions.

**Permutation invariance**

Permutation invariance means that when we change the order of parts of a composition, it will give the same results.

### 1.1.3 Operations on the simplex

Aitchison (1986) introduced a set of operations in $\mathbf{S}^D$. For $\mathbf{x},\ \mathbf{y},\ \mathbf{z} \in \mathbf{S}^D$,

1. The perturbation of $\mathbf{x}$ and $\mathbf{y}$ is the compositional sum of $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, ..., x_D y_D)$$

where $\mathcal{C}(\mathbf{x})$ is the closure operation.
If $\mathbf{z} = \mathbf{x} \oplus \mathbf{y}, \mathbf{z} \in \mathbf{S}^D$, then $\frac{z_1/z_2}{x_1/x_2} = y_1/y_2$ represents the relative change in percentage between the first and second components.

2. The powering is the compositional scalar multiplication

$$\lambda \odot \mathbf{x} = \mathcal{C}(x_1^\lambda, ..., x_D^\lambda),\ \lambda \text{ is a scalar}, \mathbf{x} \in \mathbf{S}^D$$

For example, $\mathbf{x} \odot \mathbf{x} = \mathcal{C}(x_1^2, \ldots, x_D^2) = 2 \odot \mathbf{x}$

3. For $\mathbf{B} \in \mathbb{R}^{D \times D}$, the compositional matrix product $\mathbf{B} \boxdot \mathbf{x}$ corresponds to the matrix product in the real vector space

$$\mathbf{B} \boxdot \mathbf{x} = \mathcal{C}\left( \prod_{d=1}^{D} x_d^{b_{1d}}, \cdots, \prod_{d=1}^{D} x_d^{b_{Dd}} \right)^T$$

From the definition of perturbation and powering, we can deduct the Aitchison mean for $\mathbf{x_1},\ \ldots, \mathbf{x_n} \in \mathbf{S}^D$,

$$\frac{1}{n} \odot (\mathbf{x_1} \oplus \ldots \oplus \mathbf{x_n}) = \mathcal{C}((x_{11} \ldots x_{1n})^{1/n}, \ldots, (x_{D1} \ldots x_{Dn})^{1/n})$$

Therefore, the mean for compositions $\mathbf{z} \in \mathbf{S}^D$ is a vector of geometric means of each component

$$g(z_1, \ldots, z_n) = \sqrt[n]{z_1 \ldots z_n}$$

Besides, one can define the Aitchison geometry, i.e the compositional inner product and the compositional distance.

1. The compositional inner product (C-inner product) of $x$ and $y$ in $\mathbf{S}^D$ is defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle_c = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \log \frac{x_i}{x_j} \cdot \log \frac{y_i}{y_j} = \sum_{i=1}^{D} \log \frac{x_i}{g(\mathbf{x})} \cdot \log \frac{y_i}{g(\mathbf{y})}$$

   where $g(\mathbf{x}) = \sqrt[D]{x_1 x_2 ... x_D}$ is the geometric mean of the components of $\mathbf{x}$.

2. The compositional distance (C-distance) between $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{S}^D$ is derived from the inner product

$$\begin{aligned} \mathrm{d}_c(\mathbf{x}, \mathbf{y}) &= \|\mathbf{x} \ominus \mathbf{y}\|_c \\ &= \left( \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \left( \log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2 \right)^{1/2} \\ &= \left( \sum_{i=1}^{D} \left( \log \frac{x_i}{g(\mathbf{x})} - \log \frac{y_i}{g(\mathbf{y})} \right)^2 \right)^{1/2} \end{aligned}$$

   where $\|\mathbf{x}\|_c = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_c}$ denotes the compositional-norm (C-norm) of composition $\mathbf{x}$.

### 1.1.4 The log-ratio transformations

Classical statistical methods cannot be used directly in the simplex because the constraints that the components are positive and sum up to 1 are not compatible with their usual distributional assumptions. To overcome this difficulty, one way out is to use a log-ratio transformation from the simplex space $\mathbf{S}^D$ to the Euclidean space $\mathbb{R}^{D-1}$. The classical transformations are alr (additive log-ratio transformation), clr (centered log-ratio transformation), and ilr (isometric log-ratio transformation) (see Egozcue et al. (2012)). The coordinates in the clr transformed vector are linearly dependent, and the coordinates in the alr transformed vector are not compatible with the geometry (distance between the components in the simplex space is different from distance between the coordinates in the Euclidean space). For these reasons people generally use one of the ilr transformation.
The Additive Log-Ratio transformation (alr) is defined by

$$\mathrm{alr}(\mathbf{x}) = (\ln(x_1/x_D), \ldots, \ln(x_{D-1}/x_D))$$

The Centered Log-Ratio transformation (clr) is defined by

$$\text{clr}(\mathbf{x}) = \left(\ln\left(\frac{x_d}{g(\mathbf{x})}\right)\right) \quad \text{where} \quad g(\mathbf{x}) = \sqrt[D]{x_1, \ldots, x_D}, \ d = 1, \ldots, D$$

The additive log-ratio transformation is possibly questionable because the distances between points in the coordinate space are not the same for different reference levels $x_D$. A way to avoid the problem of choosing a reference level is to divide by the geometric mean, leading to the centered log-ratio transformation. However, the disadvantage of this is that the covariance matrix of clr coordinates is singular. To avoid the drawback of alr and clr transformations, an alternative set of transformations is known as ilr transformations.
The Isometric Log-Ratio transformation (ilr) is associated to a contrast matrix $\mathbf{V}_D$ is given by

$$\text{ilr}(\mathbf{x}) = \mathbf{x}^* = [\langle \mathbf{x}, \mathbf{e}_1 \rangle, \ldots \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle] = \mathbf{V}_D^T \ln(\mathbf{x})$$

where $\mathbf{x}$ is a column vecteur of $\mathbb{R}_+^D$ and the logarithm of $\mathbf{x}$ is understood componentwise, $\mathbf{V}_D^T$ is a $D \times (D-1)$ transposed contrast matrix associated to a given C-orthonormal basis $(\mathbf{e}_1, \ldots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by

$$\mathbf{V}_D = \text{clr}(\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}).$$

For the ilr transformation, we have the following properties:

1. $\text{ilr}(\mathbf{x} \oplus \mathbf{y}) = \text{ilr}(\mathbf{x}) + \text{ilr}(\mathbf{y}) = \mathbf{x}^* + \mathbf{y}^*$
2. $\text{ilr}(\alpha \odot \mathbf{x}) = \alpha \cdot \text{ilr}(\mathbf{x}) = \alpha \cdot \mathbf{x}^*$
3. $\langle \mathbf{x}, \mathbf{y} \rangle_c = \langle \text{ilr}(\mathbf{x}, \text{ilr}(\mathbf{y}) \rangle$
4. $\|\mathbf{x}\|_c = \|\text{ilr}(\mathbf{x})\|$
5. $\text{d}_c(\mathbf{x}, \mathbf{y}) = \text{d}(\text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{y}))$

**Sequential binary partition**

There are several ways to define C-orthonormal bases in the simplex. The main criterion to choose a C-orthonormal bases is to reinforce the interpretability of the representation in the ilr coordinates space. The C-orthonormal bases may be linked to a *sequential binary partition* (SBP) of the parts of the compositional vector.

A *sequential binary partition* is created with $D-1$ steps (see Pawlowsky-Glahn et al. (2015)). In a first step, SBP consists of dividing the composition into two groups of parts which are indicated by $+1$ and $-1$. In further steps, each previously obtained group of parts is repeatedly subdivided into two groups until all groups are made of a single part. Therefore, we get a $(D-1) \times D$ sign matrix $\mathbf{S} = (S_{dv}), \ d = 1, \ldots, (D-1), \ v = 1, \ldots, D$ with $+1$, $-1$, 0 (0 corresponds to parts which are not included in the partition). The

Table 1.1: Example of a sequential binary partition (SBP) of parts in $\mathbf{S}^3$ to build a C-orthonormal basis.

| Steps | $x_1$ | $x_2$ | $x_3$ | $\#^+$ | $\#^-$ |
|---|---|---|---|---|---|
| 1 | $+1$ | $-1$ | $-1$ | 1 | 2 |
| 2 | 0 | $+1$ | $-1$ | 1 | 1 |

Table 1.2: Example of the clr-transformed vectors of the C-orthonormal basis associated with the SBP.

| Steps | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | $+\frac{1}{1}\sqrt{\frac{1\cdot 2}{1+2}}$ | $-\frac{1}{2}\sqrt{\frac{1\cdot 2}{1+2}}$ | $-\frac{1}{2}\sqrt{\frac{1\cdot 2}{1+2}}$ |
| 2 | 0 | $+\frac{1}{1}\sqrt{\frac{1\cdot 1}{1+1}}$ | $-\frac{1}{1}\sqrt{\frac{1\cdot 1}{1+1}}$ |

matrix $\mathbf{S}$ is used to build the $(D-1)\times D$ matrix $\mathbf{\Phi}$ of the clr-transformed vectors of the C-orthonormal basis $\mathbf{e}_1,\ldots,\mathbf{e}_{D-1}$ associated with the sequential binary partition. The $\Phi_{dv}, d=1,\ldots,(D-1),\ v=1,\ldots,D$ entry of $\mathbf{\Phi}$ is defined by:

$$\begin{aligned}
&\Phi_{dv}=0 \text{ if } S_{dv}=0\\
&\Phi_{dv}=+\frac{1}{\#_d^+}\sqrt{\frac{\#_d^+\cdot\#_d^-}{\#_d^++\#_d^-}} \text{ if } S_{dv}>0\\
&\Phi_{dv}=-\frac{1}{\#_d^-}\sqrt{\frac{\#_d^+\cdot\#_d^-}{\#_d^++\#_d^-}} \text{ if } S_{dv}<0
\end{aligned} \tag{1.1}$$

where $\#_d^+$ and $\#_d^-$ are the number of parts in the $d^{\text{th}}$ raw of $\mathbf{S}$ coded by $+1$ and $-1$, respectively. Thus, $\text{clr}(\mathbf{e}_d)=\mathbf{\Phi}_d=[\Phi_{d1},\ldots,\Phi_{dD}]$. The $D\times(D-1)$ contrast matrix $\mathbf{V}$ is the transposed matrix $\mathbf{\Phi}$.

We illustrate the way to build the SBP for a compostition $\mathbf{x}\in\mathbf{S}^D,\ D=3$ in Table 1.1. We can build the SBP with two steps as in Table 1.1. After defining the SBP, we build the $(D-1)\times D$ matrix $\mathbf{\Phi}$ of the clr-transformed vectors of the C-orthonormal basis $\mathbf{e}_1,\ldots,\mathbf{e}_{D-1}$ associated with the sequential binary partition by using (1.1) as in Table 1.2

**An example of contrast matrix**

The following $D\times(D-1)$ matrix $\mathbf{V}_D$ defined by Egozcue et al (2003) [10] is an example of contrast matrix for $D=3$

$$\mathbf{V}_3=\begin{bmatrix}+2/\sqrt{6} & 0\\ -1/\sqrt{6} & +1/\sqrt{2}\\ -1/\sqrt{6} & -1/\sqrt{2}\end{bmatrix}$$

This particular matrix defines the following ilr coordinates

$$\text{ilr}_1(\mathbf{x}) = \frac{1}{\sqrt{6}}(2\ln x_1 - \ln x_2 - \ln x_3) = \frac{2}{\sqrt{6}}\ln\frac{x_1}{\sqrt{x_2 x_3}}$$
$$\text{ilr}_2(\mathbf{x}) = \frac{1}{\sqrt{2}}(\ln x_2 - \ln x_3) = \frac{1}{\sqrt{2}}\ln\frac{x_2}{x_3}$$

The first ilr coordinate contains information about the relative importance of the first component $x_1$ with respect to the geometric mean of the second and the third components $g = \sqrt{x_2 x_3}$. The second ilr coordinate contains information about the relative importance of the second component $x_2$ with respect to the third component $x_3$. The inverse ilr transformation is given by:

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^*) = \mathcal{C}(\exp(\mathbf{V}_D \mathbf{x}^{*T}))$$

where the exponential of vector $\mathbf{x}$ is understood componentwise. In fact, the contrast matrix is the transposed of matrix $\mathbf{\Phi}$ which is illustrated in Table 1.2.

**Properties of contrast matrices**

Note that such a contrast matrix $\mathbf{V}_D$ of size $D \times (D-1)$ satisfies the following properties:

1. $\mathbf{V}_D \mathbf{V}_D^T = \mathbf{G}_D = \mathbf{I}_D - \frac{1}{D}\mathbf{1}_{D\times D}$ where $\mathbf{I}_D$ ia a $D \times D$ identity matrix , $\mathbf{1}_{D\times D}$ is a $D \times D$ matrix of ones.

2. $\mathbf{V}_D^T \mathbf{V}_D = \mathbf{I}_{D-1}$ where $\mathbf{I}_{D-1}$ is the $(D-1) \times (D-1)$ identity matrix.

3. $\mathbf{V}_D^T \mathbf{j}_D = \mathbf{0}$ where $\mathbf{j}_D$ is a $D \times 1$ column vector of ones.

4. $\mathbf{j}_D^T \mathbf{V}_D = \mathbf{0}$ where $\mathbf{j}_D$ is a $D \times 1$ column vector of ones.

## 1.2 Distributions for random compositions

Distribution of simplex value random variables can be found in the literature. The main ones are Dirichlet, Aitchison, Logistic Normal and Student. Initially, people introduced the Dirichlet distribution as a distribution on the simplex. However, it is restrictive because of complete subcompositional independence. It implies the fact that all its subcompositions must be independent for each possible partition of the composition. Therefore, it is impossible to model any dependent structure for compositional data using the Dirichlet distribution. On the contrary, the Normal distribution on the simplex imply no constraint of complete subcompositional independence.

Table 1.3: Notations

| Variable | Notation | Coordinates |
|---|---|---|
| Dependent | $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iL})$ | $\text{ilr}(\mathbf{Y}_i) = \mathbf{Y}_i^*$ |
| Independent compositional | $\mathbf{X}_i^{(q)} = (X_{1i}^{(q)}, ..., X_{D_q i}^{(q)})$ | $\text{ilr}(\mathbf{X}_i^{(q)}) = \mathbf{X}_i^{(q)*}$ |
| Independent classical | $Z_i$ | |
| **General notations** | | |
| $L$ | Number of components of dependent variable | |
| $D_q$ | Number of components of covariate $\mathbf{X}^{(q)}$ | |
| $i, j = 1, \ldots, n$ | Index of observations | |
| $l, m = 1, \ldots, L$ | Index of components of compositional data | |
| $q = 1, \ldots, Q$ | Index of independent compositional variables | |
| $k = 1, \ldots, K$ | Index of independent classical variables | |

### 1.2.1 The normal distribution on the simplex

The additive logistic-normal distribution (also called Normal on the simplex) was introduced in Aitchison and Shen (1980). A random composition $\mathbf{x}$ follows the Normal distribution $\mathcal{N}_{\mathbf{S}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on the simplex $\mathbf{S}^D$ with the mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ if the coordinates $\mathbf{x}^* = \text{ilr}(\mathbf{x})$ of the random composition $\mathbf{x}$ follow the multivariate normal distribution with the density function

$$f^*(\mathbf{x}^*) = \frac{1}{(2\pi)^{(D-1)/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}^* - \boldsymbol{\mu}^*)\boldsymbol{\Sigma}^{*-1}(\mathbf{x}^* - \boldsymbol{\mu}^*)^t\right]$$

where $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ are the mean of the random ilr-coordinates and their covariance matrix, respectively. Therefore, if $\mathbf{x}$ follows a normal distribution on the simplex $\mathbf{S}^D$, then it is equivalent to say that $\text{ilr}(\mathbf{x})$ follows a normal distribution on $\mathbb{R}^{D-1}$. The logistic normal distribution can be estimated by the OLS method with the packages "compositions " and "robCompositions" in R.

## 1.3 The CODA regression models

### 1.3.1 Notations

In this work, we will use the notations defined in Table 1.3.

### 1.3.2 The CODA regression models: expression in the simplex space and in the coordinates space.

Let $\mathbf{Y}_i \in \mathbf{S}^L$ denote the compositional response value of the $i$th observation, and $\mathbf{X}_i^{(q)} \in \mathbf{S}^{D_q}$, $q = 1, \ldots, Q$, denote the value of the $q$th compositional covariate for the $i$th observation, $Z_{ik}$, $k = 1, \ldots, K$ denote the $k$th classical covariate of the $i$th observation, $\mathbf{B} = (b_{ld})$, $l = 1, \ldots, L$, $d = 1, \ldots, D$, is a parameter matrix such that $\mathbf{j}_L^T\mathbf{B} = \mathbf{0}_D$,

$\mathbf{B}\mathbf{j}_D = \mathbf{0}_L$ (see Kynčlová et al. (2015)), where $\mathbf{j}_L$ is a $L \times 1$ column vector of ones, and $\mathbf{j}_L^T$ is the transposed of $\mathbf{j}_L$.

### The CODA regression model in the simplex

The linear CODA regression model in the simplex is defined by

$$\mathbf{Y}_i = \mathbf{b_0} \bigoplus_{q=1}^{Q} \mathbf{B}^{(q)} \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_{ki} \odot \mathbf{c}_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{1.2}$$

where $\mathbf{b_0}, \mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(Q)}, \mathbf{c}_1, \ldots, \mathbf{c}_K$ are the parameters. The distributional assumption is that $\boldsymbol{\epsilon}_i \in \mathbf{S}^L$ follows the multivariate normal distribution on the simplex.

### The CODA regression model in the ilr coordinate space

The CODA regression model in the ilr coordinate space is defined by

$$\text{ilr}(\mathbf{Y}_i) = \mathbf{b_0}^* + \sum_{q=1}^{Q} \text{ilr}(\mathbf{X}_i^{(q)}) \mathbf{B}^{*(q)} + \sum_{k=1}^{K} Z_{ki} \mathbf{c}_k^* + \text{ilr}(\boldsymbol{\epsilon}_i) \tag{1.3}$$

where $\text{ilr}(\mathbf{Y_i})$, $\text{ilr}(\mathbf{X}_i^{(q)})$ are the ilr coordinates of $\mathbf{Y}_i$, $\mathbf{X}_i^{(q)}$ $(q = 1, \ldots, Q)$ respectively; $\mathbf{b_0}^*$, $\mathbf{B}_q^*$, $\mathbf{c}_k^*$ are the parameters , and $\text{ilr}(\boldsymbol{\epsilon}_i)$ are the residuals on the ilr coordinate space. The distributional assumption is that $\text{ilr}(\boldsymbol{\epsilon})$ follows the multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}^*$.

Thereby, if a random composition $\mathbf{x}$ follows a Normal distribution on the simplex $\mathbf{S}^D$, then it is equivalent to say that $\mathbf{x}^*$ follows a Normal distribution on $\mathbb{R}^{D-1}$.

# Chapter 2

# Abstract

Many papers focus on studying the two-party electoral system. However, more and more interest is given to the exploration of the multiparty system nowadays. There are at least fifteen political parties in France. In order to illustrate our method, we concentrate on the outcome of the departmental election in France and we aggregate the outcome of the election into three large groups of parties which are Left, Right, and Extreme Right. The proportions of votes for these three parties and for each department form a vector called composition (mathematically, a vector belonging to a simplex). Therefore, the electoral outcome data in the 2015 French departmental election has the following properties: the vectors have positive components which sum up to 1 for each statistical unit (department here). According to the definition of compositions, it is thus equivalent to say that the vectors of proportions of votes by party on a given department are compositional data. The objective of this chapter is to use CODA regression models to generalize political economy models to more than two parties and to study the impact of the characteristics of the territorial units on the outcome of the election. The models are fitted on French electoral data of the 2015 departmental elections with departmental level data. In the political economy literature, such regression models are generally restricted to the case of two political parties. In the statistical literature, there are regression models adapted to share vectors including CODA models (for COmpositional Data Analysis), but also Dirichlet models, Student models and others. In this chapter, we propose a CODA regression model for analyzing how the geographic electoral result depends on the socio-economic characteristic of the departments in France. This model will contain both compositional explanatory variables and classical explanatory variables. Besides, the classical statistical methods cannot be used directly for compositional data. Some log-ratio transformations will be carried out. We use the ilr transformation in the coordinate space in this thesis. We first build a CODA regression model in coordinate space and then rewrite this model in the simplex. We apply the maximum likelihood estimation to fit this model. However, the interpretation of parameters in regression model is quite complex. We show how to use instead the vote share predictions to understand the impact of socio-economic factors using some graphical techniques.

# Résumé

De nombreux articles portent sur l'étude du système électoral à deux partis. Cependant, on s'intéresse de plus en plus à l'exploration du système multipartite. Il y a au moins quinze partis politiques en France. Afin d'illustrer notre méthode, nous nous concentrons sur le résultat de l'élection départementale en France et nous agrégons le résultat de l'élection en trois grands groupes de partis qui sont la gauche, la droite et l'extrême droite. Les proportions de votes pour ces trois partis et pour chaque département forment un vecteur appelé composition (mathématiquement, un vecteur appartenant à un simplex). Par conséquent, les données sur les résultats des élections de 2015 dans les départements français ont les propriétés suivantes: les vecteurs ont des composantes positives qui totalisent 1 pour chaque unité statistique (département ici). Selon la définition des compositions, il est donc équivalent de dire que les vecteurs de proportions de votes par parti sur un département donné sont des données de composition. L'objectif de ce chapitre est d'utiliser les modèles de régression CODA pour généraliser les modèles d'économie politique à plus de deux partis et d'étudier l'impact des caractéristiques des unités territoriales sur le résultat des élections. Les modèles sont adaptés aux données électorales françaises des élections départementales de 2015 avec des données au niveau départemental. Dans la littérature sur l'économie politique, de tels modèles de régression sont généralement limités au cas de deux partis politiques. Dans la littérature statistique, il existe des modèles de régression adaptés au partage de vecteurs, notamment les modèles CODA (pour l'analyse des données de substitution), mais également les modèles de Dirichlet, les modèles de Student et autres. Dans ce chapitre, nous proposons un modèle de régression CODA pour analyser la manière dont le résultat électoral géographique dépend des caractéristiques socio-économiques des départements en France. Ce modèle contiendra à la fois des variables explicatives de composition et des variables explicatives classiques. De plus, les méthodes statistiques classiques ne peuvent pas être utilisées directement pour les données de composition. Certaines transformations de log-ratio seront effectuées. Nous utilisons la transformation ilr dans l'espace de coordonnées dans cette thèse. Nous construisons d'abord un modèle de régression CODA dans un espace de coordonnées, puis nous réécrivons ce modèle dans le simplexe. Nous appliquons l'estimation du maximum de vraisemblance pour correspondre à ce modèle. Cependant, l'interprétation des paramètres dans le modèle de régression est assez complexe. Nous montrons comment utiliser plutôt les prévisions de partage des voix pour comprendre l'impact des facteurs socio-économiques à l'aide de certaines techniques graphiques.

# Analyzing the impacts of socio-economic factors on French departmental elections with CODA methods[1][2]

**Abstract**
The proportions of votes by party on a given subdivision of a territory form a vector called composition (mathematically, a vector belonging to a simplex). It is interesting to model these proportions and study the impact of the characteristics of the territorial units on the outcome of the elections. In the political economy literature, such regression models are generally restricted to the case of two political parties. In the statistical literature, there are regression models adapted to share vectors including CODA models (for COmpositional Data Analysis), but also Dirichlet models, Student models and others. Our goal is to use CODA regression models to generalize political economy models to more than two parties. The models are fitted on French electoral data of the 2015 departmental elections.

**Keywords**
political economy, compositional regression models, vote shares, departmental election, Gaussian distribution

**Résumé**
Les proportions de votes par parti forment un vecteur de données dites de composition (mathématiquement, un vecteur appartenant à un simplexe) sur une subdivision de territoire. Il est intéressant de modéliser ces proportions en étudiant l'impact des caractéristiques des unités territoriales sur l'issue des élections. Dans la littérature d'économie politique, il existe des modèles de régression qui sont restreints généralement au cas de deux partis politiques. Dans la littérature statistique, il existe des modèles de régression adaptés à des vecteurs de parts dont les modèles CODA (pour "COmpositional Data Analysis"), mais aussi les modèles de Dirichlet et d'autres. Notre objectif est d'utiliser les

[1] Submitted to Journal of Applied Statistics
[2] Joint work with T. Laurent, C. Thomas-Agnan and A. Ruiz-Gazen

modèles de régression de type CODA pour généraliser les modèles d'économie politique à plus de deux partis. Les modèles sont ajustés sur des données électorales françaises des élections départementales de 2015.

**Mots-clés.**
économie politique, modéles de régression pour données de composition, proportions de vote, élection départmental.

## 2.1 Introduction

Recently, models for elections focus on analyzing impacts of socio-economic factors for two-party systems using classical regression models( see Lewis and Linzer (2005)). In this paper, we propose a statistical model for studying the multiparty system using compositional data analysis (CODA) with departmental level data. The dependent variable will be the vector of votes shares for the French departmental election in 2015. The explanatory variables include some compositional and continuous socio-economic variables.

Among papers concentrating on the relationship between socio-economic variables and election results, Beauguitte and Colange (2013) study a linear regression at three levels of aggregation (polling stations, cities and electoral districts) and show that the socio-economic variables are significant.Kavanagh et al. (2006) use geographically weighted regression, which produces parameter estimates for each data point, i.e. for each electoral division. On the other hand in the statistical literature, people have developped CODA regression models where the dependent and independent variables may be compositional variables (see Mert et al. (2018) for a review). Morais (2017) study the impact of media investments on brand's market shares with a CODA regression model. Trinh and Morais (2017) use a CODA regression model to highlight the nutrition transition and to explain it according to household characteristics. Honaker et al. (2002), Katz and King (1999) use a statistical model for multiparty electoral data assuming that the territorial units yield independent observations.

Vote share data of the 2015 French departmental election for 95 departments in France are collected from the Cartelec website [3] and corresponding socio-economic data (for 2014) have been downloaded from the INSEE website [4]. Table 2.1 summarizes our data set. In Section 2.2, we present the data set. Subsection 2.3.1 (resp: 2.3.2) recalls the principles of compositional data analysis (resp: of compositional regression models). In subsection 2.3.3, we implement the CODA model on the election data set and present several plots to explore the impact of explanatory variables of a classical type illustrated by the case of unemployment rate as well as variables of a compositional type illustrated by the diploma variable.

[3] https://www.data.gouv.fr/fr/datasets/elections-departementales-2015-resultats-par-bureaux-de-vote/

[4] https://www.insee.fr/fr/statistiques

Table 2.1: Description for departmental level data

| Variable name | Description | Averages |
|---|---|---|
| Vote share | Left(L), Right(R), Extreme Right(XR) | 0.370, 0.388, 0.242 |
| Age | Age_1840, Age_4064, Age_65. | 0.313, 0.432, 0.255 |
| Diploma | <BAC, BAC, SUP. | 0.591, 0.16, 0.239 |
| Employment | AZ, BE, FZ, GU, OQ | 0.031, 0.099, 0.049, 0.439, 0.382 |
| unemp | The unemployment rate | 0.117 |
| employ_evol | Mean annual growth rate of employment (2009-2014) | -0.145 |
| owner | The proportion of people who own assets | 0.616 |
| income | The proportion of people who pay income tax | 0.552 |
| foreign | The proportion of foreigners | 0.050 |

## 2.2 Data

The database in this chapter contains the vote shares and all socio-economics charateristics for 95 departments in France. These data are illustrated on Table 2.2. Employment has five categories: AZ (agriculture, fisheries), BE (manufacturing industry, mining industry and others), FZ (construction), GU (business, transport and services) and OQ (public administration, teaching, human health). Diploma has three levels: <BAC for people with at most some secondary education, BAC for people with at least some secondary education and at most a high school diploma, and SUP for people with a university diploma. The Age variable has three levels: Age_1840 for people from 18 to 40 years old, Age_4064 for people from 40 to 64 years old, and Age_65 for elderly. For the vote share variable, the Cartelec website provides a very detailed information. The list of political parties which present candidates at that election is higher than 15. However, at the end of the election, it is common to present the results by grouping the political parties into three main components : Left, Right and Extreme-Right[5]. The third column in Table 2.1 indicates the geometric means for compositional variables and the averages for classical variables.

From the CODA point of view, when compositional data have three components, they can be represented in a ternary diagram. For instance, the vote shares of the 95 departments for the Left and Right wings and the Extreme Right party are the blue points in Figure 2.1. The red triangle corresponding to the Aube department on Figure 2.1 shows that its vote shares of the Left wing, the Right wing and the Extreme Right party are respectively 17.4%, 54.6%, and 28%. Figure 2.2 illustrates the positions of the French departments on the ternary diagram whose components correspond to the three levels of the diploma variable, and the red triangle figures the geometric mean (adapted mean for compositional data) of all departments.

[5]for more details, see https://fr.wikipedia.org/wiki/Elections_départementales_francaises_de_2015

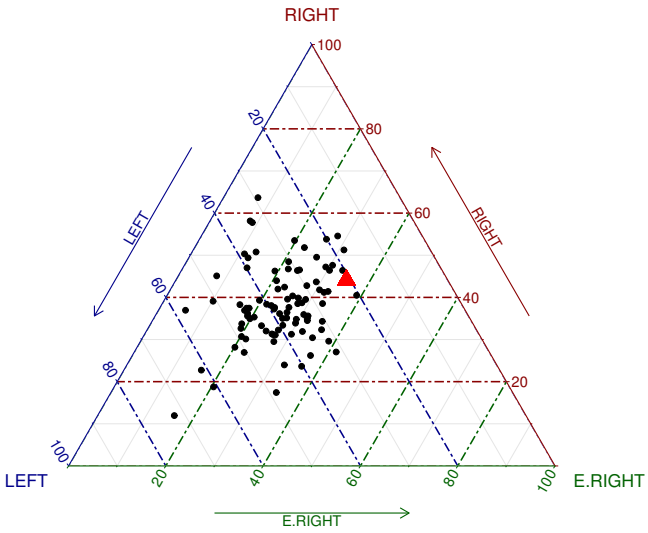


Figure 2.1: Vote shares in the 95 departments (blue points) with the Aube department as the red triangle

## 2.3 Compositional data analysis approach

In order to analyze the impacts of the socio-economic factors on the election results, a CODA regression model is proposed where the dependent variable is a compositional variable (vote shares) and the independent variables are compositional or classical variables or a mixture of both. This model is based on the log-ratio transformation approach.

### 2.3.1 Principles of compositional data analysis

A composition $\mathbf{x}$ is a vector of $D$ parts of some whole which carries relative information. A $D$-composition $\mathbf{x}$ lies in the so-called simplex space $\mathbf{S}^D$ defined by:

$$\mathbf{S}^D = \{\mathbf{x} = (x_1, ..., x_D)' : x_j > 0, j = 1, ..., D; \sum_{j=1}^{D} x_j = 1\}$$

The simplex $\mathbf{S}^D$ can be equipped with the Aitchison inner product (Aitchison (1985) and Pawlowsky-Glahn et al. (2015)) in order to define distances. Classical regression models cannot be used directly in the simplex because the constraints that the components are positive and sum up to 1 are not compatible with their usual distributional assumptions. To overcome this difficulty, one way out is to use a log-ratio transformation from the simplex space $\mathbf{S}^D$ to the Euclidean space $\mathbb{R}^{D-1}$. The classical transformations are alr (additive log-ratio transformation), clr (centered log-ratio transformation), and ilr (isometric log-ratio transformation). The coordinates in the clr transformed vector are

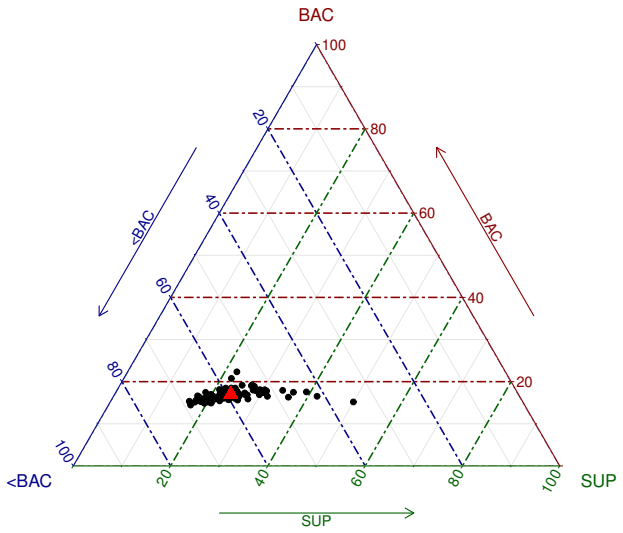


Figure 2.2: Components of Diploma in the 95 departments (blue points) and their geometric mean (red triangle)

linearly dependent, and the coordinates in the alr transformed vector are not compatible with the geometry (distance between the components in the simplex space is different from distance between the coordinates in the Euclidean space). For these reasons people generally use one of the ilr transformation for compositional regression models.
An isometric log-ratio transformation (ilr) is defined by:

$$\text{ilr}(\mathbf{x}) = \mathbf{V}_D^T \ln(\mathbf{x})$$

where the logarithm of $\mathbf{x}$ is understood componentwise, $\mathbf{V}_D^T$ is a transposed contrast matrix Pawlowsky-Glahn et al. (2015) associated to a given orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by

$$\mathbf{V}_D = \text{clr}(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1}).$$

Note that such a contrast matrix $\mathbf{V}_D$ of size $D \times (D-1)$ satisfies the following property:

1. $\mathbf{V}_D \mathbf{V}_D^T = \mathbf{I}_D - \frac{1}{D}\mathbf{1}_{D\times D}$ where $\mathbf{I}_D$ is the $D \times D$ identity matrix, $\mathbf{1}_{D\times D}$ is a $D \times D$ matrix of ones.

2. $\mathbf{V}_D^T \mathbf{V}_D = \mathbf{I}_{D-1}$ where $\mathbf{I}_{D-1}$ is the identity matrix with dimension $(D-1)$.

3. $\mathbf{V}_D^T \mathbf{j}_D = \mathbf{0}_{D-1}$ where $\mathbf{j}_D$ is a $D \times 1$ column vectors of ones.

The following $D \times (D-1)$ matrix $\mathbf{V}_D$ defined by Egozcue et al. (2003) is an example of contrast matrix for $D = 3$

$$\mathbf{V}_3 = \begin{bmatrix} 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}$$

This particular matrix defines the following ilr coordinates

$$\begin{aligned} \text{ilr}_1(\mathbf{x}) &= \frac{1}{\sqrt{6}}(2\log x_1 - \ln x_2 - \log x_3) = \frac{2}{\sqrt{6}}\log\frac{x_1}{\sqrt{x_2 x_3}} \\ \text{ilr}_2(\mathbf{x}) &= \frac{1}{\sqrt{2}}(\log x_2 - \log x_3) = \frac{1}{\sqrt{2}}\log\frac{x_2}{x_3} \end{aligned}$$

The first ilr coordinate contains information about the relative importance of the first component $x_1$ with respect to the geometric mean of the second and the third components $g = \sqrt{x_2 x_3}$. The second ilr coordinate contains information about the relative importance of the second component $x_2$ with respect to the third component $x_3$. In our case, the first ilr coordinate opposes the Left wing to the group of the Right wing and the Extreme Right party and the second opposes the Right wing to the Extreme Right party. The inverse ilr transformation is given by:

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^*) = \mathcal{C}(\exp(V_D\mathbf{x}^*)) \text{ for } x^* \in R^{D-1}$$

where the exponential of vector $\mathbf{x}$ is understood componentwise and

$$\mathcal{C}(\mathbf{x}) = \left( x_1/\sum_{j=1}^{D} x_j, \cdots, x_D/\sum_{j=1}^{D} x_j \right) \text{ is the closure operation.}$$

The vector space structure of the simplex $\mathbf{S}^D$ is defined by the perturbation and powering operations:

$$\begin{aligned} &\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D),\ \mathbf{x},\ \mathbf{y} \in \mathbf{S}^D \\ &\lambda \odot \mathbf{x} = \mathcal{C}(x_1^\lambda, \ldots, x_D^\lambda),\ \lambda \text{ is a scalar}, \mathbf{x} \in \mathbf{S}^D. \end{aligned}$$

The compositional inner product (C-inner product) of $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{S}^D$ is defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle_c = \frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D} \log\frac{x_i}{x_j} \cdot \log\frac{y_i}{y_j} = \sum_{i=1}^{D} \log\frac{x_i}{g(\mathbf{x})} \cdot \log\frac{y_i}{g(\mathbf{y})}$$

where $g(\mathbf{x}) = \sqrt[D]{x_1 x_2 \ldots x_D}$ is the geometric mean of the components.
The compositional distance (C−distance) between $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{S}^D$ is derived from the inner product

$$\begin{aligned} \text{d}_c(\mathbf{x}, \mathbf{y}) &= \left( \frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D} \left( \log\frac{x_i}{x_j} - \log\frac{y_i}{y_j} \right)^2 \right)^{1/2} \\ &= \left( \sum_{i=1}^{D} \left( \log\frac{x_i}{g(\mathbf{x})} - \log\frac{y_i}{g(\mathbf{y})} \right)^2 \right)^{1/2} \end{aligned}$$

Table 2.2: Notations.

| **Variable** | **Notation** | **Coordinates** |
|---|---|---|
| Dependent | $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iL})$ | $\text{ilr}(\mathbf{Y}_i) = \mathbf{Y}_i^*$ |
| Compositional explanatory | $\mathbf{X}_i^{(q)} = (X_{i1}^{(q)}, \ldots, X_{iD_q}^{(q)})$ | $\text{ilr}(\mathbf{X}_{ip}^{(q)}) = \mathbf{X}_{ip}^{(q)*}$ |
| Classical explanatory | $Z_{ki}$ | |
| **General notations** | | |
| $L$ | Number of components of the dependent variable | |
| $i = 1, \ldots, n$ | Index of observations ($n = 95$ ) | |
| $q = 1, \ldots, Q$ | Index of compositional explanatory variables ($Q = 3$) | |
| $p = 1, \ldots, D_q$ | Index of the coordinates for the compositional explanatory variables | |
| $k = 1, \ldots, K$ | Index of classical explanatory variables ($K = 5$) | |

The expected value $\mathbb{E}^{\oplus}\mathbf{Y}$ of a simplex-valued random composition $\mathbf{Y} \in \mathbf{S}^D$ (see Pawlowsky-Glahn et al. (2015)) is defined by

$$\text{argmin}_{\mathbf{z}\in\mathbf{S}^D}\mathbb{E}(d_c^2(\mathbf{Y}, \mathbf{z}))$$

and it is equal to

$$\mathbb{E}^{\oplus}\mathbf{Y} = \mathcal{C}(\exp(\mathbb{E}\log\mathbf{Y})) = \text{clr}^{-1}(\mathbb{E}\,\text{clr}(\mathbf{Y})) = \text{ilr}^{-1}(\mathbb{E}\,\text{ilr}(\mathbf{Y})) = \text{ilr}^{-1}(\mathbb{E}\mathbf{Y}^*)$$

where $\mathbf{Y}^* = \text{ilr}(\mathbf{Y})$.

### 2.3.2 Compositional regression models

The notations used in this paper are summarized in Table 2.2

We now describe the CODA regression model. $\mathbf{Y}_i \in \mathbf{S}^L$ denotes the compositional response value of the $i$th observation, and $\mathbf{X}_i^{(q)} \in \mathbf{S}^{D_q}$, $q = 1, \ldots, Q$, denotes the value of the $q$th compositional covariate for the $i$th observation, where $\mathbf{Y} \in \mathbf{S}^L$ and $\mathbf{X}^{(q)} \in \mathbf{S}^{D_q}$, $q = 1, \ldots, Q$, $Z_{ki}$, $k = 1, \ldots, K$, denotes the $k$th classical covariate of the $i$th observation. Let $\boxdot$ be the compositional matrix product, which corresponds to the matrix product in the coordinate space through the ilr transformation

$$\mathbf{B} \boxdot \mathbf{x} = \mathcal{C}\left(\prod_{j=1}^{D} x_j^{b_{1j}}, \cdots, \prod_{j=1}^{D} x_j^{b_{Lj}}\right)^T$$

where $\mathbf{x} \in \mathbf{S}^D$ and $\mathbf{B} = ((b_{ij}))$, $i = 1, \ldots, L$, $j = 1, \ldots, D$, is a parameter matrix such that the column vectors belong to $\mathbf{S}^D$, $\mathbf{j}_L^T\mathbf{B} = \mathbf{0}_D$, $\mathbf{B}\mathbf{j}_D = \mathbf{0}_L$, where $\mathbf{j}_L$ is a $L \times 1$ column vector of ones, and $\mathbf{j}_L^T$ is the transposed of $\mathbf{j}_L$.

Let us first introduce the CODA regression model in the ilr coordinate space as follows:

$$\text{ilr}(\mathbf{Y}_i) = \mathbf{b_0}^* + \sum_{q=1}^{Q} \text{ilr}(\mathbf{X}_i^{(q)})\mathbf{B}_q^* + \sum_{k=1}^{K} Z_{ki}\mathbf{c}_k^* + \text{ilr}(\boldsymbol{\epsilon}_i) \tag{2.1}$$

where $\text{ilr}(\mathbf{Y_i})$, $\text{ilr}(\mathbf{X}_i^{(q)})$ are the ilr coordinates of $\mathbf{Y}_i$, $\mathbf{X}_i^{(q)}$ $(q = 1, \ldots, Q)$ respectively; $\mathbf{b_0}^*$, $\mathbf{B}_q^*$, $\mathbf{c}_k^*$ are the parameters in the coordinate space, and $\text{ilr}(\boldsymbol{\epsilon}_i)$ are the residuals. The distributional assumption is that $\text{ilr}(\boldsymbol{\epsilon})$ follows the multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$.
This regression model can be written in the simplex as

$$\mathbf{Y}_i = \mathbf{b_0} \bigoplus_{q=1}^{Q} \mathbf{B}_q \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_{ki} \odot \mathbf{c}_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{2.2}$$

where $\mathbf{b_0}, \mathbf{B}_1, \ldots, \mathbf{B}_Q, \mathbf{c}_1, \ldots, \mathbf{c}_K$ are the parameters satisfying $\mathbf{b_0} \in \mathbf{S}^L$, $\mathbf{B}_q \in \mathbf{S}^{D_q}$, $q = 1, \ldots, Q$, $\mathbf{c}_k \in \mathbf{S}^L$, $k = 1, \ldots, K$, $\mathbf{j}_L^T \mathbf{B}_q = \mathbf{0}_{D_q}$, $\mathbf{B}_q \mathbf{j}_{D_q} = \mathbf{0}_L$, The distributional assumption is that $\boldsymbol{\epsilon}_i \in \mathbf{S}^L$ follows the normal distribution on the simplex (see Aitchison (1985)).

It is classical to estimate model (2.1) using OLS thus assuming the independence between the ilr coordinates. Chen et al. (2017) give different formulas relating the parameters in the simplex to the parameters in the coordinate space. Following Chen et al. (2017) (Property 2.1 and Property 2.3(3)), and Van den Boogaart and Tolosana-Delgado (2013), in model (2.1)-(2.2), the relationship between the parameters in the simplex and their counterpart in coordinate space is given by

$$\begin{cases} \mathbf{b_0} = \exp(\mathbf{V}_L \mathbf{b}_0^*) = \text{ilr}^{-1}(\mathbf{b}_0^*) \\ \mathbf{B}_q = \mathbf{V}_L \mathbf{B}_q^* \mathbf{V}_{D_q}^T \\ \mathbf{c}_k = \exp(\mathbf{V}_L \mathbf{c}_k^*) = \text{ilr}^{-1}(\mathbf{c}_k^*) \end{cases} \tag{2.3}$$

where $\mathbf{V}_L$ and $\mathbf{V}_{D_q}$, $q = 1, \ldots, Q$ are contrast matrices associated to the selected ilr transformations.

### 2.3.3 Impact of compositional and classical explanatory variables

Because the interpretation of the parameters of these models is not so straightforward (see Morais et al. (2017)), we rather concentrate on illustrating graphically the relationship between the predicted vote shares and the explanatory variables. The prediction of the dependent variable for the above models are given by:

$$\hat{\mathbf{Y}}_i = \hat{\mathbf{b}}_\mathbf{0} \bigoplus_{q=1}^{Q} \hat{\mathbf{B}}_q \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_{ki} \odot \hat{\mathbf{c}}_k \quad i = 1, ..., n \tag{2.4}$$

where $\hat{\mathbf{b}}_{\mathbf{0}}, \hat{\mathbf{B}}_q$ and $\hat{\mathbf{c}}_k$ are the estimated parameters. We can rewrite (2.4) as

$$\hat{\mathbf{Y}}_i = \mathcal{C}\left(\hat{\mathbf{b}}_{\mathbf{0}} \cdot \left(\prod_{q=1}^{Q} \mathbf{X}_{\mathbf{i}}^{(\mathbf{q})^{\hat{\mathbf{B}}_q}}\right) \cdot \left(\prod_{k=1}^{K} \hat{\mathbf{c}}_k^{Z_{ki}}\right)\right) \quad i = 1, ..., n \tag{2.5}$$

In order to illustrate these formulas, we will focus on graphing the predicted values of the dependent variable as a function of one specific variable of interest: two cases must be considered depending on whether the specific variable is classical or compositional. In both cases, we will create a grid of potential values of the specific explanatory and fix the other explanatory variables at the values they take for one selected point of the dataset (we repeat for several selected points). For the sake of simplicity let us take $L = 3$.

For the case when the specific variable is a classical covariate $Z_{ki}$, from (2.5) there exists $\hat{\mathbf{b}}_0 \in \mathbf{S}^L$ (this term contains the impacts of all other explanatory but is constant when $Z_{ki}$ alone varies) such that

$$\hat{\mathbf{Y}}_i = \hat{\mathbf{b}}_{\mathbf{0}} \bigoplus Z_{ki} \odot \hat{\mathbf{c}} = \mathcal{C}\left(\hat{b}_{01}\hat{c}_1^{Z_{ki}}, \cdots, \hat{b}_{0L}\hat{c}_L^{Z_{ki}}\right)$$

With $T = \hat{b}_{01}\hat{c}_1^{Z_{ki}} + \cdots + \hat{b}_{0L}\hat{c}_L^{Z_{ki}}$, we get

$$\hat{Y}_{i1} = \frac{\hat{b}_{01}\hat{c}_1^{Z_{ki}}}{T};\ \hat{Y}_{i2} = \frac{\hat{b}_{02}\hat{c}_2^{Z_{ki}}}{T}; \cdots ; \hat{Y}_{iL} = \frac{\hat{b}_{0L}\hat{c}_L^{Z_{ki}}}{T}.$$

Now for the case when the specific variable is a compositional variable $\mathbf{X}_{\mathbf{i}}^{(\mathbf{q})}$, let us take for the sake of simplicity $D_q = 3$. As before, from (2.5), there exists $\hat{\mathbf{b}}_{\mathbf{0}} \in \mathbf{S}^L$ (this term contains the impacts of all other explanatory but is constant when $X_i^{(q)}$ alone varies) such that

$$\begin{aligned}
\hat{\mathbf{Y}}_i &= \hat{\mathbf{b}}_{\mathbf{0}} \bigoplus \mathbf{X}_{\mathbf{i}}^{(\mathbf{q})^{\hat{\mathbf{B}}_q}} \\
&= \mathcal{C}(\hat{b}_{01} {X_{i1}^{(q)}}^{\hat{b}_{11}^{(q)}} {X_{i2}^{(q)}}^{\hat{b}_{12}^{(q)}} {X_{i3}^{(q)}}^{\hat{b}_{13}^{(q)}}, \hat{b}_{02} {X_{i1}^{(q)}}^{\hat{b}_{21}^{(q)}} {X_{i2}^{(q)}}^{\hat{b}_{22}^{(q)}} {X_{i3}^{(q)}}^{\hat{b}_{23}^{(q)}}, \\
&\quad \hat{b}_{03} {X_{i1}^{(q)}}^{\hat{b}_{31}^{(q)}} {X_{i2}^{(q)}}^{\hat{b}_{32}^{(q)}} {X_{i3}^{(q)}}^{\hat{b}_{33}^{(q)}})
\end{aligned}$$

We now fit a CODA regression model describing the impacts of socio-economic factors on vote shares in the 2015 French departmental election.
After including all explanatory variables from our data set in the regression model, and eliminating one by one the variables which are not significant, we obtain the results in Table 2.3. This model shows that the age of people, the proportion of people who own assets, the proportion of foreigners do not have any impact on the vote shares. However, the levels of education, the working areas, the unemployment rate and the proportion of people who pay income tax really affect the result of the French departmental election in 2015.

Table 2.3: Regression with compositional and classical variables.

| | *Dependent variable:* | |
|---|---|---|
| | y_ilr[, 1] | y_ilr[, 2] |
| Diplome_ilr1 | $-2.06(0.54)^{***}$ | $-1.51(0.46)^{**}$ |
| Diplome_ilr2 | $-1.28(0.80)$ | $-2.07(0.67)^{**}$ |
| Employ_ilr1 | $-0.05(0.30)$ | $-2.12(0.34)$ |
| Employ_ilr2 | $+0.12(0.37)$ | $-2.62(0.46)^{**}$ |
| Employ_ilr3 | $+0.30(0.30)$ | $-2.12(0.34)$ |
| Employ_ilr4 | $+0.13(0.11)$ | $-2.62(0.46)^{\cdot}$ |
| unemp | $-7.65(3.16)^{*}$ | $-2.12(0.34)^{***}$ |
| income | $+2.04(1.37)$ | $-2.62(0.46)^{***}$ |
| Constant | $-2.32(1.15)^{*}$ | $-4.80(0.97)^{***}$ |
| $R^2$ | 0.30 | 0.62 |
| Adjusted $R^2$ | 0.23 | 0.59 |
| Residual Std. Error (df = 86) | 0.30 | 0.26 |
| F Statistic (df = 8; 86) | $4.602^{***}$ | $17.85^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

In order to illustrate the impact of unemployment on the predicted shares, we choose three departments Ariège, Cantal and Bas-Rhin with different characteristics: Ariège has the maximum Left wing share, Cantal the maximum Right wing share and Bas-Rhin has the minimum Left wing share. We fix the values of the covariates at the values of each of the three departments and create a grid of fictive values of unemployment rates.

Figure 2.3 shows the predictions of vote shares in these departments Ariège, Cantal and Bas-Rhin as a function of unemployment rate (its minimum and maximum in the data base are figured by the dotted vertical lines). We first of all see the non linear nature of the relationship, and the fact that they differ from one department to the other. Note that the predicted shares using this model satisfy the constraint of unit sum and it clearly shows on the graph. In all cases, when the unemployment rate increases up to a given threshold of around 15%, the Left wing and the Extreme Right party gain votes at the expense of the Right wing. However, if unemployment keeps increasing beyond 15%, the Left wing starts loosing votes while the Right wing keeps decreasing and the Extreme Right keeps increasing. Overall, we can say that as unemployment rate varies, the Left wing proportion is more stable than the other two parties and that the other two parties affect each other like interconnecting pipes. Even though the three departments curves have the same general shape, we note differences: the maximum of the Left wing share is the highest in Ariège and lowest in Bas-Rhin; it is striking that the point at which the Left wing share and the Right wing share are equal is obtained at approximately the same value of unemployment rate in the three department but corresponds to different

values of the common Left wing- Right wing share; this value is lower than the maximum Left wing share in Ariège whereas it is slightly higher in Bas-Rhin. A major difference between the three departments is revealed when one looks at the highest of the three predictions: in Ariège, all realistic scenarios (between two vertical lines) result in a victory of the Left wing, in Cantal, all three parties may win depending on the value of unemployment and finally in Bas-Rhin there is no scenario leading to a victory of the Left wing . To represent this differently, we plot on Figure 2.4 a ternary diagram showing the curve of predicted shares as a function of unemployment rates together with a small square figuring the observed position of the given department in the triangle and a small diamond the corresponding prediction on the curve. The curve, a line in the simplex, is colored according to the value of unemployment rate. The Cantal department is better predicted than the Ariège and Bas-Rhin departments. We also note that the maximum predicted proportion for the Left wing is lower in the Bas-Rhin than in the other two departments. Finally, the triangle is divided in three parts with respect to the highest shares to highlight the winning party as in Figure 2.3.

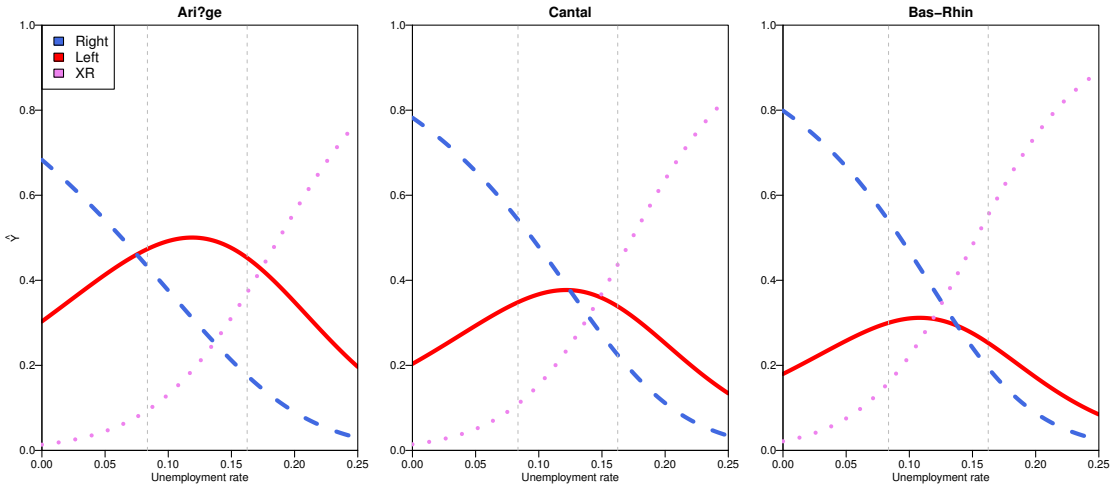


Figure 2.3: The vote share prediction curves in three departments: Ariège, Cantal, Bas-Rhin respectively (the grey dotted line show the minimum and the maximum observed unemployment rates)

Let us now turn attention to the case of a compositional explanatory variable impact. Figure 2.5 presents the vote share predictions according to the Diploma variable in the same three departments (Ariège, Cantal and Bas-Rhin). The principle is the same: all explanatory variables are fixed to the value of the given department except Diploma. We create of grid in the Diploma triangle and compute the predicted shares at each point of this grid. However, since it is impossible to plot a function from the simplex to the simplex, we choose to summarize the predicted shares by the winning party (corresponding to the highest share) and color the triangle in the Diploma space according to the winning party color. The observed shares are also figured by black points in this

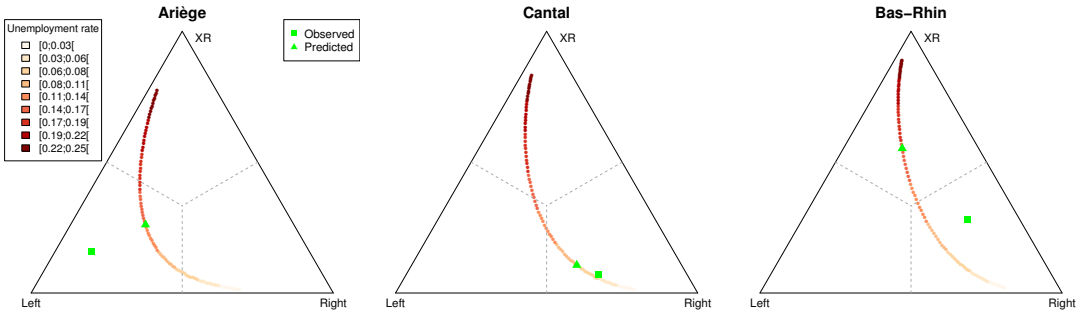


Figure 2.4: The vote share prediction ternary diagrams for fixed covariates given by three departments: Ariège, Cantal, Bas-Rhin respectively as a function of the unemployment rate. The green squares show the observed vote share of these departments and the green triangles on the red curve the corresponding predictions.

ternary diagram thus showing the realistic values. This figure shows that there is a large proportion of fictive situations (in terms of diploma proportions) where the Left party would win.

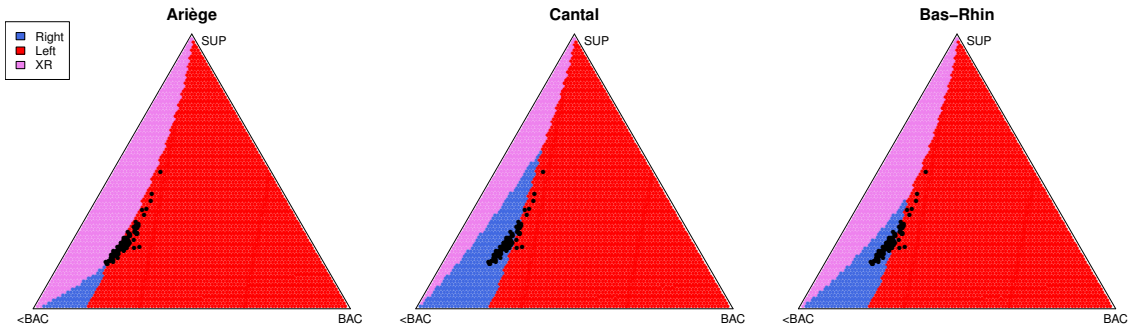


Figure 2.5: Predictions of vote shares according to Diploma for fixed covariates given by the departments Ariège (left plot), Cantal (middle plot) and Bas-Rhin (right plot)

## 2.4 Conclusion

The above analysis demonstrates that the CODA regression models can be useful in the context of political economy. We analyze how the predicted values in these models vary with the predictors and propose new graphical tools to explore the impact of some socio-economic variables on election results. Our future perspectives are to introduce the geographical dimension in the model and to use the Student distribution (Katz and King (1999)) instead of the normal distribution. At last, we plan to compute the

elasticities as in Morais et al. (2017) to characterize the impacts of the covariates in a more quantitative way.

# Chapter 3

# Abstract

In the previous chapter, we build a multivariate Normal regression model to study the impact of socio-economic factors on the outcome of an election. However, political data often exhibit heavy tail behavior. Katz and King Katz and King (1999) propose to replace the multivariate Normal distribution by the multivariate Student distribution. The multivariate Student distribution is a useful and popular distribution and provides a robust estimation procedure. This distribution is highly appreciated for heavy tails data, especially with data in finance and in political economy. There are several ways to define the multivariate Student distribution: the independent multivariate Student distribution (IT) and the uncorrelated multivariate Student (UT) distribution. The IT distribution considers that the components of the random vector are independent with the same marginal Student distribution while the UT distribution postulates a joint multivariate Student distribution for the vector of interest with uncorrelated vectors of individuals. In order to model multivariate, possibly heavy-tailed data, we compare the multivariate normal model (N) with two versions of the multivariate Student model. After recalling some facts about these distributions and models, known but scattered in the literature and fixed degree of freedom, we prove that the maximum likelihood estimator of the covariance matrix in the UT model is asymptotically biased and propose an unbiased version. We provide implementation details for an iterative reweighted algorithm to compute the maximum likelihood estimators of the parameters of the IT model. We present a simulation study to compare the bias and root mean squared error of the ensuing estimators of the regression coefficients and covariance matrix under several scenarios of the potential data-generating process, misspecified or not. The UT model is simpler to fit than the IT model but the single realization assumption is a limitation. Therefore, in this chapter we propose a graphical tool and a test based on the Mahalanobis distance to guide the choice between the gaussian N and the IT Student models. The advantage of this approach is that it is simple based on a one dimensional variable whereas the original observations have a multidimensional nature. We also present an application to model vectors of financial assets returns.

# Résumé

Dans le chapitre précédent, nous construisons un modèle de régression normal à plusieurs variables pour étudier l'impact des facteurs socioéconomiques sur le résultat d'une élection. Cependant, les données politiques montrent souvent un comportement de queue lourde. Katz et King Katz and King (1999) proposent de remplacer la distribution multivariée de Normal par la distribution multivariée de Student. La distribution multivariée de Student est une distribution utile et populaire qui fournit une procédure d'estimation robuste. Cette distribution est très appréciée pour les données sur les queues lourdes, en particulier pour les les données en finance et en économie politique. Il existe plusieurs façons de définir la distribution de Student multivariée: la distribution de Student multivariée indépendante et la distribution de Student multivariée non corrélée. La distribution IT considère que les composants du vecteur aléatoire sont indépendants avec la même distribution de Student marginale, tandis que la distribution UT postule une distribution de Student commune à plusieurs variables pour le vecteur d'intérêt avec des vecteurs d'individus non corrélés. Afin de modéliser des données multivariées, éventuellement lourdes, nous comparons le modèle normal multivarié (N) à deux versions du modèle de Student multivarié. Après avoir rappelé quelques faits sur ces distributions et modèles, connus mais éparpillés dans la littérature et à degré de liberté fixe, nous montrons que l'estimateur du maximum de vraisemblance de la matrice de covariance dans le modèle UT est asymptotiquement biaisé et proposons une version non biaisée. Nous fournissons des détails de mise en œuvre pour un algorithme itératif repondéré permettant de calculer les estimateurs de vraisemblance maximale des paramètres du modèle IT. Nous présentons une étude de simulation pour comparer le biais et l'erreur quadratique moyenne fondamentale des estimateurs des coefficients de régression et de la matrice de covariance résultants dans plusieurs scénarios du processus potentiel de génération de données, mal spécifié ou non. Le modèle UT est plus simple à adapter que le modèle informatique, mais l'hypothèse de réalisation unique est une limitation. Dans ce chapitre, nous proposons donc un outil graphique et un test basés sur la distance de Mahalanobis pour guider le choix entre les modèles Gaussien N et Student IT. L'avantage de cette approche est qu'elle est simple, basée sur une variable unidimensionnelle, alors que les observations originales ont un caractère multidimensionnel. Nous présentons également une application permettant de modéliser les vecteurs des rendements des actifs financiers.

# Multivariate Student versus Multivariate Gaussian Regression Models with Application to Finance[1][2]

**Abstract**
To model multivariate, possibly heavy-tailed data, we compare the multivariate normal model (N) with two versions of the multivariate Student model: the independent multivariate Student (IT) and the uncorrelated multivariate Student (UT). After recalling some facts about these distributions and models, known but scattered in the literature, we prove that the maximum likelihood estimator of the covariance matrix in the UT model is asymptotically biased and propose an unbiased version. We provide implementation details for an iterative reweighted algorithm to compute the maximum likelihood estimators of the parameters of the IT model. We present a simulation study to compare the bias and root mean squared error of the ensuing estimators of the regression coefficients and covariance matrix under several scenarios of the potential data-generating process, misspecified or not. We propose a graphical tool and a test based on the Mahalanobis distance to guide the choice between the competing models. We also present an application to model vectors of financial assets returns.

**Keywords**
multivariate regression models; heavy-tailed data; Mahalanobis distances; maximum likelihood estimator; independent multivariate Student distribution; uncorrelated multivariate Student distribution

## 3.1 Introduction

Many applications involving models for multivariate data underline the limitations of the classical multivariate Gaussian model, mainly due to its inability to model heavy

[1] Published to Journal of Risk and Financial Management
[2] Joint work with A. Ruiz-Gazen, C. Thomas-Agnan and T. Laurent

tails. It is then natural to turn attention to a more flexible family of distributions, for example the multivariate Student distribution.

In one dimension, the generalized Student distribution encompasses the Gaussian distribution as a limit when the number of degrees of freedom or shape parameter tends to infinity, allowing for heavier tails when the shape parameter is small. As we will see, a first difficulty in higher dimensions is that there are several kinds of multivariate Student distributions; see for example Johnson and Kotz (1972) and more recently Kotz and Nadarajah (2004). A nice summary of the properties of the multivariate Student distribution that we will use later on in this paper, and its comparison with the Gaussian multivariate, can be found in Roth (2012).

Before going further, let us mention that it is not so easy to have a clear overview of the results in terms of Student regression models for at least three reasons. The first reason is that this topic is scattered, with some papers in the statistical literature and others in the econometrics literature, sometimes without cross-referencing. The second reason is that the word "multivariate" is sometimes misleading since, as we will see, the multivariate Student is used to define a univariate regression model. At last, the distinction between models UT and IT (see below) is not always clearly announced in the papers. Other miscellaneous reasons are that some authors just fit the distribution without covariates and finally that some authors consider the degrees of freedom as fixed, whereas others estimate it. Our first purpose here is to lead the reader through this literature and gather the results concerning the maximum likelihood estimators of the parameters in the multivariate UT and IT models with a common notation. In the present paper, we consider a multivariate dependent vector and a linear regression model with different assumptions on the error term distribution. The most common and convenient assumption is the Gaussian distribution. For a Gaussian vector, the assumption of independent coordinates is equivalent to the assumption of uncorrelated coordinates. Such an equivalence is no longer true when considering a multivariate Student distribution. We thus consider two cases: uncorrelated (UT) on the one hand and independent Student (IT) error vectors on the other hand.

The purpose of this paper is to contribute to the UT and IT models as well as to their comparisons. First of all, for the UT model, we extend to the multivariate case the results of Zellner (1976) for the derivation of the maximum likelihood estimators and Zellner's formula (Zellner (1976)) for the bias of the covariance matrix estimator, and we prove that it does not vanish asymptotically. For the multivariate IT model, in the same spirit as Lange and Sinsheimer (1993), we provide details for the implementation of an iterative reweighted algorithm to compute the maximum likelihood estimators of the parameters. We devise a simulation study to measure the impact of misspecification on the bias, variance, and mean squared error of these different parameters' estimates under several data-generating processes (Gaussian, UT ,and IT) and try to answer the question: what are the consequences of a wrong specification? Finally we introduce a new procedure for model selection based on the knowledge of the distribution of the Mahalanobis distances under the different data-generating processes (DGP).

One application attracted our attention in the finance literature. The work in Platen

and Rendek (2008) identified the Student distribution with between three and five degrees of freedom, with a concentration around four, as the typical distribution for modeling the distribution of log-returns of world stock indices. They embedded the Student $t$ in the class of generalized hyperbolic distributions, itself a subclass of the normal/independent family. For bivariate returns, the work in Fung and Seneta (2010) compared a multivariate Student IT model with an alternative model obtained by a more complex mixing representation from the point of view of asymptotic tail dependence. The work in Huber and Ronchetti (2009) insisted on the fact that the choice of distribution matters when optimizing the portfolio. They found that the Student UT model performs the best in the class of symmetric generalized hyperbolic distributions. The work in Kan and Zhou (2017) advocated using a multivariate IT model for fitting the joint distribution of stock returns for a few fixed values of the degrees of freedom parameter and showed that this model outperforms the multivariate Gaussian.

In Section 3.2, after recalling the univariate results, we extend the results of Zellner (1976) for the derivation of the maximum likelihood estimators and its properties in the UT model and propose an iterative implementation for the IT model. We present the results of the simulation study in Section 3.3 and of the model selection strategy in Section 3.4 using a toy example and a dataset from finance. Section 3.5 summarizes the findings and gives recommendations.

## 3.2 Multivariate Regression Models

### 3.2.1 Literature Review

In order to define a Student regression model, even in the univariate case (single dependent variable), one needs to use the multivariate Student distribution to describe the joint distribution of the vector of observations for the set of statistical units. There are mainly two options, which were described in Kelejian and Prucha (1985) for the case of univariate regression. Indeed, the property of the equivalence between the independence and uncorrelatedness for components of a Gaussian vector are not satisfied anymore for a multivariate Student vector. One option, which we will call the IT model (for independent $t$-distribution) in the sequel, considers that the components of the random disturbance vector of the regression model are independent with the same marginal Student distribution. The second option, which we will call the UT model (for uncorrelated $t$-distribution), postulates a joint multivariate Student distribution for the vector of disturbances. Note that in both models, the marginal distribution of each component still is Student univariate.

The work in Zellner (1976) introduced a univariate Student regression model of the type UT with known degrees of freedom and studied the corresponding maximum likelihood and Bayesian estimators (with some adapted priors). The work in Singh (1988) considered the case of univariate Student regression with the UT model and with unknown degrees of freedom and derived an estimator of the degrees of freedom and subsequent estimators of the other parameters. However, Fernandez and Steel (1999) showed that this estimator was not consistent. Using one possible representation of the

multivariate Student distribution, Lange and Sinsheimer (1993) embedded univariate Student regression with the UT model in a larger family of regression models (with normal/independent error distributions) and developed EM algorithms to compute their maximum likelihood estimates, as in Dempster (1980).

In the framework of the spherical error distribution, which includes the Student error model as a special case, the work in Fraser and Ng (1980) proved an extension to the multivariate case of Zellner's result stating that inference about the parameters corresponds closely to that under normal theory. Motivated by a financial application, the work in Sutradhar and Ali (1986) used a multivariate UT Student regression model with moment estimators instead of maximum likelihood and allowing the degrees of freedom to be unknown.

The univariate IT model was introduced in Fraser (1979) and compared to the UT model in Kelejian and Prucha (1985).

Concerning multivariate IT Student distributions, there was first a collection of results or applications for the case without regressors. The work in McNeil et al. (2005) used a representation of the multivariate IT Student distribution to derive an algorithm of the EM type for computing the maximum likelihood parameter estimators. They used the framework of normal mixture distributions in which the Student distribution can be expressed as a combination of a Gaussian random variable and an inverse gamma random variable. More recently, the work in Dogru et al. (2018) proposed a more robust extension, replacing maximum likelihood by a kind of M-estimation method based on the minimization of a q-entropy criterion. For the multivariate Student IT model, the work in Prucha and Kelejian (1984) derived the normal equations for the maximum likelihood estimators and their asymptotic properties with known degrees of freedom in a framework that encompasses our multivariate Student regression case. The work in Lange et al. (1989) illustrated this multivariate IT model on several examples. The work in Lange and Sinsheimer (1993) considered the framework of normal/independent error distributions (same as normal variance mixtures) and derived the EM algorithm for the maximum likelihood estimators in a model with covariates. The works in Liu and Rubin (1995) and Liu (1997) developed extensions of the EM algorithm for the multivariate IT model with known or unknown degrees of freedom, with or without covariates and with or without missing data. The work in Katz and King (1999) fit a multivariate IT distribution to multiparty electoral data. The work in Fernandez and Steel (1999) attracted attention to the fact that maximum likelihood inference can encounter problems of unbounded likelihood when the number of degrees of freedom is considered unknown and has to be estimated. Before engaging in the use of the multivariate Student distribution, it is wise to read Hofert (2013), which explained some traps to be avoided. One difficulty indeed is to be aware that some authors parametrize the multivariate Student distribution using the covariance matrix, while others use the scatter matrix, sometimes with the same notation for either one.

We consider the following version of the Student $p$-multivariate distribution denoted by $\mathbf{T}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with $\boldsymbol{\mu}$ being the $p$-vector of means, $\boldsymbol{\Sigma}$ being the $p \times p$ covariance matrix, and $\nu > 2$ the degrees of freedom. It is defined, for a $p$-vector $\mathbf{z}$, by the probability

density function:

$$p(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{f(\nu)}{\det(\boldsymbol{\Sigma})^{1/2}}\left[1+\frac{1}{\nu-2}(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right]^{-(\nu+p)/2}, \tag{3.1}$$

where $^T$ denotes the transpose operator, $f(\nu) = \dfrac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)(\nu-2)^{p/2}\pi^{p/2}}$ and $\Gamma$ is the usual Gamma function.

Note that the assumption $\nu > 2$ implies the existence of the first two moments of the distribution and that the above density function is parametrized in terms of the covariance matrix. In most of the literature on multivariate Student distributions, the density is rather parametrized as a function of the scatter matrix $((\nu-2)/\nu)\boldsymbol{\Sigma}$. Using the covariance matrix parametrization facilitates the comparison with the Gaussian distribution. We first recall some results in the univariate regression context.

### 3.2.2 Univariate Regression Case Reminder

In the univariate regression case and for a sample of size $n$, we have a one-dimensional dependent variable $\boldsymbol{\mathcal{Y}}_i$, $i = 1,\ldots,n$, whose values are stacked in a vector $\boldsymbol{\mathcal{Y}}$, and $K$ explanatory variables defining a $n \times (K+1)$ design matrix $\boldsymbol{\mathcal{X}}$ including the constant.

The regression model is written as $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_K)^T$ is a $(K+1)$-dimensional vector of parameters and the error term $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^T$ is an $n$-dimensional vector. If we consider that the design matrix is fixed with rank $K+1$ or look at the distribution of $\boldsymbol{\epsilon}$ conditional on $\boldsymbol{\mathcal{X}}$, the usual assumptions are the following. The errors $\epsilon_i$, $i = 1,\ldots,n$, are independent and identically distributed (i.i.d.) with expectation zero and equal variance $\sigma^2$. In this context, it is well known that the least squares estimator of $\boldsymbol{\beta}$ is equal to:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{Y}} \tag{3.2}$$

while the classical $\sigma^2$ estimator is $\hat{\sigma}^2 = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}/(n-K-1)$ where $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\beta}}$. These estimators are unbiased. In the case of a Gaussian error distribution, the estimator $\hat{\boldsymbol{\beta}}$ coincides with the maximum likelihood estimator of $\boldsymbol{\beta}$, while the maximum likelihood estimator of $\sigma^2$ is equal to $\hat{\sigma}^2$ multiplied by $(n-K-1)/n$ and is only asymptotically unbiased. In the Gaussian case, there is an equivalence between the $\epsilon_i$ being independent or uncorrelated. However, this property is no longer true for a Student distribution. This means that one should distinguish the case of uncorrelated errors from the case of independent errors. The case where the errors $\epsilon_i$, $i = 1,\ldots,n$, follow a joint $n$-dimensional Student distribution with diagonal covariance matrix and equal variance is called the UT model, and its coordinates are uncorrelated, but not independent. Interestingly, the maximum likelihood method for the UT model with known degrees of freedom leads to the least squares estimator (3.2) of $\boldsymbol{\beta}$ (see Zellner (1976)). This property is true for more general distributions as long as the likelihood is a decreasing function of $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$. Concerning the error variance, the maximum likelihood estimator is $(n-K-1)\nu\,\hat{\sigma}^2/(n(\nu-2))$ and is biased even asymptotically Zellner (1976). For the independent

case, we assume that the errors $\epsilon_i$, $i = 1, \ldots, n$, are i.i.d. with a Student univariate distribution and known degrees of freedom. The maximum likelihood estimators belong to the class of M-estimators, which are studied in detail in Chapter 7 of Huber and Ronchetti (2009). These estimators are defined through implicit equations and can be computed using an iterative reweighted algorithm.

In what follows, we consider the case of a multivariate dependent variable and propose to gather and complete the results from the literature. As we will see, the results derived in the multivariate case are very similar to their univariate counterpart. In particular, the maximum likelihood estimator of the error covariance matrix is biased for the uncorrelated Student model, while there is a need to define an iterative algorithm for the independent Student model.

### 3.2.3 The Multivariate Regression Model

Let us consider a sample of size $n$, and for $i = 1, \ldots, n$, let us denote the $L$-dimensional dependent vector by:

$$\boldsymbol{\mathcal{Y}}_i = (y_{i1}, \ldots, y_{iL})^T.$$

For $K$ explanatory variables, the design matrix is of size $L \times (K+1)L$ and is given by:

$$\boldsymbol{\mathcal{X}}_i = \boldsymbol{I}_L \otimes \boldsymbol{x}_i^T$$

for $i = 1, \ldots, n$, with the $(K+1)$-vector $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{iK})^T$, $\boldsymbol{I}_L$ the identity matrix with dimension $L$, and $\otimes$ the usual Kronecker product. The parameter of interest is a $(K+1)L$ vector given by:

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_L^T)^T,$$

where $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{Kj})^T$, for $j = 1, \ldots, L$, and the $L$-vector of errors is denoted by:

$$\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iL})^T$$

for $i = 1, \ldots, n$. We consider the linear model:

$$\boldsymbol{\mathcal{Y}}_i = \boldsymbol{\mathcal{X}}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{3.3}$$

with $\mathbb{E}(\boldsymbol{\epsilon}_i) = 0$ and $i = 1, \ldots, n$. Using matrix notations, we can write Model (3.3) as:

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.4}$$

with the $nL$-vectors:

$$\boldsymbol{\mathcal{Y}} = (\boldsymbol{\mathcal{Y}}_1^T, \ldots, \boldsymbol{\mathcal{Y}}_n^T)^T,$$
$$\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_n^T)^T$$

and the $nL \times (K+1)L$ matrix:

$$\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_1^T, \ldots, \boldsymbol{\mathcal{X}}_n^T)^T.$$

In what follows, we make different assumptions on the distribution of $\boldsymbol{\epsilon}$ and recall (for Gaussian and IT) or derive (for UT) the maximum likelihood estimators of the parameter $\boldsymbol{\beta}$ and of the covariance matrix of $\boldsymbol{\epsilon}$.

### 3.2.4 Multivariate Normal Error Vector

Let us first consider Model (3.4) with independent and identically distributed error vectors $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, following a multivariate normal distribution $\mathcal{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ with an $L$-vector of means equal to zero and an $L \times L$ covariance matrix $\boldsymbol{\Sigma}$. This model is denoted by $N$, and the subscript $N$ is used to denote the error terms $\boldsymbol{\epsilon}_{Ni}$, $i = 1, \ldots, n$, and the parameters $\boldsymbol{\beta}_N$ and $\boldsymbol{\Sigma}_N$ of the model. The maximum likelihood estimators of $\boldsymbol{\beta}_N$ and $\boldsymbol{\Sigma}_N$ are:

$$\hat{\boldsymbol{\beta}}_N = (\boldsymbol{\mathcal{X}}^T \boldsymbol{\mathcal{X}})^{-1} \boldsymbol{\mathcal{X}}^T \boldsymbol{\mathcal{Y}}, \tag{3.5}$$

$$\hat{\boldsymbol{\Sigma}}_N = \frac{\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_{Ni} \hat{\boldsymbol{\epsilon}}_{Ni}^T}{n}, \tag{3.6}$$

where $\hat{\boldsymbol{\epsilon}}_{Ni} = \boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}_i \hat{\boldsymbol{\beta}}_N$ (see, e.g., Theorem 8.4 from Seber (2009)).

The estimator $\hat{\boldsymbol{\beta}}_N$ is an unbiased estimator of $\boldsymbol{\beta}_N$ while the bias of $\hat{\boldsymbol{\Sigma}}_N$ is equal to $-((K+1)/n)\boldsymbol{\Sigma}_N$ and tends to zero when $n$ tends to infinity (see, e.g., Theorems 8.1 and 8.2 from Seber (2009)).

For data such as financial data, it is well known that the Gaussian distribution does not fit the error term well. Student distributions are known to be more appropriate because they have heavier tails than the Gaussian. As for the univariate case, for Student distributions, the independence of coordinates is not equivalent to their uncorrelatedness, and we consider below two types of Student distributions for the error term. In Section 3.2.5, the error vector $\boldsymbol{\epsilon}$ is assumed to follow a Student distribution with $nL$ dimensions and a particular block diagonal covariance matrix. More precisely, we assume that the error vectors $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, are identically distributed and uncorrelated but are not independent. In Section 3.2.6, however, we consider independent and identically distributed error vectors $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, with an $L$-dimensional Student distribution.

### 3.2.5 Uncorrelated Multivariate Student (UT) Error Vector

Let us consider Model (3.4) with uncorrelated and identically distributed error vectors $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, such that the vector $\boldsymbol{\epsilon}$ follows a multivariate Student distribution $\mathbf{T}_{nL}(\mathbf{0}, \boldsymbol{\Omega}, \nu)$ with known degrees of freedom $\nu > 2$ and covariance matrix $\boldsymbol{\Omega} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$. The $L \times L$ matrix $\boldsymbol{\Sigma}$ is the common covariance matrix of the $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$. This model is denoted by UT, and the subscript $UT$ is used to denote the error terms $\boldsymbol{\epsilon}_{UTi}$, $i = 1, \ldots, n$, and the parameters $\boldsymbol{\beta}_{UT}$, $\boldsymbol{\Omega}_{UT}$, and $\boldsymbol{\Sigma}_{UT}$ of the model. This model generalizes the model proposed by Zellner (1976) to the case of multivariate $\boldsymbol{\epsilon}_i$s. We derive the maximum likelihood estimators of $\boldsymbol{\beta}_{UT}$ and $\boldsymbol{\Sigma}_{UT}$ in Proposition 1 and give the bias of the covariance estimator in Proposition 2. The proofs of the propositions are given in the Appendix.

**Proposition 1.** *The maximum likelihood estimators of $\boldsymbol{\beta}_{UT}$ and $\boldsymbol{\Sigma}_{UT}$ are given by:*

$$\begin{aligned} \hat{\beta}_{UT} &= \left(\boldsymbol{\mathcal{X}}^T \boldsymbol{\mathcal{X}}\right)^{-1} \boldsymbol{\mathcal{X}}^T \boldsymbol{\mathcal{Y}}, \\ \hat{\boldsymbol{\Sigma}}_{UT} &= \frac{\nu}{\nu - 2} \frac{\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_{UTi} \hat{\boldsymbol{\epsilon}}_{UTi}^T}{n}, \end{aligned} \tag{3.7}$$

*where* $\hat{\boldsymbol{\epsilon}}_{UTi} = \boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}_i\hat{\boldsymbol{\beta}}_{UT}$.

The next proposition gives the bias of the maximum likelihood estimators and generalizes Zellner's result (Zellner (1976), p. 402) to the multivariate UT model. The maximum likelihood estimator of $\boldsymbol{\beta}_{UT}$ coincides with the least squares and with the method of moment estimators and is unbiased. This is no longer the case for the maximum likelihood estimator of $\boldsymbol{\Sigma}_{UT}$, which is biased even asymptotically. This gives an example of a maximum likelihood estimator that is not asymptotically unbiased in a context where the random variables are not independent. It illustrates that the independence assumption is crucial to derive the usual properties of the maximum likelihood estimators. Note that the method of moments estimator is a consistent estimator of $\boldsymbol{\Sigma}_{UT}$ (see Sutradhar and Ali (1986)).

**Proposition 2.** *The estimator* $\hat{\boldsymbol{\beta}}_{UT}$ *is unbiased for* $\boldsymbol{\beta}_{UT}$*. The estimator* $\hat{\boldsymbol{\Sigma}}_{UT}$ *is biased for* $\boldsymbol{\Sigma}_{UT}$ *even asymptotically. More precisely,*

$$\mathbb{E}(\hat{\boldsymbol{\Sigma}}_{UT}) = \frac{n-K}{n}\frac{\nu}{\nu-2}\boldsymbol{\Sigma}_{UT}$$

A consequence of Proposition 2 is that an asymptotically unbiased estimator of $\boldsymbol{\Sigma}_{UT}$ is given by $\tilde{\boldsymbol{\Sigma}}_{UT} = \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_{UTi}\hat{\boldsymbol{\epsilon}}_{UTi}^T/n$.

### 3.2.6 Independent Multivariate Student (IT) Error Vector

Let us consider Model (3.4) using the notations of Section 3.2.3 with i.i.d. $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, following a Student distribution with $L$ dimensions and known degrees of freedom $\nu > 2$. We denote this model by IT and the parameters of the model by $\boldsymbol{\beta}_{IT}$ and $\boldsymbol{\Sigma}_{IT}$. The IT model is a particular case of Prucha and Kelejian (1984) where the $B$ matrix in Expression (2.1) in Prucha and Kelejian (1984) is equal to zero.

Following Prucha and Kelejian (1984), we derive the maximum likelihood estimators for the IT model.

**Proposition 3.** *The maximum likelihood estimators of* $\boldsymbol{\beta}_{IT}$ *and* $\boldsymbol{\Sigma}_{IT}$ *in the IT regression model satisfy the following implicit equations:*

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{IT} &= \left(\sum_{i=1}^{n} \hat{w}_{ITi}\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\Sigma}}_{IT}^{-1}\boldsymbol{\mathcal{X}}_i\right)^{-1} \sum_{i=1}^{n} \hat{w}_{ITi}\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\Sigma}}_{IT}^{-1}\boldsymbol{\mathcal{Y}}_i \\
\hat{\boldsymbol{\Sigma}}_{IT} &= \frac{1}{n}\sum_{i=1}^{n} \hat{w}_{ITi}\hat{\boldsymbol{\epsilon}}_{ITi}\hat{\boldsymbol{\epsilon}}_{ITi}^T
\end{aligned} \tag{3.8}$$

$$\textit{with:} \quad \hat{\boldsymbol{\epsilon}}_{ITi} = \boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}_i\hat{\boldsymbol{\beta}}_{IT} \quad \textit{and} \quad \hat{w}_{ITi} = \frac{\nu+L}{\nu-2+\hat{\boldsymbol{\epsilon}}_{ITi}^T\hat{\boldsymbol{\Sigma}}_{IT}^{-1}\hat{\boldsymbol{\epsilon}}_{ITi}}.$$

These estimators are consistent estimators of $\boldsymbol{\beta}_{IT}$ and $\boldsymbol{\Sigma}_{IT}$ (see Theorem 3.2 in Prucha and Kelejian (1984)). In order to compute them, we propose to implement the following iterative reweighted algorithm in the same spirit as in Huber and Ronchetti (2009) for the univariate case (see also Lange et al. (1989)).

Step 0: Let:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{IT}^{(0)} &= (\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{Y}}\\ \hat{\boldsymbol{\epsilon}}_{IT}^{(0)} &= \boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\beta}}_{IT}^{(0)}\\ \hat{\boldsymbol{\Sigma}}_{IT}^{(0)} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_{ITi}^{(0)}\hat{\boldsymbol{\epsilon}}_{ITi}^{(0)T}\end{aligned}$$

Step $k \to$ Step $(k+1)$, $k > 0$:

$$\begin{aligned}\hat{w}_{ITi}^{(k+1)} &= \frac{\nu + L}{\nu - 2 + \hat{\boldsymbol{\epsilon}}_{ITi}^{(k)}\hat{\boldsymbol{\Sigma}}_{IT}^{(k)-1}\hat{\boldsymbol{\epsilon}}_{ITi}^{(k)}}\\ \hat{\boldsymbol{\beta}}_{IT}^{(k+1)} &= \left(\sum_{i=1}^{n}\hat{w}_{ITi}^{(k+1)}\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\Sigma}}_{IT}^{(k)-1}\boldsymbol{\mathcal{X}}_i\right)^{-1}\sum_{i=1}^{n}\hat{w}_{ITi}^{(k+1)}\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\Sigma}}_{IT}^{(k)-1}\boldsymbol{\mathcal{Y}}_i\\ \hat{\boldsymbol{\epsilon}}_{IT}^{(k+1)} &= \boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\beta}}_{IT}^{(k+1)}\\ \hat{\boldsymbol{\Sigma}}_{IT}^{(k+1)} &= \frac{1}{n}\sum_{i=1}^{n}\hat{w}_{ITi}^{(k+1)}\hat{\boldsymbol{\epsilon}}_{ITi}^{(k+1)}\hat{\boldsymbol{\epsilon}}_{ITi}^{(k+1)T}\end{aligned}$$

The process is iterated until convergence. Note that this algorithm is given in detail in Section 7.8 of Huber and Ronchetti (2009) for a general class of univariate regression M-estimators. It is also sometimes called IRLS for iteratively-reweighted least squares and can be seen as a particular case of the EM algorithm (see Lange et al. (1989)).

Table 3.2.6 gathers the likelihoods and thus summarizes the three models of interest.

Table 3.1: Distribution of the error vector $\boldsymbol{\epsilon}$ in the Gaussian, UT, and IT models.

| Model | Distribution |
|---|---|
| $\mathrm{N}(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)$ | $\mathcal{N}_{nL}(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}_N) = \prod_{i=1}^{n}\mathcal{N}_L(\mathbf{0}, \boldsymbol{\Sigma}_N)$ |
| $\mathrm{UT}(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)$ | $\mathbf{T}_{nL}(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}_{UT}, \nu)$ |
| $\mathrm{IT}(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)$ | $\prod_{i=1}^{n}\mathbf{T}_L(\mathbf{0}, \boldsymbol{\Sigma}_{IT}, \nu)$ |

## 3.3 Simulation Study

### 3.3.1 Design

This study aims at comparing the properties of the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ as defined in the previous section for the multivariate Gaussian (N), the uncorrelated multivariate Student (UT), and the independent multivariate Student (IT) error distributions, under several scenarios for the DGP. Note that for the UT model, we used the asymptotically unbiased estimator $\tilde{\boldsymbol{\Sigma}}_{UT}$ to estimate $\boldsymbol{\Sigma}_{UT}$. We considered a variety of degrees of freedom $\nu_{DGP}$ for the Student IT and UT models with a focus on values between three and five. We used the function rmvt from the R package mvnfast to simulate the Student distributions. For a sample size $n = 1000$ and a number of replications $N = 10{,}000$, we simulated an explanatory variable $\boldsymbol{\mathcal{X}}$ following a Gaussian distribution $\mathcal{N}(45, 10)$. The parameter vector $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$ are respectively chosen to be:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{02} \\ \beta_{12} \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ -3 \end{bmatrix} ; \ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

Note that similar results are obtained with other choices of parameters.

For each DGP, we calculate a number of Monte Carlo performance measures of the estimators proposed in Section 3.2. The performances are measured by the Monte Carlo relative bias (RB) and the mean squared error (MSE), which are defined for an estimator $\hat{\theta}$ of a parameter $\theta$ by:

$$\begin{aligned} \text{Bias}(\hat{\theta}) &= \frac{1}{N}\sum_{i=1}^{n} \hat{\theta}^{(i)} - \theta \\ \text{RB}(\hat{\theta}) &= 100\frac{\text{Bias}(\hat{\theta})}{\theta} \\ \text{MSE}(\hat{\theta}) &= \frac{1}{N}\sum_{i=1}^{N} \left(\hat{\theta}^{(i)} - \theta\right)^2. \end{aligned} \tag{3.9}$$

We also compute a relative root mean squared error (RRMSE) with respect to a baseline estimator $\tilde{\theta}$ as:

$$\text{RRMSE}(\hat{\theta}) = \left(\frac{\text{MSE}(\hat{\theta})}{\text{MSE}(\tilde{\theta})}\right)^{1/2}.$$

In our case, the baseline estimator is the maximum likelihood estimator (MLE) corresponding to the DGP. For example, in Table 3.2, the RRMSE of the $\hat{\boldsymbol{\beta}}_{IT}$ for the Gaussian DGP is the ratio of the MSE of $\hat{\boldsymbol{\beta}}_{IT}$ with the degrees of freedom $\nu_{MLE}$ and the MSE of $\hat{\boldsymbol{\beta}}_{N}$. Note that if $\hat{\theta} = \tilde{\theta}$, then the RRMSE of $\hat{\theta}$ is equal to one.

Table 3.2: Relative bias and relative root mean squared error of the estimators of $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}}_N$, $\hat{\boldsymbol{\beta}}_{UT}$,$\hat{\boldsymbol{\beta}}_{IT}$) for the corresponding DGP (Gaussian, UT, and IT).

| **DGP** | | **N** | | **UT** ($\boldsymbol{\nu_{DGP} = 3}$) | | **IT** ($\boldsymbol{\nu_{DGP} = 3}$) | |
|---|---|---|---|---|---|---|---|
| **Methods** | **Estimators** | **RB (%)** | **RRMSE** | **RB (%)** | **RRMSE** | **RB (%)** | **RRMSE** |
| $\hat{\boldsymbol{\beta}}_N$,$\hat{\boldsymbol{\beta}}_{UT}$ | $\hat{\beta}_{01}$ | −0.07 | 1.00 | −0.06 | 1.00 | −0.09 | 1.48 |
| | $\hat{\beta}_{02}$ | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.48 |
| | $\hat{\beta}_{11}$ | −0.02 | 1.00 | −0.01 | 1.00 | −0.07 | 1.46 |
| | $\hat{\beta}_{12}$ | −0.00 | 1.00 | −0.00 | 1.00 | −0.00 | 1.46 |
| $\hat{\boldsymbol{\beta}}_{IT}(\nu_{MLE} = 3)$ | $\hat{\beta}_{01}$ | −0.09 | 1.04 | −0.09 | 1.09 | −0.03 | 1.00 |
| | $\hat{\beta}_{02}$ | 0.00 | 1.04 | 0.00 | 1.09 | 0.00 | 1.00 |
| | $\hat{\beta}_{11}$ | −0.04 | 1.07 | −0.02 | 1.08 | −0.03 | 1.00 |
| | $\hat{\beta}_{12}$ | −0.00 | 1.07 | −0.00 | 1.08 | −0.00 | 1.00 |

### 3.3.2 Estimators of the $\beta$ Parameters

Table 3.3 reports the bias and the MSE of the Gaussian MLE estimator $\hat{\boldsymbol{\beta}}_N$, the UT MLE estimator $\hat{\boldsymbol{\beta}}_{UT}$ ($\nu_{\text{DGP}} = 3$), and the IT MLE estimator $\hat{\boldsymbol{\beta}}_{IT}$ ($\nu_{\text{DGP}} = 3$) when the model is well specified, i.e., under the corresponding DGP. The bias and MSE of the estimators of $\boldsymbol{\beta}$ are small and comparable under the Gaussian and the UT DGP, but smaller for the IT DGP. Note that, in our implementation, the results of the algorithm for the IT estimators are very similar to those obtained using the function heavyLm from the R package heavy.

Table 3.3: Bias and MSE of the maximum likelihood estimators of $\boldsymbol{\beta}$ for the corresponding DGP (Gaussian, UT, and IT).

| **DGP** | **N** | | **UT** ($\boldsymbol{\nu_{DGP} = 3}$) | | **IT** ($\boldsymbol{\nu_{DGP} = 3}$) | |
|---|---|---|---|---|---|---|
| Estimators | Bias | MSE | Bias | MSE | Bias | MSE |
| $\hat{\beta}_{01}$ | $-1.39 \times 10^{-3}$ | $4.57 \times 10^{-2}$ | $-1.27 \times 10^{-3}$ | $3.72 \times 10^{-2}$ | $6.65 \times 10^{-4}$ | $1.99 \times 10^{-2}$ |
| $\hat{\beta}_{02}$ | $2.41 \times 10^{-5}$ | $2.18 \times 10^{-5}$ | $1.47 \times 10^{-5}$ | $1.76 \times 10^{-5}$ | $9.90 \times 10^{-6}$ | $9.50 \times 10^{-6}$ |
| $\hat{\beta}_{11}$ | $-6.62 \times 10^{-4}$ | $2.16 \times 10^{-2}$ | $3.23 \times 10^{-4}$ | $2.05 \times 10^{-2}$ | $-1.02 \times 10^{-3}$ | $9.84 \times 10^{-3}$ |
| $\hat{\beta}_{12}$ | $1.87 \times 10^{-5}$ | $1.02 \times 10^{-5}$ | $3.90 \times 10^{-6}$ | $9.60 \times 10^{-6}$ | $2.14 \times 10^{-5}$ | $4.70 \times 10^{-6}$ |

In Table 3.2, we start considering misspecifications and report the corresponding relative values RB and RRMSE of the same estimators and the same DGP as in Table 3.3 with all possible combinations of DGP and estimation methods. The results indicate that the RB of $\hat{\boldsymbol{\beta}}$ are all very small. If the DGP is Gaussian and the estimator is IT, the RRMSE of coordinates of $\hat{\boldsymbol{\beta}}$ is about 1.09. However, if the DGP is IT and the estimator is Gaussian, the RRMSE of coordinates of $\hat{\boldsymbol{\beta}}$ is higher (from 1.46–1.48). Hence for the Gaussian DGP, we do not loose too much efficiency using the IT estimator $\hat{\boldsymbol{\beta}}_{IT}$ with three degrees of freedom. Inversely, we loose much more efficiency when using $\hat{\boldsymbol{\beta}}_N$ for the IT DGP with three degrees of freedom.

In order to consider more degrees of freedom (3, 4, and 5), we now drop the bias and focus on the RRMSE. Table 3.4 indicates that the RRMSE of $\hat{\boldsymbol{\beta}}$ is very similar

and close to one, with a maximum of 1.09, except for the case of the N estimator under the IT DGP, where it can reach 1.48. The work in Maronna (1976) provided theoretical asymptotic efficiencies of the Student versus the Gaussian estimators, the ratio of asymptotic variances being equal to $\frac{(\nu-2)(\nu+L+2)}{\nu(\nu+L)}$. The values obtained in Table 3.2 are very similar to these asymptotic values.

Table 3.4: The root relative mean squared errors of $\hat{\boldsymbol{\beta}}$.

| **Methods** | **DGP** | **N** | **UT** | | | **IT** | | |
|---|---|---|---|---|---|---|---|---|
| | **RRMSE** | | $\nu_{DGP}=3$ | $\nu_{DGP}=4$ | $\nu_{DGP}=5$ | $\nu_{DGP}=3$ | $\nu_{DGP}=4$ | $\nu_{DGP}=5$ |
| N | $\hat{\beta}_{01}$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.48 | 1.22 | 1.14 |
| | $\hat{\beta}_{02}$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.48 | 1.23 | 1.14 |
| | $\hat{\beta}_{11}$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.46 | 1.22 | 1.13 |
| | $\hat{\beta}_{12}$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.46 | 1.22 | 1.13 |
| IT ($\nu_{MLE}=3$) | $\hat{\beta}_{01}$ | 1.04 | 1.09 | 1.09 | 1.08 | 1.00 | 1.00 | 1.01 |
| | $\hat{\beta}_{02}$ | 1.04 | 1.09 | 1.09 | 1.08 | 1.00 | 1.00 | 1.01 |
| | $\hat{\beta}_{11}$ | 1.07 | 1.08 | 1.10 | 1.08 | 1.00 | 1.00 | 1.01 |
| | $\hat{\beta}_{12}$ | 1.07 | 1.08 | 1.09 | 1.09 | 1.00 | 1.00 | 1.01 |
| IT ($\nu_{MLE}=4$) | $\hat{\beta}_{01}$ | 1.02 | 1.07 | 1.06 | 1.06 | 1.00 | 1.00 | 1.00 |
| | $\hat{\beta}_{02}$ | 1.01 | 1.06 | 1.06 | 1.05 | 1.00 | 1.00 | 1.00 |
| | $\hat{\beta}_{11}$ | 1.04 | 1.06 | 1.07 | 1.06 | 1.00 | 1.00 | 1.00 |
| | $\hat{\beta}_{12}$ | 1.04 | 1.05 | 1.07 | 1.06 | 1.00 | 1.00 | 1.00 |
| IT ($\nu_{MLE}=5$) | $\hat{\beta}_{01}$ | 1.00 | 1.05 | 1.05 | 1.04 | 1.01 | 1.00 | 1.00 |
| | $\hat{\beta}_{02}$ | 1.00 | 1.05 | 1.05 | 1.04 | 1.01 | 1.00 | 1.00 |
| | $\hat{\beta}_{11}$ | 1.03 | 1.04 | 1.05 | 1.05 | 1.01 | 1.00 | 1.00 |
| | $\hat{\beta}_{12}$ | 1.03 | 1.04 | 1.05 | 1.05 | 1.01 | 1.00 | 1.00 |

Figure 3.1 shows the performances in terms of RRMSE of the IT estimators $\hat{\beta}_{12}^{IT}$ under different DGP as a function of the degrees of freedom of the IT estimator ($\nu_{\text{MLE}}$). The considered DGP are the Gaussian, UT, and IT DGP with the degrees of freedom $\nu_{\text{DGP}} = 3$ (respectively, $\nu_{\text{DGP}} = 4$, $\nu_{\text{DGP}} = 5$) on the left (respectively, middle, right) plot. Overall, the RRMSE of $\hat{\beta}_{12}^{IT}$ for the IT DGP has a down trend and then an up trend, while for the Gaussian and the UT DGP, the RRMSE are decreasing when $\nu_{\text{MLE}}$ increases. The maximum RRMSE of $\hat{\beta}_{12}^{IT}$ is around 1.09 under the UT DGP and is around 1.08 under the Gaussian DGP. It decreases then to one when $\nu_{\text{MLE}}$ increases to twenty under the Gaussian and the UT DGP; thus, the risk under misspecification is not very high. The curve is U-shaped under the IT DGP with a minimum when $\nu_{\text{MLE}} = \nu_{\text{DGP}}$. The worst performance is when $\nu_{\text{DGP}}$ is small and $\nu_{\text{MLE}}$ is large. The RRMSE of $\hat{\beta}_{12}^{IT}$ with $\nu_{\text{DGP}} = 4$ is similar than the one with $\nu_{\text{DGP}} = 5$.

### 3.3.3 Estimators of the Variance Parameters

Table 3.5 reports the biases and the MSE of $\hat{\rho}$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ for the Gaussian DGP, the UT ($\nu_{\text{DGP}} = 3$) DGP, and the IT ($\nu_{\text{DGP}} = 3$) DGP. The bias and the MSE of $\hat{\rho}$ are very similar and small for all cases. The MSE of the Gaussian estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are small

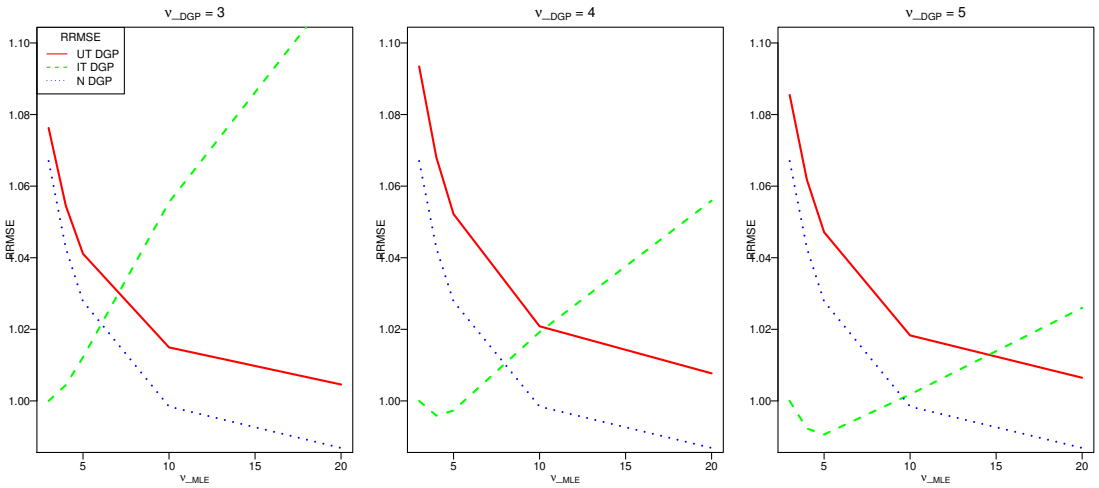


Figure 3.1: The RRMSE of the IT estimator of $\hat{\beta}_{12}$ for the UT DGP in solid line, for the IT DGP in dashed line, and for the Gaussian DGP in dotted line with $\nu_{DGP} = 3$ (respectively, $\nu_{DGP} = 4$, $\nu_{DGP} = 5$) on the left (respectively, middle, right) plot.

under the Gaussian DGP, but they are higher under the UT and IT DGP. The biases and MSE of the IT estimator $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are small under the IT DGP, but high under the Gaussian and the UT DGP. Besides, Table 3.5 also indicates that there is no method that estimates the variances well under the UT DGP.

Table 3.5: The bias and the MSE of $\hat{\rho}$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$.

| **Methods** | **DGP** | **N** | | **UT** ($\nu_{DGP} = 3$) | | **IT** ($\nu_{DGP} = 3$) | |
|---|---|---|---|---|---|---|---|
| | | **Bias** | **MSE** | **Bias** | **MSE** | **Bias** | **MSE** |
| **N** | $\hat{\rho}$ | $-4.85 \times 10^{-4}$ | $9.46 \times 10^{-4}$ | $-2.08 \times 10^{-4}$ | $7.68 \times 10^{-4}$ | $-3.99 \times 10^{-3}$ | $1.17 \times 10^{-2}$ |
| | $\hat{\sigma}_1^2$ | $-3.89 \times 10^{-3}$ | $8.33 \times 10^{-3}$ | $-1.05 \times 10^{-1}$ | 58 | $6.94 \times 10^{-3}$ | 3.17 |
| | $\hat{\sigma}_2^2$ | $-1.75 \times 10^{-3}$ | $2.01 \times 10^{-3}$ | $-5.17 \times 10^{-2}$ | 14.93 | $-1.77 \times 10^{-2}$ | $2.85 \times 10^{-1}$ |
| **IT** | $\hat{\rho}$ | $-1.70 \times 10^{-4}$ | $8.94 \times 10^{-4}$ | $-2.18 \times 10^{-4}$ | $9.05 \times 10^{-4}$ | $-2.03 \times 10^{-4}$ | $1.07 \times 10^{-3}$ |
| | $\hat{\sigma}_1^2$ | 2.00 | 4.06 | 1.80 | 244.87 | $-1.43 \times 10^{-2}$ | $1.54 \times 10^{-2}$ |
| $\nu_{MLE} = 3$ | $\hat{\sigma}_2^2$ | 1.00 | 1.02 | 0.91 | 64.75 | $-7.30 \times 10^{-3}$ | $3.94 \times 10^{-3}$ |

As before, we now consider misspecified cases and focus on relative bias in Table 3.6. We observe that the relative bias for $\hat{\rho}$ is negligible in all situations. The RB for $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are also quite small (less than around 5%) when using the Gaussian estimator for all DGP. This is also true when using the IT estimator for the IT DGP with the same degrees of freedom $\nu_{MLE} = \nu_{DGP}$. There are some biases for $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ if the DGP is Gaussian or UT and the estimator is IT. For this estimator, the relative bias of $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ is around 100% for the Gaussian DGP, 96% for the UT DGP with $\nu_{\text{DGP}} = 5$ and $\nu_{\text{MLE}} = 3$, and 22% for the UT DGP with $\nu_{\text{DGP}} = 5$ and $\nu_{\text{MLE}} = 5$. The RB for $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are also quite high (up to 50%) for the IT estimator when the DGP is IT

with $\nu_{MLE} \neq \nu_{DGP}$. To summarize, in terms of the RB of the variance estimators, the Gaussian estimator yields better results than the IT estimator.

Table 3.6: The RB of $\hat{\rho}$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ with $\nu = 3, 4, 5$.

| **Methods** | **DGP** | **N** | **UT** | | | **IT** | | |
|---|---|---|---|---|---|---|---|---|
| | **RB (%)** | | $\nu_{DGP} = 3$ | $\nu_{DGP} = 4$ | $\nu_{DGP} = 5$ | $\nu_{DGP} = 3$ | $\nu_{DGP} = 4$ | $\nu_{DGP} = 5$ |
| **N** | $\hat{\rho}$ | −0.14 | −0.06 | −0.06 | −0.06 | −1.13 | −0.24 | 0.02 |
| | $\hat{\sigma}_1^2$ | −0.21 | −5.23 | −3.34 | −2.31 | 0.35 | −0.08 | −0.12 |
| | $\hat{\sigma}_2^2$ | −0.18 | −5.17 | −3.33 | −2.20 | −1.77 | −0.30 | −0.09 |
| **IT**, $\nu_{MLE} = 3$ | $\hat{\rho}$ | −0.05 | −0.06 | −0.06 | −0.06 | −0.06 | −0.04 | −0.02 |
| | $\hat{\sigma}_1^2$ | 99.99 | 90.25 | 93.89 | 95.80 | −0.72 | 32.79 | 50.12 |
| | $\hat{\sigma}_2^2$ | 100.05 | 90.60 | 93.90 | 96.03 | −0.73 | 32.79 | 50.13 |
| **IT**, $\nu_{MLE} = 4$ | $\hat{\rho}$ | −0.05 | −0.06 | −0.06 | −0.06 | −0.06 | −0.04 | −0.01 |
| | $\hat{\sigma}_1^2$ | 42.62 | 35.80 | 38.32 | 39.68 | −24.66 | −0.24 | 11.18 |
| | $\hat{\sigma}_2^2$ | 42.66 | 36.01 | 38.34 | 39.85 | −24.67 | −0.23 | 11.19 |
| **IT**, $\nu_{MLE} = 5$ | $\hat{\rho}$ | −0.06 | −0.06 | −0.06 | −0.06 | −0.06 | −0.04 | −0.00 |
| | $\hat{\sigma}_1^2$ | 24.71 | 18.85 | 21.03 | 22.23 | −31.75 | −10.13 | −0.14 |
| | $\hat{\sigma}_2^2$ | 24.74 | 19.02 | 21.04 | 22.38 | −31.76 | −10.13 | −0.14 |

Finally, Table 3.7 presents the RRMSE in the same cases. It shows that the RRMSE of $\hat{\rho}$ varies from 0.94–1.09 for all DGP except for the case of the IT DGP with the Gaussian estimator, which ranges between 1.42 and 3.21. Besides, if the DGP is Gaussian and the estimator is IT or if the DGP is IT and the estimator is Gaussian, the RRMSE of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are high in particular for $\nu_{DGP} = 3$ or $\nu_{MLE} = 3$: we loose a lot of efficiency in these misspecified cases. To conclude, we have seen from Table 3.6 that the RB of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are smaller for the Gaussian estimator than for the IT estimator. However, in terms of RRMSE, there is no clear advantage in using the Gaussian estimator with respect to the IT estimator.

It should be noted that for $\nu \leq 4$, the Student distribution has no fourth-order moment, which may explain the fact that the covariance estimators have large MSE.

Table 3.7: The RRMSE of $\hat{\rho}$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ in the Gaussian DGP, the UT DGP ($\nu_{\text{DGP}} = 3, 4, 5$), and the IT DGP ($\nu_{\text{DGP}} = 3, 4, 5$).

| **Methods** | **DGP** | **N** | **UT** | | | **IT** | | |
|---|---|---|---|---|---|---|---|---|
| | **RRMSE** | | $\nu_{DGP} = 3$ | $\nu_{DGP} = 4$ | $\nu_{DGP} = 5$ | $\nu_{DGP} = 3$ | $\nu_{DGP} = 4$ | $\nu_{DGP} = 5$ |
| **N** | $\hat{\rho}$ | 1.00 | 1.00 | 1.00 | 1.00 | 3.21 | 1.91 | 1.42 |
| | $\hat{\sigma}_1^2$ | 1.00 | 1.00 | 1.00 | 1.00 | 14.33 | 2.65 | 1.64 |
| | $\hat{\sigma}_2^2$ | 1.00 | 1.00 | 1.00 | 1.00 | 8.50 | 2.24 | 1.78 |
| **IT**, $\nu_{MLE} = 3$ | $\hat{\rho}$ | 0.97 | 1.09 | 1.09 | 1.09 | 1.00 | 1.00 | 1.01 |
| | $\hat{\sigma}_1^2$ | 22.07 | 2.05 | 2.11 | 2.16 | 1.00 | 5.89 | 9.18 |
| | $\hat{\sigma}_2^2$ | 22.45 | 2.08 | 2.11 | 2.16 | 1.00 | 5.77 | 9.13 |
| **IT**, $\nu_{MLE} = 4$ | $\hat{\rho}$ | 0.95 | 1.06 | 1.06 | 1.06 | 1.01 | 1.00 | 1.00 |
| | $\hat{\sigma}_1^2$ | 9.49 | 1.46 | 1.47 | 1.48 | 4.04 | 1.00 | 2.31 |
| | $\hat{\sigma}_2^2$ | 9.65 | 1.48 | 1.47 | 1.48 | 4.00 | 1.00 | 2.30 |
| **IT**, $\nu_{MLE} = 5$ | $\hat{\rho}$ | 0.94 | 1.05 | 1.05 | 1.05 | 1.01 | 1.00 | 1.00 |
| | $\hat{\sigma}_1^2$ | 5.58 | 1.27 | 1.27 | 1.28 | 5.16 | 1.99 | 1.00 |
| | $\hat{\sigma}_2^2$ | 5.68 | 1.28 | 1.28 | 1.27 | 5.10 | 1.95 | 1.00 |

In order to allow the reproducibility of the empirical analyses contained in the present and the following sections, some Supplementary Material is available at the following link: `http://www.thibault.laurent.free.fr/code/jrfm/`.

## 3.4 Selection between the Gaussian and IT Models

In this section, we propose a methodology to select a model between the Gaussian and independent Student models and to select the degrees of freedom for the Student in a short list of possibilities. Following the warnings of Fernandez and Steel (1999) and the empirical results of Katz and King (1999), Platen and Rendek (2008), and Kan and Zhou (2017), we decided to focus on a small selection of degrees of freedom and fit our models without estimating this parameter, considering that a second step of model selection will make the choice. Indeed, there is a limited number of interesting values, which are between three and eight (for larger values, the distribution gets close to being Gaussian). The work in Lange et al. (1989), p.883, proposed the likelihood ratio test for the univariate case. In what follows, we use the fact that the distribution of the Mahalanobis distances is known under the two DGP, which allows building a Kolmogorov–Smirnov test and using Q-Q plots. Unfortunately, this technique does not apply to the UT model for which the $n$ observations are a single realization of the multivariate distribution. One advantage of this approach is that the Mahalanobis distance is a one-dimensional variable, whereas the original observations have $L$ dimensions.

### 3.4.1 Distributions of Mahalanobis Distances

For an $L$-dimensional random vector $\mathbf{Y}$, with mean $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$, the squared Mahalanobis distance is defined by:

$$d^2 = (\mathbf{Y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu})$$

If $\mathbf{Y}_1, \dots \mathbf{Y}_n$ is a sample of size $n$ from the $L$-dimensional Gaussian distribution $\mathcal{N}_L(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$, the squared Mahalanobis distance of observation $i$, denoted by $d^2_{Ni}$, follows a $\chi^2_L$ distribution. If $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ are unknown, then the squared Mahalanobis distance of observation $i$ can be estimated by:

$$\hat{d}^2_{Ni} = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_N)^T \hat{\boldsymbol{\Sigma}}_N^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_N)$$

where $\hat{\boldsymbol{\mu}}_N = \bar{\mathbf{Y}} = \dfrac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i$ and $\hat{\boldsymbol{\Sigma}}_N$ is the sample covariance matrix. The work in Gnanadesikan and Kettenring (1972) (see also Bilodeau and Brenner (2008)) proved that this square distance follows a Beta distribution, up to a multiplicative constant:

$$\frac{n}{(n-1)^2} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_N)^T \hat{\boldsymbol{\Sigma}}_N^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_N) \sim Beta\left(\frac{L}{2}, \frac{n-L-1}{2}\right)$$

where $L$ is the dimension of $\mathbf{Y}$. For large $n$, this Beta distribution can be approximated by the chi-square distribution $d_{Ni}^2 \sim \chi_L^2$. According to Gnanadesikan and Kettenring (1972) (p. 172), $n = 25$ already provides a sufficiently large sample for this approximation, which is the case in all our examples below.
If we now assume that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is a sample of size $n$ from the $L$-dimensional Student distribution $\mathbf{Y}_i \sim \mathbf{T}(\boldsymbol{\mu}_{IT}, \boldsymbol{\Sigma}_{IT}, \nu)$, then the squared Mahalanobis distance of observation $i$, denoted by $d_{ITi}^2$ and properly scaled, follows a Fisher distribution (see Roth (2012)):

$$\frac{1}{L}\frac{\nu}{\nu-2} d_{ITi}^2 \sim \mathcal{F}(L, \nu)$$

If $\boldsymbol{\mu}_{IT}$ and $\boldsymbol{\Sigma}_{IT}$ are unknown, then the squared Mahalanobis distance of observation $i$ can be estimated by:

$$\hat{d}_{ITi}^2 = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{IT})^T \hat{\boldsymbol{\Sigma}}_{IT}^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{IT}),$$

where $\hat{\boldsymbol{\mu}}_{IT}$ and $\hat{\boldsymbol{\Sigma}}_{IT}$ are the MLE of $\boldsymbol{\mu}_{IT}$ and $\boldsymbol{\Sigma}_{IT}$. Note that in the IT model, $\hat{\boldsymbol{\mu}}_{IT}$ is no longer equal to $\bar{\mathbf{Y}}$. Up to our knowledge, there is no result about the distribution of $\hat{d}_{ITi}^2$.

In the elliptical distribution family, the distribution of Mahalanobis distances characterizes the distribution of the observations. Thus, in order to test the normality of the data, we can test whether the Mahalanobis distances follow a chi-square distribution. Similarly, testing the Student distribution is equivalent to testing whether the Mahalanobis distances follow the Fisher distribution. There are two difficulties with the approach. The first one is that the estimated Mahalanobis distances are not a sample from the chi-square (respectively, the Fisher) distribution because there is dependence due to the estimation of the parameters. The second one is that, in our case, we not only estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, but we are in a regression framework where $\boldsymbol{\mu}$ is linear combination of regressors, and we indeed estimate its coefficients. In what follows, we will ignore these two difficulties and consider that, for large $n$, the distributions of the estimated Mahalanobis distances behave as if $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were known.

We propose to implement several Kolmogorov–Smirnov tests in order to test different null hypothesis: Gaussian, Student with three degrees of freedom, and Student with four degrees of freedom. As an exploratory tool, we also propose drawing Q-Q plots of the Mahalanobis distances with respect to the chi-square and the Fisher distribution (see Small (1978)).

### 3.4.2 Examples

This section illustrates some applications of the proposed methodology for selecting a model. We use a real dataset from finance and three simulated datasets with the same DGP as in Section 3.3.

The real dataset is the daily closing share price of IBM and MSFT, which are imported from Yahoo Finance from 3 January 2007–27 September 2018 using the quantmod

package in R. It contains $n = 2955$ observations. Let $\mathbf{S}_t,\ t = 1, \ldots, n$ be the daily share price of IBM and MSFT and $\mathbf{Y}_t$ be the log-price increment (return) (see Fung and Seneta (2010)) over a day period, then:

$$\mathbf{Y}_t = \log \mathbf{S}_t - \log \mathbf{S}_{t-1}.$$

The three other datasets are simulated using the same model as in Section 3.3 with the Gaussian DGP, the IT DGP with $\nu_{DGP} = 3$, and the IT DGP with $\nu_{DGP} = 4$ and with sample size $n = 1000$. Figure 3.2 (respectively, Figure 3.3) displays the scatterplots of the financial data (respectively, the three toy data).

We compute the Gaussian and the IT estimators as in Section 3.3. We then calculate the squared Mahalanobis distances of the residuals and use a Kolmogorov–Smirnov test for deciding between the models. For the financial data, we have no predictor. We test the Gaussian (respectively the Student with three degrees of freedom, the Student with four degrees of freedom) null hypothesis. When testing one of the null hypotheses, we use the estimator corresponding to the null. Moreover, when the null hypothesis is Student, we use the corresponding degrees of freedom for computing the maximum likelihood estimator. We do reject the null hypothesis if the $p$-value is smaller than $\alpha = 5\%$. Note that we could adjust the level of $\alpha$ by taking into account multiple testing.

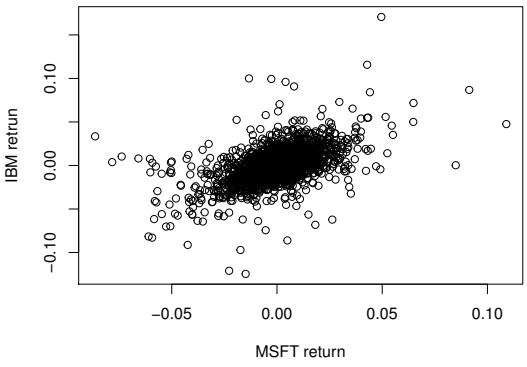


Figure 3.2: Financial data: scatterplot of returns .

Table 3.8 shows the $p$-values of these tests. For the simulated data, at the 5% level, we do not reject the Gaussian assumption when the DGP is Gaussian. Similarly, we do not reject the Student distribution with three (respectively, four) degrees of freedom when the DGP is the IT with degrees of freedom $\nu_{DGP} = 3$ (respectively, $\nu_{DGP} = 4$). For the financial data, we do not reject the Student distribution with three degrees of freedom, but we do reject the Gaussian distribution and the Student distribution with four degrees of freedom.

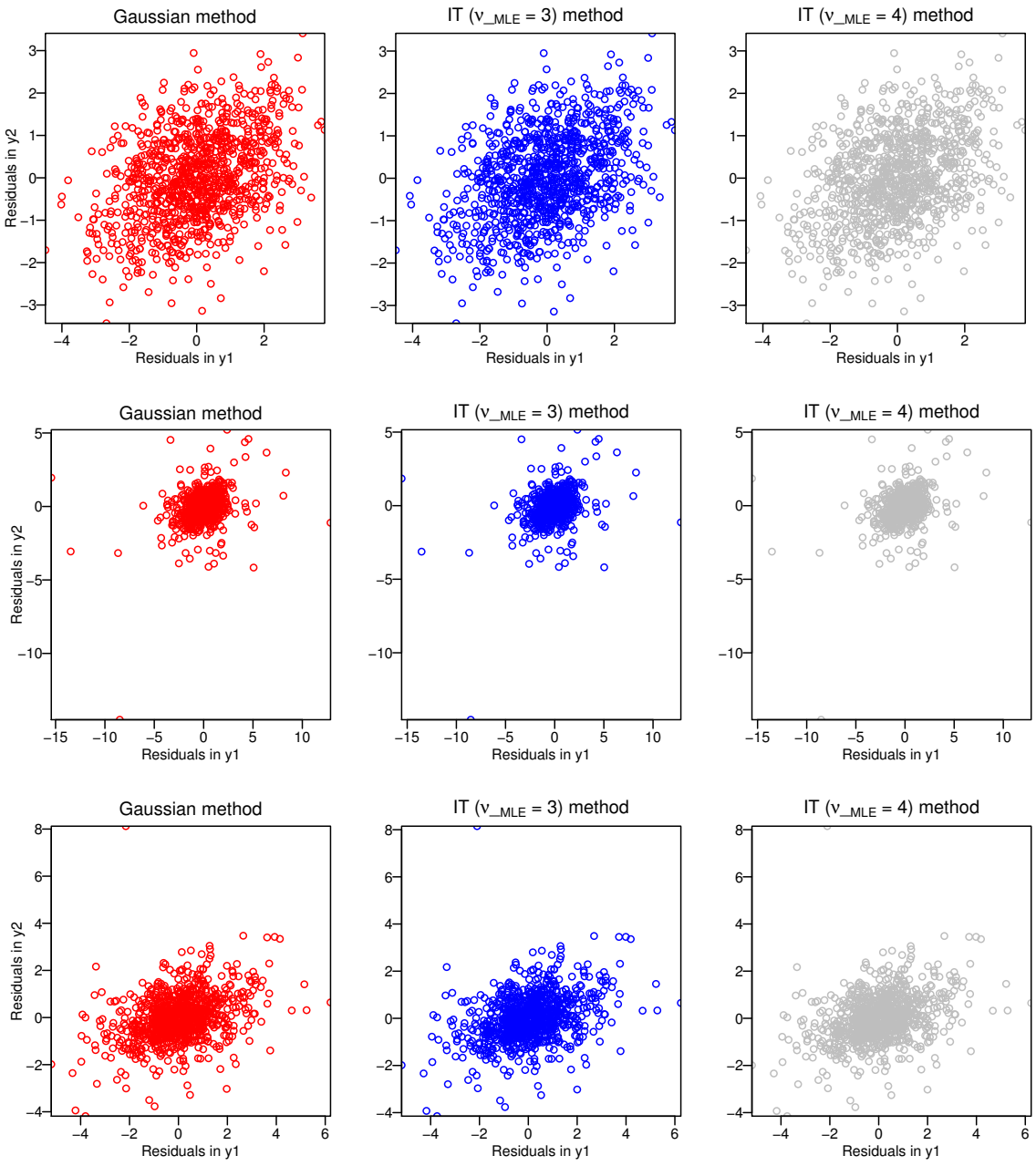


Figure 3.3: Toy data: scatterplots of residuals in the Gaussian DGP (respectively, the IT DGP with $\nu_{DGP} = 3$, the IT DGP with $\nu_{DGP} = 4$) on the first row (respectively, the second row, the third row).

Table 3.8: All datasets: the $p$-values of the Mahalanobis distances tests with the null hypothesis and the corresponding estimators.

| **Hypothesis $H_0$** | **Toy DGP** | | | **Financial Data** |
|---|---|---|---|---|
| **Methods** | **N** | **IT, $\nu_{DGP} = 3$** | **IT, $\nu_{DGP} = 4$** | |
| **N** | 0.546 | $2.2 \times 10^{-16}$ | $2.2 \times 10^{-16}$ | $2.2 \times 10^{-16}$ |
| **IT, $\nu_{MLE} = 3$** | $2.2 \times 10^{-16}$ | 0.405 | 0.033 | 0.882 |
| **IT, $\nu_{MLE} = 4$** | $2.2 \times 10^{-16}$ | 0.023 | 0.303 | 0.049 |

Figures 3.4 shows the Q-Q plots comparing the empirical quantiles of the Mahalanobis distances for the normal (respectively, the IT ($\nu_{MLE} = 3$), the IT ($\nu_{MLE} = 4$)) estimators on the horizontal axis to the theoretical quantiles of the Mahalanobis distances for the normal (respectively, the IT ($\nu_{MLE} = 3$), the IT ($\nu_{MLE} = 4$)) on the vertical axis for the financial data. These Q-Q plots are coherent with the results of the tests in Table 3.8. The IT model with three degrees of freedom fits our financial data well.

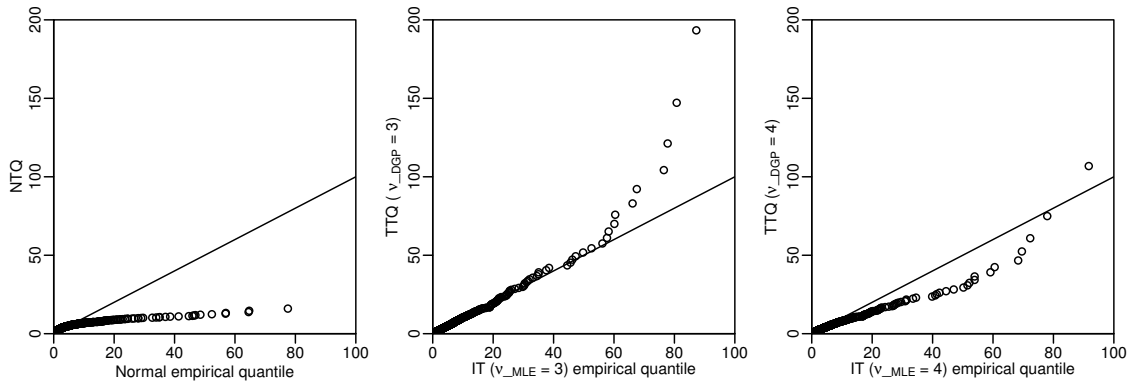


Figure 3.4: Financial data: Q-Q plots of the Mahalanobis distances for the normal, IT ($\nu_{MLE} = 3$), and IT ($\nu_{MLE} = 4$) estimators.

Figure 3.5 displays the Q-Q plots for the toy DGP: the Gaussian DGP in the first column, the IT DGP with $\nu_{DGP} = 3$ in the second column, and the IT DGP with $\nu_{DGP} = 4$ in the third column. The first row compares the empirical quantiles to the normal case quantiles, the second to the Student case quantiles with $\nu_{DGP} = 3$, and the third row to Student case quantiles with $\nu_{DGP} = 4$. The Q-Q plots on the diagonal confirm that the fit is good when the model is correct. The other Q-Q plots outside the diagonal correctly reveal a clear deviation from the hypothesized model.

To summarize the findings of this study, let us first say that there may be an abusive use of the Gaussian distribution in applications due to its simplicity. We have seen that considering the Student distribution instead is just slightly more complex, but feasible,

and that one can test this choice. Concerning the two Student models, we have seen that the UT model is simpler to fit than the IT model, but has limitations due to the fact that it assumes a single realization, which restricts the properties of the maximum likelihood estimators and prevents the use of tests against the other two models.

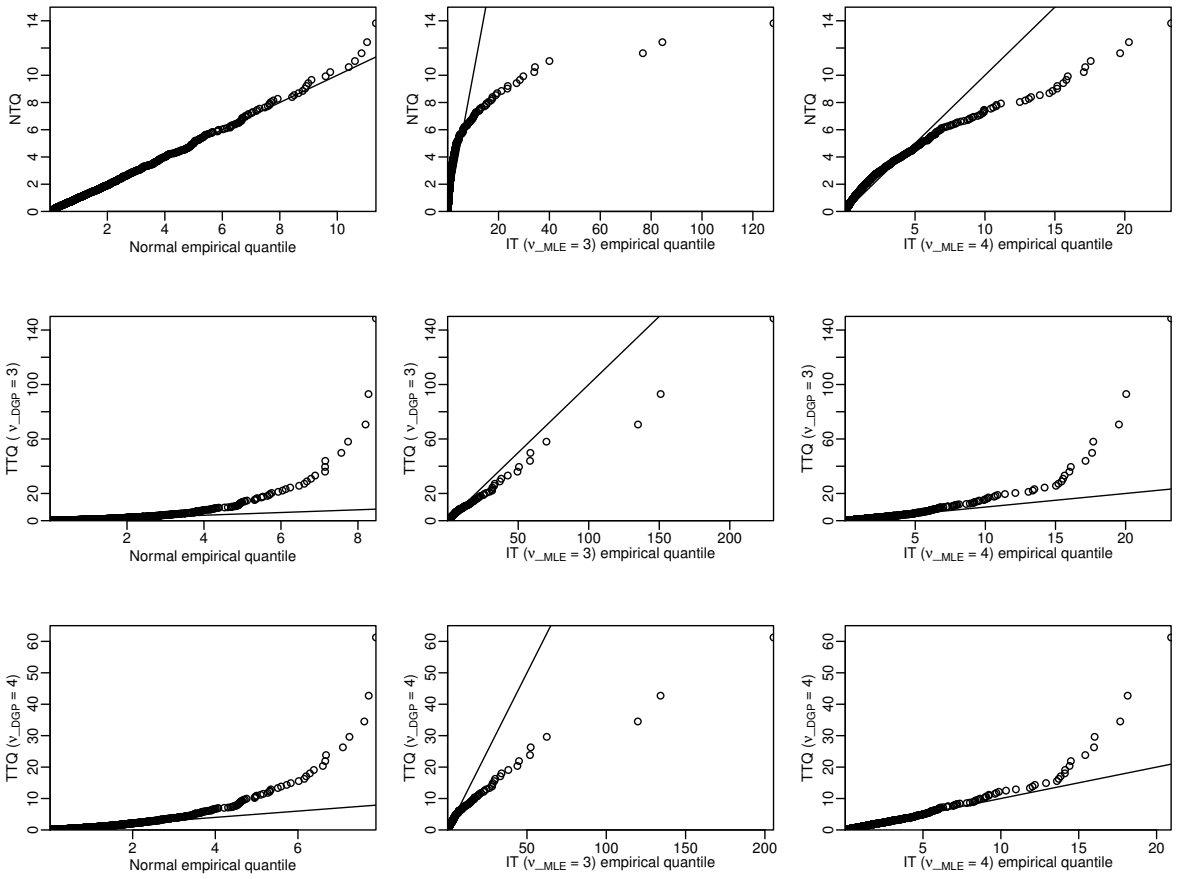


Figure 3.5: Toy data: Q-Q plots of the Mahalanobis distances of the residuals for the normal (respectively, the IT with $\nu_{DGP} = 3$, the IT with $\nu_{DGP} = 4$) case empirical quantiles against the normal (respectively, the IT with $\nu_{MLE} = 3$, the IT with $\nu_{MLE} = 4$) case theoretical quantiles in the first row (respectively, the second row, the third row).

## 3.5 Conclusions

We have compared three different models: the multivariate Gaussian model and two different multivariate Student models (uncorrelated or independent). We have derived some theoretical properties of the Student UT model and proposed a simple iterative reweighted algorithm to compute the maximum likelihood estimators in the IT model.

Our simulations show that using a multivariate Student IT model instead of a multivariate Gaussian model for heavy tail data is simple and can be viewed as a safeguard against misspecification in the sense that there is more to loose if the DGP is Student and one uses a Gaussian model than in the reverse situation. Finally, we have proposed some graphical tools and a test to choose between the Gaussian and the IT models. The IT model fits our finance dataset quite well. There is still work to do in the direction of improving the model selection procedure to overcome the fact that the parameters are estimated and hence the hypothetical distribution is only approximate. Let us mention that it is also possible to adapt our algorithm for the IT model to the case of missing data. We intend to work in the direction of allowing different degrees of freedom for each coordinate. It may be also relevant to consider an alternative estimation method by generalizing the one proposed in Kent et al. (1994) to the multivariate regression case. Finally, another perspective is to consider multivariate errors-in-variables models, which allow incorporating measurement errors in the response and the explanatory variables. A possible approach is proposed in Croux et al. (2010).

# Chapter 4

# Abstract

Political economists are interested by the impact of the characteristics of the geographical units on the outcome of an election. Because vote shares data often exhibit heavy tail behavior, a proposal found in the literature is to replace the Gaussian distribution by the Student distribution. In the previous chapter, we have seen two versions of the multivariate Student distribution the UT and the IT model. Because the restrictions imposed by the UT version (observations are a single realization) of multivariate distribution has consequences on the properties of the maximum likelihood estimators and prevent the use of tests, we will focus on the multivariate Independent Student (IT) distribution in this chapter. The classical CODA regression model assumes that a set of log ratios of the compositional vector follows a Normal distribution in coordinate space, thus defining the logistic normal distribution in the simplex. We describe how to adapt it to replace the normal distribution by the multivariate Independent Student error distribution thus defining the logistic Student distribution. We first recall facts about the multivariate Student distribution and the multivariate Normal distribution for the error term. We concentrate on building a CODA regression model using multivariate Student distributed error vectors with fixed degree of freedom. We compare this model to the multivariate Gaussian model. These models are fitted on French electoral data of the 2015 departmental elections for canton level data. Besides, in order to choose the best model, we apply the method of Chapter 3 for selecting between the Gaussian and the Student models based on the Mahalanobis distance. Finally, we present the impact of socio-economics factors using graphs of the predicted values of the dependent variable as a function of one specific variable of interest: two cases must be considered depending on whether the specific variable is classical or compositional.

# Résumé

Les économistes politiques s'intéressent à l'impact des caractéristiques des unités géographiques sur le résultat d'une élection. Étant donné que les données sur les votes partagés présentent souvent un comportement à queue lourde, une proposition trouvée dans la littérature consiste à remplacer la distribution Gaussienne par la distribution de Student. Dans le chapitre précédent, nous avons vu deux versions de la distribution multivariée de Student, le modèle UT et IT. Comme les restrictions imposées par la version UT (les observations forment une unique réalisation) de la distribution multivariée ayant des conséquences sur les propriétés des estimateurs du maximum de vraisemblance et empêchant l'utilisation de tests, nous allons nous concentrer sur la distribution multivariée de Student indépendant (IT) dans ce chapitre. Le modèle de régression CODA classique suppose qu'un ensemble de log-ratios du vecteur de composition suit une distribution Normale dans un espace de coordonnées, définissant ainsi la distribution logistique normale dans le simplexe. Nous décrivons comment l'adapter pour remplacer la distribution Normale par la distribution d'erreur multivariée des Student indépendants, définissant ainsi la distribution logistique de Student. Nous rappelons d'abord des faits sur la distribution multivariée de Student et la distribution multivariée Normale pour le terme d'erreur. Nous nous concentrons sur la construction d'un modèle de régression CODA utilisant des vecteurs d'erreur distribués de Student multivariés à degré de liberté fixe. Nous comparons ce modèle au modèle multivarié Gaussien. Ces modèles sont adaptés aux données électorales Françaises des élections départementales de 2015 pour les données cantonales. De plus, afin de choisir le meilleur modèle, nous appliquons la méthode du chapitre 3 pour choisir entre le modèle Gaussien et le modèle Student en fonction de la distance de Mahalanobis. Enfin, nous présentons l'impact des facteurs socio-économiques en utilisant des graphiques des valeurs prédites de la variable dépendante en fonction d'une variable d'intérêt spécifique: deux cas doivent être considérés selon que la variable spécifique est classique ou compositionnelle.

# Logistic Student distribution on the simplex with application to Political economy[1]

**Abstract**
In a multiparty election, the vote shares form a composition vector (mathematically, a vector belonging to a simplex). Political economists are interested by the impact of the characteristics of the geographical units on the outcome of an election. Because vote shares data often exhibit heavy tail behavior, we propose to use the Student error distribution instead of the Normal error distribution. We describe how to adapt the regression models for compositional data to the multivariate Student error distribution. For a Gaussian errors vector, the assumption of independent coordinates is equivalent to the assumption of correlated coordinates. However, this equivalence is no longer true when considering a multivariate Student distribution, which leads to two ways of using the multivariate Student distribution in the framework of a regression model: the multivariate independent Student (IT) distribution and the multivariate uncorrelated Student (UT) distribution. In this paper, we will concentrate on the IT distribution with fixed degrees of freedom. The Gaussian and Student models are then fitted on French electoral data of the 2015 departmental elections at the canton level and we use a method based on the Mahalanobis distance for model selection.

**Keywords**
compositional regression models, heavy tail distribution, independent multivariate Student distribution, logistic Student distribution, maximum likelihood estimator.

## 4.1 Introduction

Many authors in political economy concentrate on building models to understand the drivers of the outcome of a two-party electoral system (Beauguitte and Colange (2013), Ansolabehere and Leblanc (2008)). The outcome of an election can be influenced by the campaign strategies of the candidates and demographic factors such as age distribution, domain of activity distribution, rate of unemployment, and so on. In an interview

[1] Joint work with T. Laurent

with Time magazine, a group of Obama senior campaign advisers argue that there is an enormous data effort to support fundraising, micro-targeting TV ads and modeling of swing-state voters. In this work, we are interested in exploring the impact of the characteristics of the demographic and social factors on the outcome of the 2015 French departmental election for canton level data. The outcome of an election in a multi-party system consist of vectors whose components are proportions of votes per party. In what follows, our attention focuses on the relationships between votes shares and socio-economics factors such as age distribution, education levels distribution, domain of activity distribution, unemployment rate and so on, using a CODA (COmpositional Data Analysis) regression model.

Among papers concentrating on the relationship between socio-economic variables and election outcomes, Beauguitte and Colange (2013) carry out a linear regression model at three aggregation levels (polling stations, cities, and electoral districts) in France and show that the socio-economic variables are significant. Kavanagh et al. (2006) use geographically weighted regression, which produces a set of parameters estimates for each data point, i.e. for each electoral division. In the statistical literature, there are regression models adapted to share vectors including CODA models, but also Dirichlet models for example. These models, where the dependent and independent variables may be compositional variables (see Mert et al. (2018)). Honaker et al. (2002) and Katz and King (1999) assume that the territorial units yield independent observations. For example, Morais (2017) studies the impact of media investments on brand's market shares with a CODA regression model. Nguyen et al. (2018) propose a CODA multivariate regression model based on the normal distribution to study the impact of socio-economic factors on the 2015 French departmental election at the departmental level. However, election data often exhibit heavy tail behavior (see Katz and King (1999)). In order to cope with this heavy tail problem, a proposal found in the literature is to replace the Gaussian distribution by the Student distribution. There are two ways of using the Student distribution in a regression framework: the independent Student (IT) and the uncorrelated Student (UT) (see Kelejian and Prucha (1985)). Nguyen et al. (2019) show that the UT model is simpler to fit than the IT model, but the assumption of a single realization is a limitation which restricts the properties of the maximum likelihood estimators and prevents the use of tests against other models. Thus, we will concentrate here on the multivariate IT model and compare it to the multivariate Gaussian model in this chapter.

Section 4.2 describes the French departmental election data. Section 4.3 presents the considered multivariate regression models (multivariate Normal error vector and multivariate Independent Student (IT) error vector). In section 4.4, we recall the CODA principles and then build a CODA regression model using the independent multivariate Student distribution (IT) with known degrees of freedom for the error vector. As in Nguyen et al (2019), we perform a test based on the Mahalanobis distance to select between the multivariate Gaussian and the multivariate Student models in Section 4.5. Section 4.6 uses the vote shares predictions to investigate the relationships between the socio-economic factors and the outcome of this election.

Table 4.1: Data description

| Variable name | Description | Averages |
|---|---|---|
| Vote share | Left(L), Right(R), Extreme Right(XR) | 0.45, 0.29, 0.26 |
| Age | Age_1840, Age_4064, Age_65. | 0.29, 0.43, 0.28 |
| Diploma | SUP, <BAC, BAC. | 0.25, 0.57, 0.18 |
| Employment | AZ, BE, FZ, GU, OQ | 0.04, 0.08, 0.06, 0.43, 0.39 |
| unemp | The unemployment rate | 0.13 |
| nbvoter | Number of voter | 17908 |
| employ_evol | Mean annual growth rate of employment (2009-2014) | 0.39 |
| owner | The proportion of people who own assets | 0.63 |
| income | The proportion of people who pay income tax | 0.51 |
| foreign | The proportion of foreigners | 0.0048 |

## 4.2 Data

The Occitanie region has 283 cantons. However, some cantons in this region have at least one vote share (for one of three political parties) equal to zero. Eliminating these cantons results in a dataset with 207 cantons in Occitanie. Vote share data of the 2015 French departmental election with 207 cantons of Occitanie region in France are collected from the Cartelec website[2]. Corresponding socio-economic data (for 2014) are downloaded from the INSEE website[3]. Table 4.1 summarizes our data set.

Employment has five categories: AZ (agriculture, fisheries), BE (manufacturing industry, mining industry and others), FZ (construction), GU (business, transport and services) and OQ (public administration, teaching, human health). Diploma has three levels: <BAC for people with at most some secondary education, BAC for people with at least some secondary education and at most a high school diploma, and SUP for people with a university diploma. The Age variable has three levels: Age_1840 for people from 18 to 40 years old, Age_4064 for people from 40 to 64 years old, and Age_65 for elderly. For the vote share variable, the Cartelec website provides very detailed information. The number of political parties which present candidates at that election is higher than 15. For simplicity reasons, we aggregate the political parties into three main components: Left, Right and Extreme-Right[4]. Note that the averages in the last column of Table 4.1 are geometric means by component for the compositional variables.

[2]https://www.data.gouv.fr/fr/datasets/elections-departementales-2015-resultats-par-bureaux-de-vote/

[3]https://www.insee.fr/fr/statistiques

[4]for more details, see https://fr.wikipedia.org/wiki/Elections_départementales_francaises_de_2015

When compositional data have three components, they can be represented in a ternary diagram. For instance, the black points in Figure 4.1 show the vote shares of the 207 cantons for the Left and Right wings and the Extreme Right party. For compositional data, the correct way of defining the mean of a set of compositional vectors is through the vector of the geometric means of each component (see Pawlowsky-Glahn et al. (2015)). On Figure 4.1, the red triangle, corresponding to the geometric mean of vote shares, shows that average vote shares of the Left wing, the Right wing and the Extreme Right party are 45%, 29%, and 26% respectively. Figure 4.2 illustrates the positions of cantons in the Occitanie region on the ternary diagram: the components correspond to the three levels of the age variable, and the red triangle figures the geometric means of all cantons in the Occitanie region.

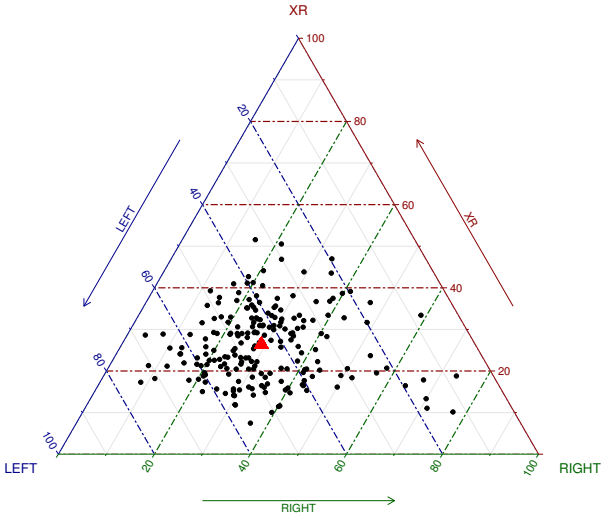


Figure 4.1: Vote shares for 207 cantons in the Occitanie region (black points) and their geometric mean of (red triangle).

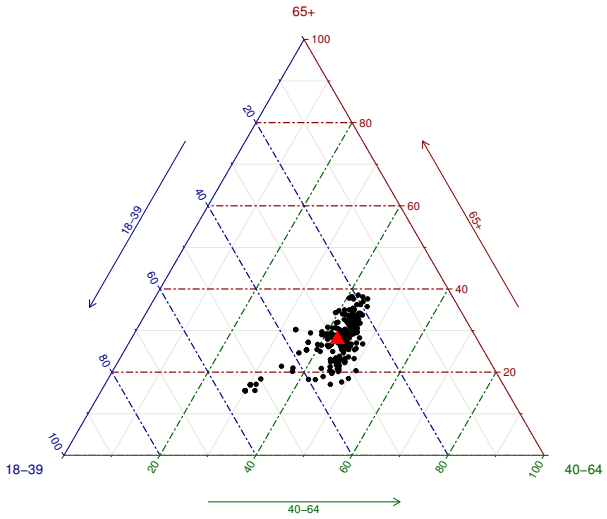


Figure 4.2: Components of Age in 207 cantons in the Occitanie region (black points) with their geometric mean (red triangle).

## 4.3 The multivariate regression models

In this section, we recall some classical facts about the multivariate Gaussian and the multivariate Student regression models. We consider a linear model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{4.1}$$

where $\mathbf{Y}$ is a $n \times L$ matrix of an $L$ dimensional dependent variable, $\mathbf{X}$ is a $n \times (K+1)$ matrix whose columns correspond to $K$ explanatory variables, $\boldsymbol{\beta}$ is the parameter matrix of size $(K+1) \times L$ and $\boldsymbol{\epsilon}$ is the error matrix of size $n \times L$.

### 4.3.1 Multivariate Normal error vector

Let us first consider model (4.1) with independent and identically distributed error vectors $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, following a multivariate normal distribution $\mathcal{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ with an $L$-vector of means equal to zero and an $L \times L$ covariance matrix $\boldsymbol{\Sigma}$. This model is denoted by $N$ and the subscript $N$ is used to denote the error terms $\boldsymbol{\epsilon}_{Ni}$, $i = 1, \ldots, n$ and the parameters $\boldsymbol{\beta}_N$ and $\boldsymbol{\Sigma}_N$ of the model. The maximum likelihood estimators of $\boldsymbol{\beta}_N$ and $\boldsymbol{\Sigma}_N$ are given by

$$\hat{\boldsymbol{\beta}}_N = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{4.2}$$

$$\hat{\boldsymbol{\Sigma}}_N = \frac{\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_{Ni}\hat{\boldsymbol{\epsilon}}_{Ni}^T}{n}, \tag{4.3}$$

where $\hat{\boldsymbol{\epsilon}}_{Ni} = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_N$ (see e.g. Theorem 8.4 from Seber (2009)).

### 4.3.2 Multivariate Independent Student error vector

Let us denote the $L$ dimensional dependent vector for the $i^{th}$ canton by:

$$\boldsymbol{\mathcal{Y}}_i = (Y_{i1}, \ldots, Y_{iL})^T.$$

For $K$ explanatory variables, the design matrix is of size $L \times (K+1)L$ and is given by:

$$\boldsymbol{\mathcal{X}}_i = \boldsymbol{I}_L \otimes \mathbf{X}_i^T$$

for $i = 1, \ldots, n$, with the $(K+1)$-vector $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{iK})^T$, $\boldsymbol{I}_L$ the identity matrix with dimension $L$ and $\otimes$ the usual Kronecker product. The parameter of interest is a $(K+1)L$ vector given by:

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_L^T)^T,$$

where $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{Kj})^T$, for $j = 1, \ldots, L$ and the $L$-vector of errors is denoted by:

$$\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iL})^T$$

for $i = 1, \ldots, n$. Given these notations, as in Nguyen et al. (2019), model (4.1) can be rewritten in vectorized form

$$\boldsymbol{\mathcal{Y}}_i = \boldsymbol{\mathcal{X}}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{4.4}$$

with $\mathbb{E}(\boldsymbol{\epsilon}_i) = 0$ and $i = 1, \ldots, n$. Note that $\boldsymbol{\mathcal{Y}} = (\boldsymbol{\mathcal{Y}}_1^T, \ldots, \boldsymbol{\mathcal{Y}}_n^T)^T$ is a $nL$ vector obtained by stacking the rows of $\boldsymbol{Y}$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_n^T)^T$ is a $nL$ vector and $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_1^T, \ldots, \boldsymbol{\mathcal{X}}_n^T)^T$ is a $nL \times (K+1)L$ matrix.

Let us consider model (4.4) with i.i.d. $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, following an independent multivariate Student (IT) distribution with $L$ dimensions and known degrees of freedom $\nu > 2$ denoted by $IT(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. In most of the literature on multivariate Student, the density is parametrized as a function of the scatter matrix $((\nu-2)/\nu)\boldsymbol{\Sigma}$ but we rather keep $\boldsymbol{\Sigma}$ as a parameter here.
The probability density function for the $L$-vector $\boldsymbol{\epsilon}$ is given by

$$p(\boldsymbol{\epsilon}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{f(\nu)}{\det(\boldsymbol{\Sigma})^{1/2}} \left[1 + \frac{1}{\nu - 2}(\boldsymbol{\epsilon} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\epsilon} - \boldsymbol{\mu})\right]^{-(\nu+p)/2}, \tag{4.5}$$

where $^T$ denotes the transpose operator, $f(\nu) = \dfrac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)(\nu-2)^{p/2}\pi^{p/2}}$ and $\Gamma$ is the Euler Gamma function. Following Prucha and Kelejian (1984), Nguyen et al. (2019) derive the maximum likelihood estimators for the IT model. The maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in the IT regression model satisfy the following implicit equations:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{IT} &= \left(\sum_{i=1}^{n} \hat{w}_{ITi} \boldsymbol{\mathcal{X}}_i^T \hat{\boldsymbol{\Sigma}}_{IT}^{-1} \boldsymbol{\mathcal{X}}_i\right)^{-1} \sum_{i=1}^{n} \hat{w}_{ITi} \boldsymbol{\mathcal{X}}_i^T \hat{\boldsymbol{\Sigma}}_{IT}^{-1} \boldsymbol{\mathcal{Y}}_i \\ \hat{\boldsymbol{\Sigma}}_{IT} &= \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ITi} \hat{\boldsymbol{\epsilon}}_{ITi} \hat{\boldsymbol{\epsilon}}_{ITi}^T \end{aligned} \tag{4.6}$$

$$\text{with} \quad \hat{\boldsymbol{\epsilon}}_{ITi} = \boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}_i \hat{\boldsymbol{\beta}}_{IT} \quad \text{and} \quad \hat{w}_{ITi} = \frac{\nu + L}{\nu - 2 + \hat{\boldsymbol{\epsilon}}_{ITi}^T \hat{\boldsymbol{\Sigma}}_{IT}^{-1} \hat{\boldsymbol{\epsilon}}_{ITi}}.$$

We use the iterative reweighted algorithm as in Nguyen et al. (2019) to estimate the covariate parameters and variance-covariance matrix.

## 4.4 Compositional regression models

### 4.4.1 Principles of compositional data analysis

#### Definition and operations

A composition $\mathbf{x}$ is a vector of $D$ parts of some whole which carries relative information. A $D$-composition $\mathbf{x}$ lies in the so-called simplex space $\mathbf{S}^D$ defined by:

$$\mathbf{S}^D = \{\mathbf{x} = (x_1, ..., x_D)' : x_j > 0, j = 1, ..., D; \sum_{j=1}^{D} x_j = 1\}$$

For $\mathbf{w} \in \mathbb{R}^{+D}$, let $\mathcal{C}(\mathbf{w}) = \left( \frac{w_1}{\sum_{j=1}^{D} w_j}, \cdots, \frac{w_D}{\sum_{j=1}^{D} w_j} \right)$ be the closure operation. The vector space structure of the simplex $\mathbf{S}^D$ is defined by the perturbation and powering operations:

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D), \ \mathbf{x}, \ \mathbf{y} \in \mathbf{S}^D$$
$$\lambda \odot \mathbf{x} = \mathcal{C}(x_1^{\lambda}, \ldots, x_D^{\lambda}), \ \lambda \text{ is a scalar}, \mathbf{x} \in \mathbf{S}^D.$$

The compositional matrix product, corresponding to the matrix product in the simplex, is defined by

$$\mathbf{B} \boxdot \mathbf{x} = \mathcal{C} \left( \prod_{j=1}^{D} x_j^{b_{1j}}, \cdots, \prod_{j=1}^{D} x_j^{b_{Lj}} \right)^T$$

where $\mathbf{B} = (b_{lj}), \ l = 1, \ldots, L, \ j = 1, \ldots, D$, is a parameter matrix such that the column vectors belong to $\mathbf{S}^D$,$\mathbf{j}_L^T \mathbf{B} = \mathbf{0}_D$, $\mathbf{B}\mathbf{j}_D = \mathbf{0}_L$, where $\mathbf{j}_L$ is a $L \times 1$ column vector of ones, and $\mathbf{j}_L^T$ is the transposed of $\mathbf{j}_L$.
The simplex $\mathbf{S}^D$ can be equipped with the Aitchison inner product (Aitchison (1985) and Pawlowsky-Glahn et al. (2015)) in order to define distances. The expected value $\mathbb{E}^{\oplus}\mathbf{Y}$ are also defined in Pawlowsky-Glahn et al. (2015).

**Log-ratio transformation**

Classical regression models cannot be used directly in the simplex because the constraints that the components are positive and sum up to 1 are not compatible with the usual distributional assumptions of normality. To overcome this difficulty, one way out is to use a log-ratio transformation from the simplex space $\mathbf{S}^D$ to the coordinate space $\mathbb{R}^{D-1}$. The classical transformations are alr (additive log-ratio transformation), clr (centered log-ratio transformation), and ilr (isometric log-ratio transformation). The coordinates in the clr transformed vector are linearly dependent, and the coordinates in the alr transformed vector are not compatible with the geometry (distance between the components in the simplex space is different from distance between the coordinates in the Euclidean space). For these reasons people generally use one of the ilr transformation for compositional regression models.
An isometric log-ratio transformation ilr is defined by:

$$\text{ilr}(\mathbf{x}) = \mathbf{V}_D^T \ln(\mathbf{x})$$

where the logarithm of $\mathbf{x}$ is understood componentwise, $\mathbf{V}_D^T$ is a transposed contrast matrix (see Pawlowsky-Glahn et al. (2015)) associated to a given orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by

$$\mathbf{V}_D = \text{clr}(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1}),$$

where clr denotes the centered log-ratio transformation (see Pawlowsky-Glahn et al. (2015)). As in Pawlowsky-Glahn et al. (2015) in our application, we use the following

contrast matrix for $D = 3$

$$\mathbf{V}_3 = \begin{bmatrix} 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}$$

This particular matrix defines the following ilr coordinates

$$\text{ilr}_1(\mathbf{x}) = \frac{1}{\sqrt{6}}(2\log x_1 - \ln x_2 - \log x_3) = \frac{2}{\sqrt{6}}\log\frac{x_1}{\sqrt{x_2 x_3}}$$
$$\text{ilr}_2(\mathbf{x}) = \frac{1}{\sqrt{2}}(\log x_2 - \log x_3) = \frac{1}{\sqrt{2}}\log\frac{x_2}{x_3}$$

The first ilr coordinate contains information about the relative importance of the first component $x_1$ with respect to the geometric mean of the second and the third components $g = \sqrt{x_2 x_3}$. The second ilr coordinate contains information about the relative importance of the second component $x_2$ with respect to the third component $x_3$. In our case, the first ilr coordinate opposes the Left wing to the group of the Right wing and the Extreme Right party and the second opposes the Right wing to the Extreme Right party. The inverse ilr transformation is given by:

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^*) = \mathcal{C}(\exp(V_D\mathbf{x}^*)) \text{ for } x^* \in \mathbb{R}^{D-1}$$

where the exponential of the vector $\mathbf{x}$ is understood componentwise.

### 4.4.2 Logistic Student regression models

Following the same construction as for the Logistic Normal distribution on the simplex (Pawlowsky-Glahn et al. (2015)), let us first define the Logistic Student distribution on the simplex.

**Definition 1.** *Given a random composition* $\mathbf{Y}$, *with sample space* $\mathbf{S}^D$, $\mathbf{Y}$ *is said to follow a Student distribution on* $\mathbf{S}^D$, *or logistic Student distribution, if the vector of random orthonormal coordinates* $\mathbf{Y}^* = ilr(\mathbf{Y})$ *follows a multivariate Student distribution* $IT(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \nu)$ *on* $\mathbb{R}^{D-1}$.

A change of ilr transformation would result in a different mean $\boldsymbol{\mu}^*$ and a different variance matrix $\boldsymbol{\Sigma}^*$ but their relationship with the former parameters remains the same as for the logistic Normal distribution (the argument of the proof of Theorem 6.19 in Pawlowsky-Glahn et al. (2015) simply goes through).

We use the notations defined in Table 4.2. Let $\mathbf{Y}_i$ be the compositional response value of the $i$th observation, $\mathbf{Y}_i \in \mathbf{S}^L$, and $\mathbf{X}_i^{(q)}$, $q = 1, \ldots, Q$, denotes the value of the $q$th compositional covariate for the $i$th observation, $\mathbf{X}_i^{(q)} \in \mathbf{S}^{D_q}$, $q = 1, \ldots, Q$, $Z_{ki}$, $k = 1, \ldots, K$, denotes the $k$th classical covariate of the $i$th observation. Let us first introduce the CODA regression model in the ilr coordinate space as follows:

$$\text{ilr}(\mathbf{Y}_i) = \mathbf{b_0}^* + \sum_{q=1}^{Q} \text{ilr}(\mathbf{X}_i^{(q)})\mathbf{B}_q^* + \sum_{k=1}^{K} Z_{ki}\mathbf{c}_k^* + \text{ilr}(\boldsymbol{\epsilon}_i) \tag{4.7}$$

Table 4.2: Notations

| **Variable** | **Notation** | **Coordinates** |
|---|---|---|
| Dependent | $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iL})$ | $\text{ilr}(\mathbf{Y}_i) = \mathbf{Y}_i^*$ |
| Compositional explanatory | $\mathbf{X}_i^{(q)} = (X_{i1}^{(q)}, \ldots, X_{iD_q}^{(q)})$ | $\text{ilr}(\mathbf{X}_{ip}^{(q)}) = \mathbf{X}_{ip}^{(q)*}$ |
| Classical explanatory | $Z_{ki}$ | |
| **General notations** | | |
| $L$ | Number of components of the dependent variable | |
| $i = 1, \ldots, n$ | Index of observations ($n = 95$ ) | |
| $q = 1, \ldots, Q$ | Index of compositional explanatory variables ($Q = 3$) | |
| $p = 1, \ldots, D_q$ | Index of the coordinates for the compositional explanatory variables | |
| $k = 1, \ldots, K$ | Index of classical explanatory variables ($K = 5$) | |

where $\text{ilr}(\mathbf{Y_i})$, $\text{ilr}(\mathbf{X}_i^{(q)})$ are the ilr coordinates of $\mathbf{Y}_i$, $\mathbf{X}_i^{(q)}$ $(q = 1, \ldots, Q)$ respectively, $\text{ilr}(\mathbf{Y_i}) \in \mathbb{R}^{L-1}$, $\text{ilr}(\mathbf{X}_i^{(q)}) \in \mathbb{R}^{D_q-1}$; $\mathbf{b_0}^*$, $\mathbf{B}_q^*$, $\mathbf{c}_k^*$ are the parameters in the coordinate space, and $\text{ilr}(\boldsymbol{\epsilon}_i)$ are the residuals in the coordinate space, $\text{ilr}(\boldsymbol{\epsilon}_i) \in \mathbb{R}^{L-1}$. The distributional assumption is $\text{ilr}(\boldsymbol{\epsilon})$ follows either the multivariate gaussian (N) distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_N$ or the independent multivariate Student (IT) distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{IT}$.
Denoting by $\bigoplus$ the addition in the simplex (see Pawlowsky-Glahn et al. (2015)), this regression model (4.7) can be written in the simplex as

$$\mathbf{Y}_i = \mathbf{b_0} \bigoplus_{q=1}^{Q} \mathbf{B}_q \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_{ki} \odot \mathbf{c}_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{4.8}$$

where $\mathbf{b_0}, \mathbf{B}_1, \ldots, \mathbf{B}_Q, \mathbf{c}_1, \ldots, \mathbf{c}_K$ are the parameters satisfying $\mathbf{b_0} \in \mathbf{S}^L$, $\mathbf{B}_q \in \mathbf{S}^{D_q}$, $q = 1, \ldots, Q$, $\mathbf{c}_k \in \mathbf{S}^L$, $k = 1, \ldots, K$, $\mathbf{j}_L^T \mathbf{B}_q = \mathbf{0}_{D_q}$, $\mathbf{B}_q \mathbf{j}_{D_q} = \mathbf{0}_L$. The distributional assumption is that $\boldsymbol{\epsilon}_i \in \mathbf{S}^L$ follows either the multivariate logistic Gaussian (N) distribution (see Aitchison (1985)) or the independent multivariate logistic Student (IT) distribution defined by (4.5).

### 4.4.3 Application to political economy

We now fit a CODA regression model describing the impacts of socio-economic factors on vote shares in the 2015 French departmental election using the data described in Section 4.2. We estimate the parameters of model (4.7) using the Multivariate Normal and the Independent Student distribution for the error vector. This is equivalent to estimate the parameters of model (4.8) with a Logistic Normal and a Logistic Student distribution.

We first estimate the models using all the explanatory variables described in Table 4.1. Then we remove the variables employ_evol, owner and foreign which are not significant and we estimate the models again with the remaining variables. Table 4.3 shows the estimated parameters for both the Gaussian model and the Independent Student model with $\nu = 4$. The population size (nbvoter) may be used to take into account the fact that cantons are not comparable in size either through the mean or the variance or both (we choose to include it in the mean value here). The distribution of education level, of age, the employment sector, the unemployment rate and the proportion of people who pay income tax really have a significant impact on the result of the French departmental election in 2015. Table 4.3 also shows that for both models, when the ratio of people between the level 18-40 and the level >40 increases, the ratio of vote shares between Left and [Right+Extreme Right] decreases and the ratio of vote shares between Right and Extreme Right increases. However, when the ratio of people between the level 40-64 and the level >64 increases, the ratio of vote shares between Right and Extreme Right will decrease. Besides, we see that when the ratio of people who pay income tax goes up, the ratio between Left and [Right+Extreme Right] will go up faster than the ratio between Right and Extreme Right. When the unemployment rate increases, the ratio between Right and Extreme Right increases.

Table 4.3: Multivariate Gaussian and Student regression models with compositional and classical explanatory variables

| | *Gaussian model* | | *Student model, $\nu = 4$* | |
|---|---|---|---|---|
| | y_ilr[, 1] | y_ilr[, 2] | y_ilr[, 1] | y_ilr[, 2] |
| Constant | $-3.69(0.89)^{***}$ | $-2.66(0.45)^{***}$ | $-4.00(1.33)^{***}$ | $-2.70(0.68)^{***}$ |
| diplome_ilr1 | $-1.27(0.50)^{**}$ | $-0.29(0.25)$ | $-1.40(0.75)^{*}$ | $-0.35(0.38)$ |
| diplome_ilr2 | $-0.03(0.61)$ | $-0.90(0.30)$ | $+0.54(0.90)$ | $-0.86(0.46)^{*}$ |
| employ_ilr1 | $-0.18(0.14)$ | $-0.13(0.07)^{*}$ | $-0.21(0.20)$ | $-0.18(0.10)^{*}$ |
| employ_ilr2 | $+0.49(0.16)^{***}$ | $-0.03(0.08)$ | $+0.38(0.24)$ | $+0.09(0.12)$ |
| employ_ilr3 | $-0.21(0.11)^{*}$ | $+0.01(0.06)$ | $-0.18(0.17)$ | $-0.05(0.09)$ |
| employ_ilr4 | $+0.21(0.06)^{***}$ | $+0.01(0.03)$ | $+0.14(0.09)$ | $+0.02(0.05)$ |
| age_ilr1 | $-1.14(0.37)^{***}$ | $+1.00(0.18)^{***}$ | $-0.81(0.55)$ | $+1.13(0.28)^{***}$ |
| age_ilr2 | $+0.48(0.30)$ | $-1.33(1.15)^{***}$ | $+0.44(0.46)$ | $-1.41(0.23)^{***}$ |
| unemp | $-0.05(2.27)$ | $+9.70(1.14)^{***}$ | $+1.80(3.37)$ | $+9.58(1.73)^{***}$ |
| income | $+4.30(0.89)^{***}$ | $+1.14(0.06)^{***}$ | $+4.10(1.34)^{***}$ | $+0.95(0.68)$ |
| nbvoter | +2e-06(5e-06) | $+1e\text{-}05(2e\text{-}05)^{***}$ | −6e-07(3e-03) | $+1e\text{-}05(2e\text{-}03)^{***}$ |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

## 4.5 Model selection

Table 4.3 shows similar results for the Gaussian and Student models for certain covariates. Nguyen et al. (2019) propose a methodology to select between the Gaussian and independent Student models based on the Mahalanobis distance. We now apply this

| **Hypothesis $\mathbf{H}_0$** | **P-value** |
|---|---|
| N | 0.03 |
| IT, $\nu_{MLE} = 3$ | 0.63 |
| IT, $\nu_{MLE} = 4$ | 0.72 |

Table 4.4: The p-values of the Mahalanobis distances tests with the null hypothesis and the corresponding estimators.

method to select between the two models.
For an $L$-dimensional random vector $\mathbf{Y}$, with mean $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$, the squared Mahalanobis distance is defined by:

$$d^2 = (\mathbf{Y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu})$$

According to Nguyen et al. (2019), we can test whether the Mahalanobis distances follow a chi-square distribution for testing the normality of the data and whether the Mahalanobis distances follow the Fisher distribution for testing the Student distribution of the data. As in Nguyen et al. (2019), we perform some Kolmogorov−Smirnov tests in order to test different null hypothesis: Gaussian, Independent Student with three and four degrees of freedom. We also produce Q-Q plots of the Mahalanobis distances with respect to the chi-square and the Fisher distribution. Note that we use the estimator corresponding to the null hypothesis when testing each null hypotheses. We do reject the null hypothesis if the p-value is smaller than $\alpha = 5\%$.

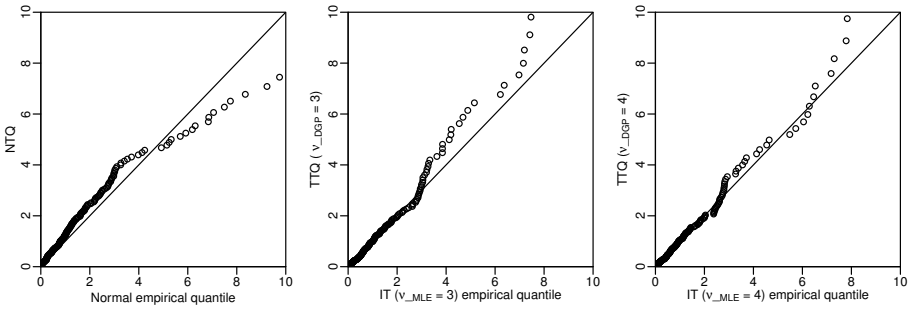


Figure 4.3: Q-Q plots of the Mahalanobis distances for the normal, IT ($\nu_{MLE} = 3$), and IT ($\nu_{MLE} = 4$) estimators

Table 4.4 shows the p-values of three tests. We do reject the Gaussian distribution and we do not reject the Student distribution with three degrees or four degrees of freedom. Figure 4.3 illustrates the Q-Q plots comparing the empirical quantiles of the Mahalanobis distances for the normal (respectively, the IT ($\nu_{MLE} = 3$), the IT ($\nu_{MLE} = 4$)) estimators on the horizontal axis to the theoretical quantiles of the Mahalanobis distances for the normal (respectively, the IT ($\nu_{MLE} = 3$), the IT ($\nu_{MLE} = 4$)) on the

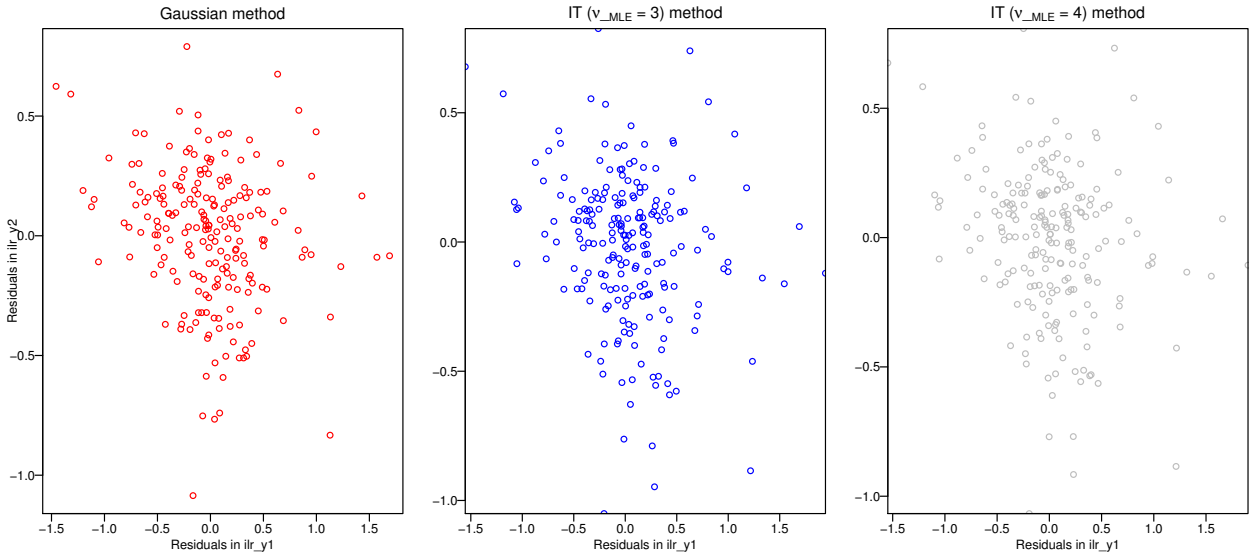


Figure 4.4: Scatterplots of residuals for the normal, IT ($\nu_{MLE} = 3$), and IT ($\nu_{MLE} = 4$) estimators

vertical axis. These Q-Q plots are coherent with the results of the tests in Table 4.4. The IT model with three or four degrees of freedom fits our data.

## 4.6 Vote shares predictions

The interpretation of the regression model parameters in the simplex is not so straightforward (see Morais et al. (2017)). Thus, in this section, we focus on illustrating graphically the relationship between the predicted vote shares and the explanatory variables to understand the impact of the socio-economic factors on the outcome of the election in France.

As in Nguyen et al. (2018), the prediction of the dependent variable for the above models for the $i^{th}$ canton is given by:

$$\hat{\mathbf{Y}}_i = \hat{\mathbf{b}}_{\mathbf{0}} \bigoplus_{q=1}^{Q} \hat{\mathbf{B}}_q \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_{ki} \odot \hat{\mathbf{c}}_k \quad i = 1, ..., n \tag{4.9}$$

where $\hat{\mathbf{b}}_{\mathbf{0}}, \hat{\mathbf{B}}_q$ and $\hat{\mathbf{c}}_k$ are the estimated parameters.

As in Nguyen et al. (2018), we now focus on graphing the predicted values of the dependent variable as a function of one specific variable of interest: two cases must be considered depending on whether the specific variable is classical or compositional. In both cases, we create a grid of potential values of the specific explanatory and fix the

other explanatory variables at the values they take for one selected point of the dataset (we repeat for several selected points). For the sake of simplicity let us take $L = 3$.

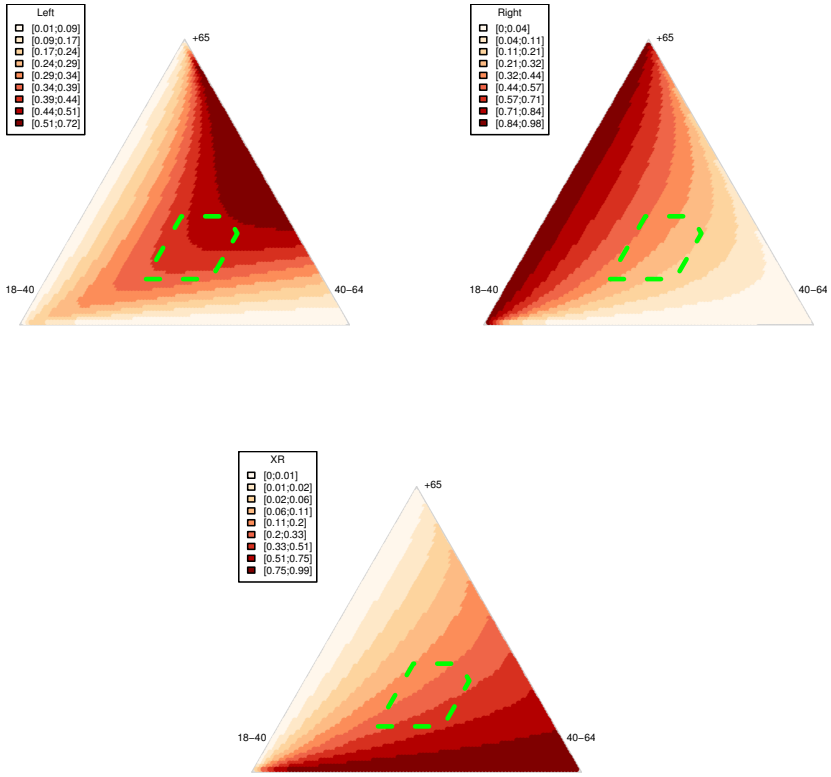


Figure 4.5: The vote share prediction (the green dotted rectangular show the position of observed age of three parties)

We now illustrate the impact of a compositional variable (Age) and a classical variable (unemployment rate) on the vote shares. We fix the values of covariates at their means except Age (resp. unemployment rate) and create a grid of fictive values of Age (resp. of unemployment rate) in order to predict the vote share according to Age (resp. unemployment rate). Then we compute the predicted shares at each of these grid points. Figure 4.5 illustrates the vote share predictions according to the Age variable. It shows that the people aged more than 64 vote for the Left wing, a part of the youngest people and a part of the oldest people vote for the Right wing and people from 40 to 64 years old vote for the Extreme Right party. On Figure 4.6, we color the triangle according to the winning party (the party with the highest proportion of votes. The Figure shows that the Left wing is the winning party because the observed points for Age (inside the

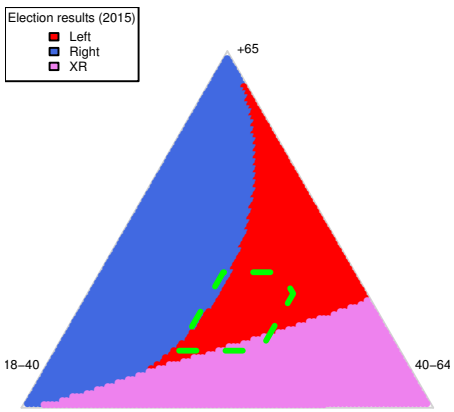


Figure 4.6: The vote share prediction (the green dotted rectangular show the position of observed age)

dotted green region) fall in the Left wing region (red part). For a classical variable - unemployment rate, Figure 4.7 shows the vote share prediction on the ilr coordinate space (on the left hand side) and the vote share prediction in the simplex (on the right hand side) according to unemployment rate. We see that these vote share predictions are neither mototone nor linear. Moreover, the vote share of Extreme Right party will go up whereas the vote share of the Left wing and the vote share of the Right wing will go down when the unemployment rate increases.

## 4.7 Conclusion

We have presented a CODA regression model with the independent multivariate Student distribution and it fits the election data quite well. This model seems to be useful for modeling heavy tailed data. We have performed Kolmogorov–Smirnov tests to select the final model. We have also illustrated the impact of compositional and classical explanatory variables on the outcome of the election by graphical techniques.

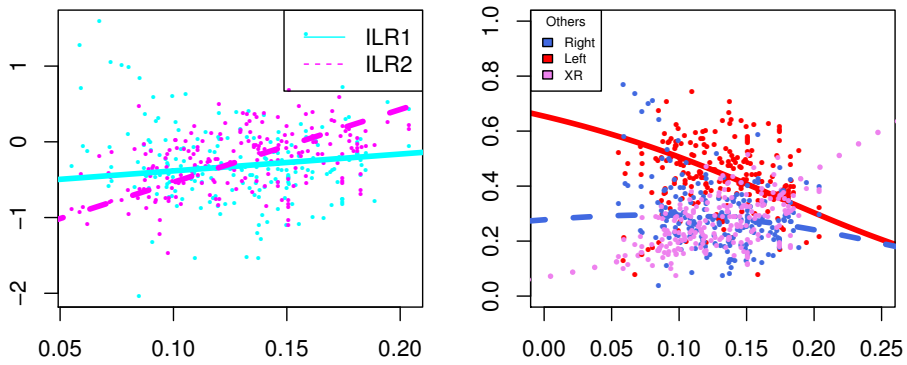


Figure 4.7: Predictions of vote shares on the Euclidean space (on the left) and on the simplex (on the right)

# Chapter 5

# Abstract

The outcome of an election in a multiparty system is a vector called composition with positive components which sum up to a constant. Besides, the vote shares are observed at a regional unit level (in our application, it is the canton level in the Occitanie region in France). These data exhibit at the same time the characteristics of compositional data as well as the characteristics of spatial data. Thus, the challenge is to build a multivariate regression model to accomodate these two aspects. For the compositional data issue, we build a multivariate regression model in which the dependent variable is a compositional variable and the explanatory variables may contain compositional and classical variables (or both of them). For the spatial issue, some authors use geostatistical approaches whereas others use rather an approach for areal data known as simultaneous spatial autocorrelation models in spatial econometrics.

We combine the compositional regression model with a multivariate spatial econometric approach in coordinate space. In what follows, after recalling some principles of compositional data, we introduce a new operation to write our model in the simplex. We then develop a simultaneous spatial autoregressive model for compositional data which allow for both spatial correlation and correlations across equations. The explanatory variables in this model are classical variables. In order to estimate the parameters, we adopt an instrumental variable (IV) method for spatial autoregressive model in a multivariate setting as in Kelejian and Prucha (2004). We devise a simulation to compare the relative root mean square error (RRMSE) of the parameters estimates under several data generating processes (DGP) for the two-stages least square (2SLS) and three-stages least squares (3SLS). We illustrate this approach with an example in political economy.

# Résumé

Le résultat d'une élection dans un système multipartite est un vecteur appelé composition avec des composantes positives dont la somme est une constante. De plus, les parts de vote sont observées au niveau de l'unité régionale (dans notre application, il s'agit du niveau du canton dans la région Occitanie en France). Ces données présentent à la fois les caractéristiques des données de composition et celles des données spatiales. Le défi consiste donc à élaborer un modèle de régression multivarié pour tenir compte de ces deux aspects. Pour le problème des données de composition, nous construisons un modèle de régression multivarié dans lequel la variable dépendante est une variable de composition et les variables explicatives peuvent contenir des variables de composition et des variables classiques (ou les deux). Pour la question spatiale, certains auteurs utilisent l'approche géostatistique, tandis que d'autres utilisent plutôt une approche pour les données surfaciques, appelée modèles d'autocorrélation spatiale simultanée en économétrie spatiale. Nous combinons le modèle de régression compositionnelle avec une approche économétrique spatiale multivariée dans un espace de coordonnées. Dans ce qui suit, après avoir rappelé quelques principes de données de composition, nous introduisons une nouvelle opération pour écrire notre modèle dans le simplexe. Nous développons ensuite un modèle autorégressif spatial simultané pour les données de composition qui permet à la fois de prendre en compte la corrélation spatiale et les corrélations entre équations. Les variables explicatives de ce modèle sont des variables classiques. Afin d'estimer les paramètres, nous adoptons une méthode de variable instrumentale (IV) pour le modèle spatial autorégressif dans un contexte multivarié comme dans Kelejian et Prucha (2004). Nous concevons une simulation pour comparer la racine carré de l'erreur quadratique moyenne relative (RRMSE) des estimations de paramètres sous plusieurs processus de génération de données (DGP) pour les moindres carrés à deux étages (2SLS) et les moindres carrés à trois étages. (3SLS). Nous illustrons cette approche avec un exemple en économie politique.

# A spatial autoregressive model for compositional data[1][2]

**Abstract.** In an election, the vote shares by party on a given subdivision of a territory form a vector with positive components adding up to 1 called a composition. Using a conventional multiple linear regression model to explain this vector by some factors is not adapted for at least two reasons. The first one is the existence of the constraint on the sum of the components and the second one is the assumption of statistical independence across territorial units which may be questionable due to potential spatial autocorrelation. We develop a simultaneous spatial autoregressive model for compositional data which allows for both spatial correlation and correlations across equations. We propose an estimation method based on two-stage and three-stage least squares. We illustrate the method with simulations and with a data set from the 2015 French departmental election.

**Keywords.** multivariate spatial autocorrelation, spatial weight matrix, three-stage least squares, two-stage least squares, simplex, electoral data.

## 5.1 Introduction

Some data present simultaneously the characteristics of compositional data (vectors with positive components adding up to a constant and conveying relative information) as well as the characteristics of spatial data (presence of spatial heterogeneity and spatial dependence). For example, land cover data contain information about different land use shares and the statistical unit is a subdivision of a territory; among the many papers that treat this type of data see Leininger et al. Leininger et al. (2013), Overmars et al. Overmars et al. (2003), Yoshida and Tsutsumi Yoshida and Tsutsumi (2018) and Pirzamanbein et al. Pirzamanbein et al. (2015). Another instance is in geochemistry where data consist of composition of mineral deposits into chemical elements at different locations in geographical space, see for example Rubio et al. Rubio et al. (2016) who study sediments in an artic lake or Filzmoser et al. Filzmoser et al. (2010) who examine the Kola moss layer composition data from the R package StatDa. This is also the case in political economy for electoral data containing the vote shares by party in a

[1] Submitted to Spatial Statistics

[2] Joint work with A. Ruiz-Gazen, C. Thomas-Agnan and T. Laurent

multiparty election for a list of administrative subdivisions of a territory as in Katz and King Katz and King (1999) or for data about turnout rates as in Borghesi and Bouchaud Borghesi and Bouchaud (2010). Other examples include the distribution of temperature data at weather stations as in Salazar et al. Salazar et al. (2015), the distribution of benthic macroinvertebrates at sampling stations in the Delaware Bay in Billheimer at al. Billheimer et al. (1997).

The challenge for modelling such data is to accomodate at the same time their compositional and spatial nature. Concerning the spatial aspect, some authors use a geostatistical approach whereas others use rather an approach for areal data known as simultaneous spatial autocorrelation models in spatial econometrics (see LeSage and Pace LeSage and Pace (2009)). For the geostatistical approach, let us mention Pawlowsky and Burger Pawlowsky and Burger (1992), Pawlowsky et al. Pawlowsky-Glahn and Egozcue (2016), Rubio et al. Rubio et al. (2016), Martins et al. Martins et al. (2016), Billheimer et al. Billheimer et al. (1997). Given the nature of our application, areal data rather than point data, we concentrate here on the spatial econometrics approach. Note that the compositional vector being the dependent variable, we will need spatial econometrics models for multivariate dependent variable as in Kelejian and Prucha Kelejian and Prucha (2004). We develop a simultaneous spatial autoregressive model for compositional data which allows for both spatial correlation and correlations across equations. We propose an estimation method based on two-stage (S2SLS) and three-stage (S3SLS) least squares.

In Section 2, we first recall some classical facts adapted to work with compositional data. We then introduce a new operation which will be necessary later to write our model in a simplex fashion and study its properties. In Section 3, we recall facts about the definition and estimation of simultaneous autoregressive models for multivariate output spatial data and combine with the tools of Section 2 to define our model for spatio-compositional data. Section 4 presents some simulations to evaluate the quality of the S2SLSL and S3SLS methods in the multivariate case. Sections 5 presents an application to election results with the question of the impact of socio-economic variables on parties vote shares with a data set from the 2015 French departmental election. Section 6 concludes.

## 5.2 Definitions and notations in compositional data analysis

A $D$-composition $\mathbf{u}$ is a vector of $D$ parts of some whole which carries relative information and therefore can be represented in the so-called simplex space $\mathbf{S}^D$ defined by:

$$\mathbf{S}^D = \left\{ \mathbf{u} = (u_1, \ldots, u_D)' : u_m > 0, m = 1, ..., D; \sum_{m=1}^{D} u_m = 1 \right\}$$

For any vector $\mathbf{w} \in \mathbb{R}^{+D}$, the closure operation is defined by

$$\mathcal{C}(\mathbf{w}) = \left( \frac{w_1}{\sum_{m=1}^{D} w_m}, \cdots, \frac{w_D}{\sum_{m=1}^{D} w_m} \right).$$

Let us recall the usual operations used to define a vector structure on the simplex space.

1. $\oplus$ is the perturbation operation, corresponding to the addition in $\mathbb{R}^D$:

$$\mathbf{u} \oplus \mathbf{v} = \mathcal{C}(u_1 v_1, \ldots, u_D v_D), \ \mathbf{u}, \mathbf{v} \in \mathbf{S}^D$$

2. $\odot$ is the power transformation, corresponding to the scalar multiplication in $\mathbb{R}^D$:

$$\lambda \odot \mathbf{u} = \mathcal{C}(u_1^{\lambda}, \ldots, u_D^{\lambda}), \ \lambda \text{ is a scalar}, \mathbf{u} \in \mathbf{S}^D$$

Moreover, the compositional product of a matrix by a vector denoted by $\boxdot$ is defined as follows

$$\mathbf{B} \boxdot \mathbf{u} = \mathcal{C} \left( \prod_{m=1}^{D} u_m^{b_{1m}}, \cdots, \prod_{m=1}^{D} u_m^{b_{Dm}} \right)^T$$

where $\mathbf{u} \in \mathbf{S}^D$, $\mathbf{B} = (b_{lm})$ with $l = 1, \ldots D$, $m = 1, \ldots, D$ is a $D \times D$ matrix and $^T$ is the transposition operator.

The simplex $\mathbf{S}^D$ can also be equipped with the compositional/Aitchison inner product (see Aitchison Aitchison (1985) and Pawlowsky-Glahn et al. Pawlowsky-Glahn et al. (2015)) in order to define distances. The expected value of a simplex valued random variable $\mathbf{Y}$, denoted by $\mathbb{E}^{\oplus}\mathbf{Y}$, is defined in Pawlowsky-Glahn et al. Pawlowsky-Glahn et al. (2015).

The analysis of compositional data makes use of log-ratio transformations which map the simplex $\mathbf{S}^D$ to $\mathbb{R}^q$ (where most often $q = D - 1$) because of their degree 0 homogeneneity (scale invariance). The classical ones are the additive log-ratio (alr), the centered log-ratio (clr) and the isometric log-ratio (ilr) transformations. In this paper, we will mainly use some ilr transformations. Because it is needed to define the ilr, let us first recall the definition of the clr transformation of a vector $\mathbf{u} \in \mathbf{S}^D$

$$\mathrm{clr}(\mathbf{u}) = \left( \ln \frac{u_m}{g(\mathbf{u})} \right)_{m=1,\ldots,D} \text{ with } g(\mathbf{u}) = \sqrt[D]{u_1 \cdot u_2 \cdots u_D}.$$

Let $\mathbf{V}_D$ be a $D \times (D-1)$ contrast matrix (Pawlowsky-Glahn et al. Pawlowsky-Glahn et al. (2015)) associated to a given orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by

$$\mathbf{V}_D = \mathrm{clr}(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1}),$$

where clr is understood componentwise. For each such matrix $\mathbf{V}_D$, an isometric log-ratio transformation (ilr) is then defined by:

$$\mathbf{u}^* = \mathrm{ilr}(\mathbf{u}) = \mathbf{V}_D^T \ln(\mathbf{u}) = \ln(\mathbf{u}) \mathbf{V}_D$$

where the logarithm of $\mathbf{u} \in \mathbf{S}^D$ is understood componentwise.

Since our data will be made of samples of composition vectors, we will store them in a $n \times D$ matrix $\mathbf{Y} = (\mathbf{Y}_{il})$, $(i = 1, \ldots, n,\ l = 1, \ldots, D)$ and each row of this matrix, denoted by $\mathbf{Y}_{i.}$, is a compositional vector of $\mathbf{S}^D$. $\mathbf{Y}_{.l}$, $l = 1, \ldots, D$ will denote the $l^{th}$ column of $\mathbf{Y}$ and we have

$$\mathbf{Y} = (\mathbf{Y}_{1.}^T, \ldots, \mathbf{Y}_{n.}^T)^T = (\mathbf{Y}_{.1}, \ldots, \mathbf{Y}_{.D}).$$

Let us define an extension of the ilr transformation of a matrix $\mathbf{Y}$ by

$$\text{ilr}(\mathbf{Y}) = \ln(\mathbf{Y})\mathbf{V}_D = \begin{bmatrix} \text{ilr}(\mathbf{Y}_{1.}) \\ \vdots \\ \text{ilr}(\mathbf{Y}_{n.}) \end{bmatrix}$$

Note that $\text{ilr}(\mathbf{Y})$ is a $n \times (D-1)$ matrix.

In spatial econometrics models, spatial weight matrices are used to specify the neighborhood structure. For $n$ spatial locations, the elements $w_{ij}$ of the $n \times n$ matrix $\mathbf{W}$ are measures of proximity between locations $i$ and $j$. See for example Bivand et al. Bivand et al. (2008) for different specifications. These matrices determine a covariance model for the data vector and play a role similar to the spatial variogram in geostatistics. For such a matrix and a given data vector $\mathbf{Z}$ of size $n$, the lagged vector $\mathbf{WZ}$ contains averages of the values of the variable $\mathbf{Z}$ in neighboring locations when $\mathbf{W}$ is row normalized. In our case, we need to apply such an operation to each column of the data matrix $\mathbf{Y}$ and we wish that the application of this process to each column of $\mathbf{Y}$ results in a matrix in the same space as the original one $(\mathbf{S}^D)^n$. As usual in compositional data analysis we use the principle of working in coordinates (log-transformed data) and inverting the transformation to go back to the simplex. We thus define the following operation.

**Definition 2.** Let $\mathbf{W}$ be a $n \times n$ matrix. The operation $\triangle$ is a map from the cartesian product of simplex spaces $(\mathbf{S}^D)^n$ to itself defined by

$$\text{ilr}(\mathbf{W} \triangle \mathbf{Y}) = \mathbf{W}\text{ilr}(\mathbf{Y}) = \mathbf{W} \ln(\mathbf{Y})\mathbf{V}_D \tag{5.1}$$

where $\mathbf{V}_D$ is a $D \times (D-1)$ contrast matrix.

Note that $(\mathbf{W} \triangle \mathbf{Y}) \in (\mathbf{S}^D)^n$ and $\mathbf{W}\text{ilr}(\mathbf{Y}) \in (\mathbb{R}^{(D-1)})^n$. This operation satisfies the following properties:

**Proposition 4.** Let $\mathbf{Y}$ be a $n \times D$ matrix such that each row, denoted by $\mathbf{Y}_{i.}$, $i = 1, \ldots, n$ is a compositional vector in $\mathbf{S}^D$. Let $\mathbf{W} = (W_{ij})$, $(i, j = 1, \ldots, n)$, a $n \times n$ matrix and $\alpha \in \mathbb{R}$. We have

1. $\mathbf{W} \triangle (\alpha \odot \mathbf{Y}) = \alpha \odot (\mathbf{W} \triangle \mathbf{Y})$.
2. $\text{ilr}(\mathbf{W} \triangle (\alpha \odot \mathbf{Y})) = \alpha \mathbf{W}\text{ilr}(\mathbf{Y}) = \alpha\ \text{ilr}(\mathbf{W} \triangle \mathbf{Y})$.

3. $(\mathbf{W}\triangle\mathbf{Y})_{i.} = \mathcal{C}\left(\prod_{j=1}^{n} Y_{j1}^{W_{ij}}; \prod_{j=1}^{n} Y_{j2}^{W_{ij}}; \ldots; \prod_{j=1}^{n} Y_{jD}^{W_{ij}}\right)$, for $i = 1, \ldots, n$, where $(\mathbf{W}\triangle\mathbf{Y})_{i.}$ denotes the $i^{th}$ row of $\mathbf{W}\triangle\mathbf{Y}$.

4. Let $\mathbf{Y}_1, \mathbf{Y}_2 \in (\mathbf{S}^D)^n$, then $\mathbf{W}\triangle(\mathbf{Y}_1 \oplus \mathbf{Y}_2) = (\mathbf{W}\triangle\mathbf{Y}_1) \oplus (\mathbf{W}\triangle\mathbf{Y}_2)$.

## 5.3 Multivariate LAG regression model

The principle of compositional regression models is to use a transformation to send the data from the simplex to some coordinate space and to postulate a gaussian regression model in the coordinate space as in Egozcue et al. Egozcue et al. (2012). The model can then be transferred back to the simplex by inverse transformation. In our case, the model in coordinate space must be a multivariate regression model because we have several response variables. For simplicity, we concentrate on the so-called LAG model which includes endogenous lagged variables on the right hand side of the model equations. An extension to a Durbin model would be immediate (LeSage and Pace LeSage and Pace (2009)). Since our model will be postulated in the coordinate space we choose to star all variables and parameters in the coordinate space Subsection 3.1, despite the fact that this section is not specific to compositional data and that we remain in the coordinate space for the whole of Subsection 3.1.

### 5.3.1 Model in coordinate space

We consider a sample of size $n$ and assume that we have $M$ endogenous variables, hence $M$ linear regression equations ($M$ will be $D-1$ in Section 3.2). For a $n \times M$ matrix $\mathbf{A}$, we will use the same notation as in Section 2, $\mathbf{A} = (\mathbf{A}_{.1}, \ldots, \mathbf{A}_{.M}) = (\mathbf{A}_{1.}^T, \ldots, \mathbf{A}_{n.}^T)^T$ with $\mathbf{A}_{.l}$ being the $l^{th}$ column of $\mathbf{A}$ and $\mathbf{A}_{i.}$ being the $i^{th}$ row of $\mathbf{A}$.
Let $\mathbf{Y}^*$ be a $n \times M$ matrix of dependent variables and $\mathbf{X}$ be a $n \times K$ matrix of explanatory variables. We will allow for using a different set of explanatory variables in each equation. For this reason, we denote by $S_l^{\mathbf{Y}^*}$, $S_l^{\mathbf{X}}$, $S_l^{\mathbf{WY}^*}$ the sets of indices of the variables which appear in the $l^{th}$ equation for $\mathbf{Y}^*$, $\mathbf{X}$, $\mathbf{WY}^*$ respectively. Accordingly $\mathbf{Y}^*_{S_l^{\mathbf{Y}^*}}$, $\mathbf{X}_{S_l^{\mathbf{X}}}$, $\mathbf{Y}^*_{S_l^{\mathbf{WY}^*}}$ will denote the columns of $\mathbf{Y}^*$, $\mathbf{X}$, $\mathbf{WY}^*$ which appear in the $l^{th}$ equation. Let $\mathbf{\Gamma}^* = (\Gamma^*_{ml})$ and $\mathbf{R}^* = (R^*_{ml}), (\ m,\ l = 1, \ldots, M)$ be $M \times M$ matrices of parameters. $\mathbf{R}^*$ contains the parameters associated to the lagged endogenous variables on the right hand side of the model equation. As in the simultaneous equations literature in econometrics, each endogenous variable may also appear in each model equation so that $\mathbf{\Gamma}^*$ contains the corresponding parameters. Finally $\boldsymbol{\beta}^*$ is a $K \times M$ matrix of parameters for the explanatory variables. $\boldsymbol{\epsilon}^*$ denotes a $n \times M$ error matrix. As in Kelejian and Prucha Kelejian and Prucha (2004), we consider the following model

$$\mathbf{Y}^*_{.l} = \sum_{m \in S_l^{\mathbf{Y}^*}} \Gamma^*_{ml} \mathbf{Y}^*_{.m} + \mathbf{X}_{S_l^{\mathbf{X}}} \boldsymbol{\beta}^*_{S_l^{\mathbf{X}}} + \sum_{m \in S_l^{\mathbf{WY}^*}} R^*_{ml} \mathbf{WY}^*_{.m} + \boldsymbol{\epsilon}^*_{.l} \tag{5.2}$$

Note that model (5.2) is written for each column of $\mathbf{Y}^*$ i.e. for each component of the composition dependent vector but the $M$ equations are linked by the covariance structure of the errors. Indeed, we assume that the errors are centered $\mathbb{E}(\boldsymbol{\epsilon}^*) = 0$ and that $\mathbb{E}(\boldsymbol{\epsilon}^*_{i.}\boldsymbol{\epsilon}^*_{j.}) = \boldsymbol{\Sigma}^*$ if $i = j$ and 0 if $i \neq j$ (individuals are independent but components of a given individual have a covariance structure). Kelejian and Prucha Kelejian and Prucha (2004) suggest and study the properties of a Spatial Two Stage Least Square (S2SLS) procedure as well as a Spatial Three Stage Least Square (S3SLS) procedure to estimate model (5.2). Following their suggestion, we consider $\mathbf{H}$ a subset of linearly independent columns of the $n \times 3K$ matrix $(\mathbf{X}, \mathbf{WX},\ \mathbf{W^2X})$. Let $\mathbf{P}_H = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ denote the projection matrix onto the space generated by the columns of $\mathbf{H}$. For the $l^{th}$ equation, we group the variables and parameters of the right hand side into a single vector $\mathbf{Z}_{.l}$ of variables and a single vector $\boldsymbol{\delta}^*_{.l}$ of parameters:

$$\mathbf{Z}_{.l} = \left[\mathbf{Y}^*_{S_l^{\mathbf{Y}^*}}; \mathbf{X}_{S_l^{\mathbf{X}}}; \mathbf{WY}^*_{S_l^{\mathbf{WY}^*}}\right]; \boldsymbol{\delta}^*_l = [\boldsymbol{\Gamma}^*_{.l}; \boldsymbol{\beta}^*_l; \mathbf{R}^*_{.l},]$$

The S2SLS estimation method for this model then proceeds as follows for each equation (i.e. each component) separately :

1. Perform a univariate regression of each column of $\mathbf{Z}$ on $\mathbf{H}$ and compute the fitted values $\tilde{\mathbf{Z}}_{.l}$:
$$\tilde{\mathbf{Z}}_{.l} = \mathbf{P}_H\mathbf{Z}_{.l} = \left[\mathbf{P}_H\mathbf{Y}^*_{S_l^{\mathbf{Y}^*}},\ \mathbf{X}_{S_l^{\mathbf{X}}},\ \mathbf{P}_H\mathbf{WY}^*_{S_l^{\mathbf{WY}^*}}\right].$$
2. Perform a univariate regression of $\mathbf{Y}^*_{.l}$ on $\tilde{\mathbf{Z}}_{.l}$:
$$\tilde{\boldsymbol{\delta}}^*_l = (\tilde{\mathbf{Z}}^T_{.l}\tilde{\mathbf{Z}}_{.l})^{-1}\tilde{\mathbf{Z}}^T_{.l}\mathbf{Y}^*_{.l}.$$

At the end of step 2, we can calculate the residuals by

$$\tilde{\boldsymbol{\epsilon}}^*_{.l} = \mathbf{Y}^*_{.l} - \hat{\mathbf{Y}}^*_{.l} = \mathbf{Y}^*_{.l} - \mathbf{Z}_{.l}\tilde{\boldsymbol{\delta}}^*_l,$$

and get an estimate of the covariance matrix $\boldsymbol{\Sigma}^*$ with

$$\tilde{\boldsymbol{\Sigma}}^*_{ml} = \frac{\tilde{\boldsymbol{\epsilon}}^{*T}_{.m}\tilde{\boldsymbol{\epsilon}}^*_{.l}}{n}.$$

Until now, the covariance structure between equations has not been taken into account and the Three Stage Least Square (S3SLS) method is supposed to correct for this. To write the expression of the S3SLS estimator, we need to vectorize $\mathbf{Y}^*$ (by stacking the columns of $\mathbf{Y}$) resulting in $\mathbf{y}^* = \text{vec}(\mathbf{Y}^*)$ and write the explanatory matrices as follows

$$\mathcal{Z} = \begin{bmatrix} \mathbf{Z}_{.1} & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \mathbf{Z}_{.M} \end{bmatrix},\ \tilde{\mathcal{Z}} = \begin{bmatrix} \mathbf{P}_H\mathbf{Z}_{.1} & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \mathbf{P}_H\mathbf{Z}_{.M} \end{bmatrix}.$$

We then get a corrected estimator $\hat{\boldsymbol{\delta}}^*$ of $\boldsymbol{\delta}^*$

$$\hat{\boldsymbol{\delta}}^* = (\tilde{\mathcal{Z}}^T(\tilde{\boldsymbol{\Sigma}}^{*-1} \otimes \mathbf{I}_n)\mathcal{Z})^{-1}\tilde{\mathcal{Z}}^T(\tilde{\boldsymbol{\Sigma}}^{*-1} \otimes \mathbf{I}_n)\mathbf{y}^*. \tag{5.3}$$

It is known that if the matrix $\mathbf{R}^*$ is not diagonal, the S2SLS $\tilde{\boldsymbol{\delta}}^*$ and S3SLS $\hat{\boldsymbol{\delta}}^*$ estimators are identical (see Greene Greene (1997) page 488).

In the application of Section 5, we consider a slightly more general model in which we include compositional variables among the explanatory (see for example Filzmoser et al. Filzmoser et al. (2018)). The additional complexity is the same as for a non-spatial model hence for the sake of simplicity we did not consider this extra layer in this section.

### 5.3.2 Writing the LAG regression model in the simplex space

Starting now with a sample of compositional vectors $\mathbf{Y}$ in $\mathbf{S}^D$, and given an ilr trnasformation, we postulate a model like (5.2) for the ilr transformed coordinates of $\mathbf{Y}$. Applying the ilr inverse transformation to each of the equations of model (5.2) with $M = D - 1$ and using Proposition 1, we easily get that the system of equations (5.2) is equivalent to the system

$$\mathbf{Y}_{i.} = \mathbf{R}^T \boxdot (\mathbf{W} \triangle \mathbf{Y})_{i.} \bigoplus_{k=1}^{K} \mathbf{X}_{ik} \odot \boldsymbol{\beta}_k \oplus \boldsymbol{\Gamma}^T \boxdot \mathbf{Y}_{i.} \oplus \boldsymbol{\epsilon}_{i.} \tag{5.4}$$

where the model is now written at the individual level for all components simultaneously whereas in (5.2) it was at the component level for all individuals simultaneously. One can write relationships between parameters in coordinate space and parameters in the simplex. The classical relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ remains the same (see for example Filzmoser et al. Filzmoser et al. (2018))

$$\boldsymbol{\beta}_k = \text{ilr}^{-1}(\boldsymbol{\beta}_k^*) = \exp(\mathbf{V}_D\boldsymbol{\beta}_k^*).$$

Considering $\mathbf{R}^T \boxdot (\mathbf{W} \triangle \mathbf{Y})_{i.}$ and $\boldsymbol{\Gamma}^T \boxdot \mathbf{Y}_{i.}$ as compositional explanatory, the relationships between $\mathbf{R}$ and $\mathbf{R}^*$ and between $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}^*$ are as follows (see Chen et al. Chen et al. (2017))

$$\mathbf{R} = \mathbf{V}_D\mathbf{R}^*\mathbf{V}_D^T \qquad \text{and} \qquad \boldsymbol{\Gamma} = \mathbf{V}_D\boldsymbol{\Gamma}^*\mathbf{V}_D^T.$$

## 5.4 Simulation

A simulation study of the performance of the S2SLS and S3SLS methods can be found in Das et al. Das et al. (2003) but it is restricted to the case of a single dependent variable. For this reason, we now investigate by simulation the properties of the estimators $\boldsymbol{\beta}^*$, $\mathbf{R}^*$ and $\boldsymbol{\Sigma}^*$ of the S2SLS and S3SLS methods in the multivariate spatial autoregressive model. We consider the $n = 283$ cantons of the Occitanie region in France with a neighborhood structure based on 10 nearest neighbors.

For a number of replications $N = 1000$, we simulate three explanatory variables $\mathbf{X_1}, \mathbf{X_2}$ and $\mathbf{X_3}$ following the Gaussian distributions $\mathcal{N}(0, 9)$, $\mathcal{N}(0, 6)$ and $\mathcal{N}(0, 9)$ respectively. When simulating the two dependent variables, we include all explanatory variables $\mathbf{X_1}, \mathbf{X_2}$ and $\mathbf{X_3}$ in each of the two equations.

The parameter $\boldsymbol{\beta}^*$, the covariance matrix $\boldsymbol{\Sigma}^*$ and the matrix $\mathbf{R}^*$ are respectively assigned the following values

$$\boldsymbol{\beta}^* = \begin{bmatrix} \beta_{01}^* & \beta_{02}^* \\ \beta_{11}^* & \beta_{12}^* \\ \beta_{21}^* & \beta_{22}^* \\ \beta_{31}^* & \beta_{32}^* \end{bmatrix} = \begin{bmatrix} +3 & -3 \\ +2 & -3 \\ +1 & -2 \\ -1 & +3 \end{bmatrix} ; \ \boldsymbol{\Sigma}_{\bar{d}}^* = \begin{bmatrix} 0.7 & 0.093 \\ 0.093 & 0.1 \end{bmatrix} ; \ \mathbf{R}_{\bar{d}}^* = \begin{bmatrix} 0.5 & 0.6 \\ 0.4 & 0.3 \end{bmatrix}$$

Alternative diagonal matrices for $\boldsymbol{\Sigma}^*$ and $\mathbf{R}^*$ are also considered

$$\boldsymbol{\Sigma}_d^* = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.1 \end{bmatrix} ; \ \mathbf{R}_d^* = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.3 \end{bmatrix}$$

We consider four data generating processes (DGP) respectively denoted by $\boldsymbol{\Sigma}_{\bar{d}}^* \mathbf{R}_{\bar{d}}^*$, $\boldsymbol{\Sigma}_{\bar{d}}^* \mathbf{R}_d^*$, $\boldsymbol{\Sigma}_d^* \mathbf{R}_{\bar{d}}^*$ and $\boldsymbol{\Sigma}_d^* \mathbf{R}_d^*$ according to the choice of matrices $\boldsymbol{\Sigma}^*$ and $\mathbf{R}^*$. For each DGP, we calculate a Monte Carlo performance measure of the estimators proposed in Section 5.3. The performance is measured by the relative root mean squared error (RRMSE), which is defined for an estimator $\hat{\theta}$ of a parameter $\theta$ by:

$$\text{RRMSE}(\hat{\theta}) = \frac{\text{RMSE}(\hat{\theta})}{|\theta|} \quad \text{with} \quad \text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \hat{\theta}^{(i)} - \theta \right)^2}.$$

Table 5.1 presents the relative root mean square error for the S2SLS method and S3SLS methods for all DGPs. We did not report the relative bias in Table 5.1 because its value is quite similar to the relative root mean square error showing that the bias dominates the error. The percentage of error is generally very small with a maximum of 10% for the variance parameters and values less than 1% for the other parameters. For all DGPs, the largest differences between the estimates occur for the estimation of the intercepts and for the estimation of the $\mathbf{R}^*$ matrix and these differences are small in all cases. Concerning the $\boldsymbol{\beta}$ and $\mathbf{R}$ estimates, only the S2SLS results are reported since S3SLS yields exactly the same results for the first two DGPs as proved by the theory. Concerning the variance estimates, note that there is no difference in its computation for the S2SLS and S3SLS. Finally, if we compare the three estimates in the case of DGP $\boldsymbol{\Sigma}_d^* \mathbf{R}_d^*$, we can see that the results are quite close showing that S2SLS is a practical alternative to maximum likelihood in the framework of this model.

Table 5.1: The RRMSE (in %) for all DGPs and parameters

| **Parameters** | **RRMSE(%)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\boldsymbol{\Sigma}^*_{\bar{d}}\mathbf{R}^*_{\bar{d}}$ | $\boldsymbol{\Sigma}^*_{d}\mathbf{R}^*_{\bar{d}}$ | $\boldsymbol{\Sigma}^*_{\bar{d}}\mathbf{R}^*_{d}$ | | $\boldsymbol{\Sigma}^*_{d}\mathbf{R}^*_{d}$ | | |
| | S2SLS | S2SLS | S2SLS | S3SLS | S2SLS | S3SLS | MLE |
| $\beta^*_{01}$ | 0.16 | 0.11 | 0.50 | 0.33 | 0.47 | 0.46 | 0.41 |
| $\beta^*_{11}$ | 0.18 | 0.15 | 0.21 | 0.21 | 0.18 | 0.18 | 0.18 |
| $\beta^*_{21}$ | 0.03 | 0.03 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 |
| $\beta^*_{31}$ | 0.88 | 0.87 | 0.86 | 0.86 | 0.86 | 0.86 | 0.86 |
| $\beta^*_{02}$ | 0.29 | 0.32 | 0.05 | 0.08 | 0.08 | 0.08 | 0.08 |
| $\beta^*_{12}$ | 0.12 | 0.14 | 0.11 | 0.11 | 0.14 | 0.14 | 0.14 |
| $\beta^*_{22}$ | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| $\beta^*_{32}$ | 0.02 | 0.03 | 0.04 | 0.04 | 0.00 | 0.00 | 0.00 |
| $\mathrm{R}^*_{11}$ | 0.17 | 0.21 | 0.76 | 0.64 | 0.68 | 0.67 | 0.64 |
| $\mathrm{R}^*_{12}$ | 0.40 | 0.39 | - | - | - | - | - |
| $\mathrm{R}^*_{21}$ | 0.41 | 0.37 | - | - | - | - | - |
| $\mathrm{R}^*_{22}$ | 0.42 | 0.58 | 0.63 | 0.67 | 0.83 | 0.83 | 0.82 |
| $\sigma^{2*}_{11}$ | 9.46 | 9.12 | 9.43 | 9.43 | 9.09 | 9.09 | 10.37 |
| $\sigma^{2*}_{12}$ | 2.27 | - | 1.67 | 1.67 | - | - | - |
| $\sigma^{2*}_{21}$ | 2.27 | - | 1.67 | 1.67 | - | - | - |
| $\sigma^{2*}_{22}$ | 9.11 | 9.38 | 9.18 | 9.18 | 9.16 | 9.16 | 7.62 |

## 5.5 Application to political economics

Vote share data of the 2015 French departmental election of the Occitanie region in France are collected from the Cartelec website[3]. Corresponding socio-economic data (for 2014) are downloaded from the INSEE website[4]. The number of political parties presenting candidates at that election is higher than 15. However for simplicity reasons, we have aggregated them into three main components: Left, Right and Extreme-Right[5]. The dependent variable is thus a compositional variable which contains the vote shares of Left, Right and Extreme Right party. Cantons with at least one missing value on one of the components of the dependent vector have been eliminated resulting in $n = 207$ cantons in the final dataset. We use the following contrast matrix for $D = 3$, built using balances (see Pawlowsky-Glahn et al. Pawlowsky-Glahn et al. (2015) page 40)

$$\mathbf{V}_3 = \begin{bmatrix} 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}.$$

[3]https://www.data.gouv.fr/fr/datasets/elections-departementales-2015-resultats-par-bureaux-de-vote/

[4]https://www.insee.fr/fr/statistiques

[5]for more details, see https://fr.wikipedia.org/wiki/Elections_départementales_francaises_de_2015

Table 5.2: Data description.

| Variable name | Description |
|---|---|
| Vote share | Left(L), Right(R), Extreme Right(XR) |
| Diploma | SUP, <BAC, BAC |
| Employment | AZ, BE, FZ, GU, OQ |
| Age | Age_1840, Age_4064, Age_65 |
| unemp | Unemployment rate |
| nbvoter | Number of voters |
| income | Proportion of people who pay income tax |

This particular matrix defines the following ilr coordinates

$$\text{ilr}_1(\mathbf{x}) = \frac{1}{\sqrt{6}}(2\log x_1 - \ln x_2 - \log x_3) = \frac{2}{\sqrt{6}}\log\frac{x_1}{\sqrt{x_2 x_3}}$$
$$\text{ilr}_2(\mathbf{x}) = \frac{1}{\sqrt{2}}(\log x_2 - \log x_3) = \frac{1}{\sqrt{2}}\log\frac{x_2}{x_3}.$$

With this choice, the first ilr coordinate opposes the Left wing to the geometric mean of the Right wing and the Extreme Right party and the second opposes the Right wing to the Extreme Right party.
Our explanatory variables, presented in Table 5.2, include both compositional and classical variables. For the three compositional variables, Diploma, Employment and Age, the contrast matrices have been built using balances as for the dependent variable.
The categories of these variables are as follows

- Employment has five levels: AZ (agriculture, fisheries), BE (manufacturing industry, mining industry and others), FZ (construction), GU (business, transport and services) and OQ (public administration, teaching, human health),
- Diploma has three levels: <BAC for people with at most some secondary education, BAC for people with at least some secondary education and at most a high school diploma, and SUP for people with a university diploma,
- Age has three levels: Age_1840 for people from 18 to 40 years old, Age_4064 for people from 40 to 64 years old, and Age_65 for elderly.

An additional variable measuring the number of voters in each canton has been included to take into account a potential size effect.

This data set has been analyzed in Nguyen and Laurent Nguyen and Laurent (2019) without taking into account the spatial structure. We use model (5.2) in the coordinate space with $\mathbf{\Gamma}^* = 0$. Indeed, the reason for including spatially lagged dependent variable in the equations is for taking into account the spatial dependence and this justifies terms

Table 5.3: Multivariate independent and spatial regression models with compositional and classical explanatory variables

| | *Independence model* | | *Spatial dependence model* | |
|---|---|---|---|---|
| | y_ilr[, 1] | y_ilr[, 2] | y_ilr[, 1] | y_ilr[, 2] |
| Constant | $-3.69(0.89)^{***}$ | $-2.66(0.45)^{***}$ | $-2.06(0.89)^{**}$ | $-1.34(0.14)^{***}$ |
| diplome_ilr1 | $-1.27(0.50)^{**}$ | $-0.29(0.25)$ | $-0.56(0.46)$ | $-0.63(0.17)^{***}$ |
| diplome_ilr2 | $-0.03(0.61)$ | $-0.90(0.30)$ | $-0.02(0.53)$ | $-0.54(0.45)$ |
| employ_ilr1 | $-0.18(0.14)$ | $-0.13(0.07)^{*}$ | $-0.13(0.12)$ | $-0.11(0.24)$ |
| employ_ilr2 | $+0.49(0.16)^{***}$ | $-0.03(0.08)$ | $+0.38(0.14)^{***}$ | $+0.02(0.27)$ |
| employ_ilr3 | $-0.21(0.11)^{*}$ | $+0.01(0.06)$ | $-0.24(0.10)^{**}$ | $-0.08(0.06)$ |
| employ_ilr4 | $+0.21(0.06)^{***}$ | $+0.01(0.03)$ | $+0.09(0.05)^{*}$ | $-0.04(0.07)$ |
| age_ilr1 | $-1.14(0.37)^{***}$ | $+1.00(0.18)^{***}$ | $-0.57(0.39)$ | $+0.33(0.05)^{***}$ |
| age_ilr2 | $+0.48(0.30)$ | $-1.33(1.15)^{***}$ | $+0.51(0.31)$ | $-0.67(0.03)^{***}$ |
| unemp | $-0.05(2.27)$ | $+9.70(1.14)^{***}$ | $-0.34(2.88)$ | $+1.43(0.20)^{***}$ |
| income | $+4.30(0.89)^{***}$ | $+1.14(0.06)^{***}$ | $+2.69(0.84)^{***}$ | $+0.70(0.16)^{***}$ |
| nbvoter | +2e-06(5e-06) | $+1e-05(2e-05)^{***}$ | −6e-06(4e-06) | +1e-05(1.47) |
| $R_1$ | - | - | $+0.99(0.43)^{**}$ | $+0.11(0.07)$ |
| $R_2$ | - | - | $-0.09(0.00)^{***}$ | $+0.70(0.09)^{***}$ |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

such as $\sum_{m \in S_l^{\mathbf{WY}^*}} R^*_{ml} \mathbf{WY}^*_{.m}$ in model (5.2). But in our case, there is no economic reason for introducing terms such as $\sum_{m \in S_l^{\mathbf{Y}^*}} \Gamma^*_{ml} \mathbf{Y}^*_{.m}$ on the right hand side of the equations in the coordinate space. In order to estimate the parameters in model (5.2), we carry out the S2SLS and the S3SLS methods from Section 5.3.

First of all, looking at the size effect, the number of voters is not significant in the spatial model whereas it was in the non-spatial one (indeed there is some heterogeneity in the distribution of the number of voters at the canton level). The spatial dependence parameters (elements of the matrix $\mathbf{R}^*$) are significant showing that a spatial dependence phenomenon is present in this data. The sign and significance of most $\boldsymbol{\beta}$ parameters are very comparable, except in two cases (the diploma and age variables) for which the significance changes from one ilr of $\mathbf{Y}$ to the other. Finally, we can say that this spatial LAG model would be necessary if we were to evaluate spillover effects across cantons (see LeSage and Pace LeSage and Pace (2009)).

## 5.6 Conclusion

Motivated by an example in political economics, we develop a simultaneous spatial autoregressive model for compositional data combining the simultaneous systems of spatially interrelated cross sectional equations of Kelejian and Prucha Kelejian and Prucha (2004) and the compositional regression models (see Filzmoser et al. Filzmoser et al. (2018)). We propose an implementation using spatial two-stage and three-stage methods

which are easy to implement.

There are several directions we could consider to go further in this framework. We could first of all consider alternative estimation methods. For example partial least squares procedures for the Spatial LAG model have been proposed in Wang et al. Wang et al. (2019) but for a single dependent variable. Similarly and with the same restriction, Spatial regression trees are developed for the LAG model in Wagner and Zeileis Wagner and Zeileis (2019). In a different direction, the aggregation of political parties in three blocks could be reconsidered. On the one hand, this aggregation avoids the zero problem due to the absence of some parties in some cantons but on the other hand it results in an information loss: imputation methods could be used to solve this as in Palarea and Martin-Fernandez Palarea-Albaladejo and Martín-Fernández (2015).

## 5.7 Acknowledgements

We acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010

# Conclusion (English version)

The objective of this thesis is to investigate the outcome of an election and the impacts of the socio-economics factors on the vote shares in the multiparty system from a mathematical point of view.

The vote shares of the departmental election in France in 2015 form a vector called composition. Thus, the classical regression model cannot be used directly to model these vote shares because of contraints of compositional data. In Chapter 2, we present a regression model in which the dependent variable is a compositional variable and the set of explanatory variables contains both classical variables and compositional variables. We analyze the impacts of socio-economic factors on the outcome of the election through predicting the vote shares according to either a classical explanatory variable or a compositional explanatory variable. Some graphical techniques are also presented. However, it would be more appreciated to interpret the coefficients of regression model on the simplex.

Furthermore, some authors show that electoral data often exhibit heavy tail behavior. Thus, we propose to replace the Normal distribution by the Student distribution. However, there are two versions of the Student distribution: the uncorrelated Student (UT) distribution and the independent Student (IT) distribution. In Chapter 3, we present a complete summary for the Student distributions which includes the univariate and multivariate Student, the IT and the UT distribution with fixed degrees of freedom. We prove that the maximum likelihood estimator of the covariance matrix in the UT model is asymptotically biased. We also provide an iterative reweighted algorithm to compute the maximum likelihood estimator of parameter of the IT model. A simulation is provided and some Kolmogorov–Smirnov tests based on the Mahalanobis distance are carried out to select the right model. However, this does not work for the UT model because of a single realization of $n$ observation of the multivariate distribution.

In Chapter 4, we apply the multivariate Student (IT) regression model to our political economy data. We then compare this model to the multivariate Normal regression model. We also apply the Kolmogorov–Smirnov tests based on the Mahalanobis distance which is proposed in chapter 3 to select a better model.

Finally, we investigate the assumption of statistical independence across territorial units which may be questionable due to potential spatial autocorrelation for compositional data. We develop a simultaneous spatial autoregressive model for compositional data which allows for both spatial correlation and correlations across equations by using

two-stage and three-stage least squares methods. We present a simulation study to illustrate these methods. An application to a data set from the 2015 French departmental election are also showed.

There is still work to continue in the direction of overcoming the problem of zeros in vote shares. This problem is already present for the departmental French elections at the canton level when aggregating the electoral parties in three categories. It would have been even more serious when considering the original political parties with no aggregation. Besides, another direction consists in considering the multivariate Student distribution for a spatial model.

# Conclusion (version française)

L'objectif de cette thèse est d'étudier le résultat d'élections et l'impact des facteurs socio-économiques sur les parts de vote dans le système multipartite d'un point de vue mathématique.

Les votes de l'élection départementale en France en 2015 forment un vecteur appelé composition. Ainsi, le modèle de régression classique ne peut pas être utilisé directement pour modéliser ces parts de vote en raison de contraintes de données de composition. Au chapitre 2, nous présentons un modèle de régression dans lequel la variable dépendante est une variable de composition et les variables explicatives contiennent à la fois des variables classiques et des variables de composition. Nous analysons les impacts des facteurs socio-économiques sur l'issue de l'élection en prédisant les parts de vote en fonction d'une variable explicative classique ou d'une variable explicative de composition. Quelques techniques graphiques sont également présentées. Néanmoins, il serait plus judicieux d'interpréter les coefficients du modèle de régression sur le simplexe.

Par ailleurs, certains auteurs montrent que les données électorales présentent souvent un comportement extrême. Nous proposons donc de remplacer la distribution Normale par la distribution de Student. Cependant, il existe deux versions de la distribution Student: la distribution Student non corrélée (UT) et la distribution independetn Student (IT). Dans la troisième partie, nous présentons un résumé complet de la distribution Student, comprenant les distributions Student univariée et multivariée, IT et UT à degrés de liberté fixes. Nous prouvons que l'estimateur de maximum de vraisemblance de la matrice de covariance dans le modèle UT est asymptotiquement biaisé. Nous fournissons également un algorithme itératif repondéré pour calculer l'estimateur du maximum de vraisemblance du paramètre du modèle IT. Une simulation est fournie et certains tests de Kolmogorov – Smirnov basés sur la distance de Mahalanobis sont effectués pour sélectionner le bon modèle. Cependant, cela ne fonctionne pas pour le modèle UT en raison d'une seule réalisation de $n$ observations de la distribution multivariée.

Enfin, nous étudions l'hypothèse d'indépendance statistique entre unités territoriales, qui peut être mise en doute du fait de l'autocorrélation spatiale potentielle des données de composition. Nous développons un modèle autorégressif spatial simultané pour les données de composition qui permet à la fois la corrélation spatiale et les corrélations entre équations en utilisant des méthodes de moindres carrés à deux étages et à trois étages. Nous présentons une étude de simulation pour illustrer ces méthodes. Une application à un ensemble de données de l'élection départementale française de 2015 est également

présentée.

Il reste encore du travail à faire pour surmonter le problème des zéros dans les parts de vote. Ce problème est déjà présent pour les élections départementales françaises au niveau cantonal lorsque l'on regroupe les partis électoraux en trois catégories. Cela aurait été encore plus grave si l'on considérait les partis politiques d'origine sans agrégation. En outre, une autre direction consiste à examiner la distribution multivariée de Student pour le modèle spatial.

# Appendix A

# Appendix for Chapter 3

*Proof of Proposition 1.* Using Expression (3.1), the joint density function of $\hat{\boldsymbol{\epsilon}}_{UT}$ is:

$$\begin{aligned} p(\boldsymbol{\epsilon}_{UT}|\mathbf{0},\boldsymbol{\Omega}_{UT},\nu) &= \frac{f(\nu)}{\det(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_{UT})^{1/2}} \left[1 + \frac{1}{\nu-2}\boldsymbol{\epsilon}_{UT}^T(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_{UT})^{-1}\boldsymbol{\epsilon}_{UT}\right]^{-\frac{\nu+nL}{2}} \\ &= \frac{f(\nu)}{\det(\boldsymbol{\Sigma}_{UT})^{n/2}} \left[1 + \frac{1}{\nu-2}\boldsymbol{\epsilon}_{UT}^T(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_{UT})^{-1}\boldsymbol{\epsilon}_{UT}\right]^{-\frac{\nu+nL}{2}} \\ &= \frac{f(\nu)}{\det(\boldsymbol{\Sigma}_{UT})^{n/2}} \left[1 + \frac{1}{\nu-2}\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^T\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}\right]^{-\frac{\nu+nL}{2}} \end{aligned}$$

Therefore, the logarithm of $p(\boldsymbol{\epsilon}_{UT}|\mathbf{0},\boldsymbol{\Omega}_{UT},\nu)$ is:

$$\log p(\boldsymbol{\epsilon}_{UT}|\mathbf{0},\boldsymbol{\Omega}_{UT},\nu) = \log f(\nu) - \frac{n}{2}\log\boldsymbol{\Sigma}_{UT} - \frac{\nu+nL}{2}\log\left[1 + \frac{1}{\nu-2}\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^T\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}\right]. \tag{A.1}$$

In order to maximize $\log p(p(\boldsymbol{\epsilon}_{UT}|\mathbf{0},\boldsymbol{\Omega}_{UT},\nu))$ as a function of $\boldsymbol{\beta}_{UT}$, we follow the same argument as in Theorem 8.4 from Seber (2009) for the Gaussian case and obtain that the minimum of $\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^T\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}$ is obtained for:

$$\hat{\boldsymbol{\beta}}_{UT} = \left(\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}}\right)^{-1}\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{Y}}.$$

Besides, taking the partial derivative of (A.1) as a function of $\boldsymbol{\Sigma}_{UT}$, we obtain:

$$\begin{aligned} \frac{\partial\log(p(\boldsymbol{\epsilon}_{UT}|\mathbf{0},\boldsymbol{\Omega}_{UT},\nu))}{\partial\boldsymbol{\Sigma}_{UT}} &= -\frac{n\boldsymbol{\Sigma}_{UT}^{-1}}{2} - \frac{(\nu+nL)}{2}\frac{\partial\log(\nu-2+\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^T\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi})}{\partial\boldsymbol{\Sigma}_{UT}} \\ &= -\frac{n\boldsymbol{\Sigma}_{UT}^{-1}}{2} - \frac{(\nu+nL)}{2}\frac{\partial(\nu-2+\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^T\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi})/\partial\boldsymbol{\Sigma}_{UT}}{\nu-2+\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^T\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}}. \end{aligned}$$

Let:

$$w_{UT} = \frac{1}{\nu-2+\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^T\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}}. \tag{A.2}$$

We have:

$$\begin{aligned}\frac{\partial \log(p(\boldsymbol{\epsilon}_{UT}|\mathbf{0},\boldsymbol{\Omega}_{UT},\nu))}{\partial \boldsymbol{\Sigma}_{UT}} &= -\frac{n\boldsymbol{\Sigma}_{UT}^{-1}}{2} - \frac{(\nu+nL)\,w_{UT}}{2}\partial(\nu-2+\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^{T}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi})/\partial\boldsymbol{\Sigma}_{UT} \\ &= -\frac{n\boldsymbol{\Sigma}_{UT}^{-1}}{2} + \frac{(\nu+nL)\,w_{UT}}{2}\sum_{i=1}^{n}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}\boldsymbol{\epsilon}_{UTi}^{T}\boldsymbol{\Sigma}_{UT}^{-1}\end{aligned}$$

Solving $\frac{\partial \log(p(\boldsymbol{\epsilon}_{UT}|\mathbf{0},\boldsymbol{\Omega}_{UT},\nu))}{\partial \boldsymbol{\Sigma}_{UT}} = 0$ and letting $\mathbf{E} = \sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}\boldsymbol{\epsilon}_{UTi}^{T}$, we have:

$$\begin{aligned}\boldsymbol{\Sigma}_{UT}^{-1} &= \frac{\nu+nL}{n}\,w_{UT}\sum_{i=1}^{n}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}\boldsymbol{\epsilon}_{UTi}^{T}\boldsymbol{\Sigma}_{UT}^{-1} \\ \boldsymbol{\Sigma}_{UT}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\Sigma}_{UT} &= \frac{\nu+nL}{n}\,w_{UT}\sum_{i=1}^{n}\boldsymbol{\Sigma}_{UT}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}\boldsymbol{\epsilon}_{UTi}^{T}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\Sigma}_{UT} \\ \boldsymbol{\Sigma}_{UT} &= (\nu+nL)\,w_{UT}\,\frac{\mathbf{E}}{n} \qquad \text{(A.3)}\end{aligned}$$

The expression of $w_{UT}$ in (A.3) can be simplified by noting that:

$$\begin{aligned}\boldsymbol{\Sigma}_{UT}^{-1} &= n((\nu+nL)w_{UT})^{-1}\mathbf{E}^{-1} \\ \sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^{T}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi} &= n((\nu+nL)w_{UT})^{-1}\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^{T}\mathbf{E}^{-1}\boldsymbol{\epsilon}_{UTi} \\ &= \frac{n}{(\nu+nL)w_{UT}}\mathrm{tr}\left(\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}\boldsymbol{\epsilon}_{UTi}^{T}\mathbf{E}^{-1}\right) \\ \sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^{T}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi} &= \frac{nL}{(\nu+nL)w_{UT}}. \qquad \text{(A.4)}\end{aligned}$$

Replacing the expression of $\sum_{i=1}^{n}\boldsymbol{\epsilon}_{UTi}^{T}\boldsymbol{\Sigma}_{UT}^{-1}\boldsymbol{\epsilon}_{UTi}$ from (A.4) into $w_{UT}$, we get:

$$w_{UT} = \frac{\nu}{(\nu-2)(\nu+nL)}.$$

Finally,

$$\hat{\boldsymbol{\Sigma}}_{UT} = \frac{\nu}{\nu-2}\frac{\sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_{UTi}\hat{\boldsymbol{\epsilon}}_{UTi}^{T}}{n}.$$

□

*Proof of Proposition 2.* The property $\mathbb{E}(\hat{\boldsymbol{\beta}}_{UT}) = \boldsymbol{\beta}_{UT}$ is immediate. In order to facilitate the derivation of the proof for $\hat{\boldsymbol{\Sigma}}_{UT}$, we write Model (3.4) as:

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon} \qquad \text{(A.5)}$$

where:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1L} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nL} \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nK} \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} \beta_{01} & \beta_{0L} \\ \beta_{11} & \beta_{1L} \\ \vdots & \vdots \\ \beta_{K1} & \beta_{KL} \end{bmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1L} \\ \vdots & \vdots & \vdots & \vdots \\ \epsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nL} \end{bmatrix}, \hat{\mathbf{B}}_{UT} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad \text{and} \quad \hat{\boldsymbol{\varepsilon}}_{UT} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{UT}.$$

Let $\mathbf{E} = \hat{\boldsymbol{\varepsilon}}_{UT}^T\hat{\boldsymbol{\varepsilon}}_{UT}$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. We have $\mathbf{MXB} = 0$, and following Seber (2009), Theorem 8.2,

$$\begin{aligned} \mathbf{E} &= (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{UT})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{UT}) = (\mathbf{MY})^T\mathbf{MY} = \mathbf{Y}^T\mathbf{MY} \\ &= (\mathbf{Y} - \mathbf{XB})^T\mathbf{M}(\mathbf{Y} - \mathbf{XB}) = \boldsymbol{\varepsilon}^T\mathbf{M}\boldsymbol{\varepsilon} = \sum_h \sum_i M_{hi}\boldsymbol{\varepsilon}_h\boldsymbol{\varepsilon}_i^T. \end{aligned}$$

Since $\mathbb{E}(\boldsymbol{\varepsilon}_h\boldsymbol{\varepsilon}_i^T) = \begin{cases} \boldsymbol{\Sigma} & \text{if } h = i \\ 0 & \text{otherwise} \end{cases}$, for $h, i = 1, \ldots, n$, $\mathbb{E}(\mathbf{E}) = \sum_h M_{hh}\boldsymbol{\Sigma} = tr(\mathbf{M})\boldsymbol{\Sigma} = (n-K)\boldsymbol{\Sigma}$ and:

$$\mathbb{E}(\hat{\boldsymbol{\Sigma}}_{UT}) = \mathbb{E}\left(\frac{\nu}{\nu-2}\frac{\mathbf{E}}{n}\right) = \frac{\nu}{\nu-2}\frac{\mathbb{E}(\mathbf{E})}{n} = \frac{\nu}{\nu-2}\frac{n-K}{n}\boldsymbol{\Sigma}_{UT}.$$

□

# Appendix B

# Appendix for Chapter 5

***Proof of Proposition 4.*** Let $\mathbf{Y} \in \mathbf{S}^D$, $\mathbf{W}$ be a $n \times n$ matrix, $\alpha$ be a scalar, and let $(\mathbf{W}\triangle\mathbf{Y})_{i.}$ denotes the $i^{th}$ row of $\mathbf{W}\triangle\mathbf{Y}$, $i, j = 1, \ldots, n,\ l, m = 1, \ldots, D$.
1. $\text{ilr}(\mathbf{W}\triangle(\alpha \odot \mathbf{Y})) = \text{ilr}(\mathbf{W}\triangle\mathbf{Y}^{\alpha}) = \alpha\mathbf{W}\ln^T(\mathbf{Y})\mathbf{V}_D = \alpha\mathbf{W}\text{ilr}(\mathbf{Y}) = \alpha\ \text{ilr}(\mathbf{W}\triangle\mathbf{Y})$.
2. We have

$$\text{ilr}(\mathbf{W}\triangle(\alpha \odot \mathbf{Y})) = \text{ilr}(\mathbf{W}\triangle\mathbf{Y}^{\alpha}) = \alpha\text{ilr}(\mathbf{W}\triangle\mathbf{Y})$$

then

$$\text{ilr}^{-1}(\text{ilr}(\mathbf{W}\triangle(\alpha \odot \mathbf{Y}))) = \text{ilr}^{-1}(\alpha\text{ilr}(\mathbf{W}\triangle\mathbf{Y})) = \alpha \odot (\mathbf{W}\triangle\mathbf{Y})$$

Thus,

$$\mathbf{W}\triangle(\alpha \odot \mathbf{Y}) = \alpha \odot (\mathbf{W}\triangle\mathbf{Y}).$$

3. We have

$$\begin{aligned}(\mathbf{W}\triangle\mathbf{Y})_{i.} &= \text{ilr}^{-1}(\text{ilr}((\mathbf{W}\triangle\mathbf{Y})_{i.})) \\ &= \text{ilr}^{-1}(\mathbf{W}\text{ilr}(\mathbf{Y}))_{i.} \\ &= \mathcal{C}(\exp(\mathbf{W}\text{ilr}(\mathbf{Y})\mathbf{V}_{\mathbf{D}}^{\mathbf{T}})_{i.})\end{aligned}$$

where

$$(\mathbf{W}\text{ilr}(\mathbf{Y})\mathbf{V}_{\mathbf{D}}^{\mathbf{T}})_{i.} = \left(\ln\prod_{j=1}^{n}\left(\frac{Y_{j1}}{\prod_{l=1}^{D}Y_{jl}^{1/D}}\right)^{W_{ij}}; \ln\prod_{j=1}^{n}\left(\frac{Y_{j2}}{\prod_{l=1}^{D}Y_{jl}^{1/D}}\right)^{W_{ij}}; \ldots; \ln\prod_{j=1}^{n}\left(\frac{Y_{jD}}{\prod_{l=1}^{D}Y_{jl}^{1/D}}\right)^{W_{ij}}\right).$$

Thus,

$$\begin{aligned}(\mathbf{W}\triangle\mathbf{Y})_{i.} &= \mathcal{C}\left(\prod_{j=1}^{n}\left(\frac{Y_{j1}}{\prod_{l=1}^{L}Y_{jl}^{1/D}}\right)^{W_{ij}}; \prod_{j=1}^{n}\left(\frac{Y_{j2}}{\prod_{l=1}^{L}Y_{jl}^{1/D}}\right)^{W_{ij}}; \ldots; \prod_{j=1}^{n}\left(\frac{Y_{jD}}{\prod_{l=1}^{L}Y_{jl}^{1/D}}\right)^{W_{ij}}\right) \\ &= \mathcal{C}\left(\prod_{j=1}^{n}Y_{j1}^{W_{ij}}; \prod_{j=1}^{n}Y_{j2}^{W_{ij}}; \ldots; \prod_{j=1}^{n}Y_{jD}^{W_{ij}}\right)\end{aligned}$$

4. Let $\mathbf{Y}_1, \mathbf{Y}_2 \in (\mathbf{S}^D)^n$, and let $\mathbf{Y}_1^* = \text{ilr}(\mathbf{Y}_1)$, $\mathbf{Y}_2^* = \text{ilr}(\mathbf{Y}_2)$, then

$$\begin{aligned}
\text{ilr}^{-1}(\mathbf{W} \triangle (\mathbf{Y}_1 \oplus \mathbf{Y}_2)^*) &= \text{ilr}^{-1}(\mathbf{W}(\mathbf{Y}_1^* + \mathbf{Y}_2^*)) \\
&= \text{ilr}^{-1}(\mathbf{W}\mathbf{Y}_1^* + \mathbf{W}\mathbf{Y}_2^*) \\
&= \text{ilr}^{-1}(\mathbf{W}\text{ilr}(\mathbf{Y}_1) + \mathbf{W}\text{ilr}(\mathbf{Y}_2)) \\
&= \mathbf{W}(\text{ilr}^{-1}(\text{ilr}(\mathbf{Y}_1) + \text{ilr}(\mathbf{Y}_2))) \\
&= \mathbf{W}(\text{ilr}^{-1}(\text{ilr}(\mathbf{Y}_1))) + \mathbf{W}(\text{ilr}^{-1}(\text{ilr}(\mathbf{Y}_2))) \\
&= (\mathbf{W} \triangle \mathbf{Y}_1) \oplus (\mathbf{W} \triangle \mathbf{Y}_2)
\end{aligned}$$

Thus,

$$\mathbf{W} \triangle (\mathbf{Y}_1 \oplus \mathbf{Y}_2) = (\mathbf{W} \triangle \mathbf{Y}_1) \oplus (\mathbf{W} \triangle \mathbf{Y}_2)$$

□

# Bibliography


J. Aitchison. A general class of distributions on the simplex. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 136–146, 1985.

J. Aitchison. *The Statistical Analysis of Compositional Data (Monographs on Statistics and Applied Probability)*. Chapman and Hall, 2011.

J. Aitchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

S. Ansolabehere and W. Leblanc. A spatial model of the relationship between seats and votes. *Mathematical and Computer Modelling*, 48(9-10):1409–1420, 2008.

L. Beauguitte and C. Colange. Analyser les comportements électoraux à l'échelle du bureau de vote. *Sciences de l'Homme et de la Société*, 2013.

D. Billheimer, T. Cardoso, E. Freeman, P. Guttorp, H.-W. Ko, and M. Silkey. Natural variability of benthic species composition in the delaware bay. *Environmental and Ecological Statistics*, 4(2):95–115, 1997.

M. Bilodeau and D. Brenner. *Theory of multivariate statistics*. Springer Science & Business Media, 2008.

R. S. Bivand, E. J. Pebesma, V. Gomez-Rubio, and E. J. Pebesma. *Applied spatial data analysis with R*, volume 747248717. Springer, 2008.

C. Borghesi and J.-P. Bouchaud. Spatial correlations in vote statistics: a diffusive field model for decision-making. *The European Physical Journal B-Condensed Matter and Complex Systems*, 75(3):395–404, 2010.

J. Chen, X. Zhang, and S. Li. Multiple linear regression with compositional response and covariates. *Journal of Applied Statistics*, 44(12):2270–2285, 2017.

C. Croux, M. Fekri, and A. Ruiz-Gazen. Fast and robust estimation of the multivariate errors in variables model. *Test*, 19(2):286–303, 2010.

D. Das, H. H. Kelejian, and I. R. Prucha. Finite sample properties of estimators of spatial autoregressive models with autoregressive disturbances. *Papers in Regional Science*, 82(1):1–26, 2003.

A. P. Dempster. Iterative reweighted least squares for linear regression when errors are normal/inde-pendent distributed. *Multivariate analysis*, pages 35–57, 1980.

F. Z. Dogru, Y. M. Bulut, and O. Arslan. Doubly reweighted estimators for the parameters of the multivariate t-distribution. *Communications in Statistics-Theory and Methods*, pages 1–21, 2018.

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

J. J. Egozcue, J. Daunis-I-Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser. *Simplicial regression. The normal model.* na, 2012.

C. Fernandez and M. F. Steel. Multivariate student-t regression models: Pitfalls and inference. *Biometrika*, 86(1):153–167, 1999.

P. Filzmoser, K. Hron, and C. Reimann. The bivariate statistical analysis of environmental (compositional) data. *The Science of the total environment*, 408 19:4230–8, 2010.

P. Filzmoser, K. Hron, and M. Templ. Applied compositional data analysis. *With Worked*, 2018.

D. A. S. Fraser. *Inference and linear models*. McGraw-Hill Companies, 1979.

D. A. S. Fraser and K. W. Ng. Multivariate regression analysis with spherical error. *Multivariate analysis*, 5:369–386, 1980.

T. Fung and E. Seneta. Modelling and estimation for bivariate financial returns. *International statistical review*, 78(1):117–133, 2010.

R. Gnanadesikan and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, pages 81–124, 1972.

W. Greene. *Econometric Analysis*. Prentice-Hall international editions. Prentice Hall, 1997. ISBN 9780023466021. URL `https://books.google.fr/books?id=y\_crAQAAIAAJ`.

M. Hofert. On sampling from the multivariate t distribution. *The R Journal*, 5(2): 129–136, 2013.

J. Honaker, J. N. Katz, and G. King. A fast, easy, and efficient estimator for multiparty electoral data. *Political Analysis*, 10(1):84–100, 2002.

P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley, 2009.

N. L. Johnson and S. Kotz. Distributions in statistics: Continuous multivariate distributions, john wiley & sons. *Inc. New York*, 1972.

R. Kan and G. Zhou. Modeling non-normality using multivariate t: implications for asset pricing. *China Finance Review International*, 7(1):2–32, 2017.

J. N. Katz and G. King. A statistical model for multiparty electoral data. *American Political Science Review*, 93(1):15–32, 1999.

A. Kavanagh, S. Fotheringham, and M. Charlton. A geographically weighted regression analysis of the election specific turnout behaviour in the republic of ireland. In *Elections, Public Opinion and Parties Conference, Nottingham 8th to 10th September*, 2006.

H. H. Kelejian and I. R. Prucha. Independent or uncorrelated disturbances in linear regression: An illustration of the difference. *Economics Letters*, 19(1):35–38, 1985.

H. H. Kelejian and I. R. Prucha. Estimation of simultaneous systems of spatially interrelated cross sectional equations. *Journal of econometrics*, 118(1-2):27–50, 2004.

H. H. Kelejian, I. R. Prucha, and Y. Yuzefovich. Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: Large and small sample results. In *Spatial and spatiotemporal econometrics*, pages 163–198. Emerald Group Publishing Limited, 2004.

J. T. Kent, D. E. Tyler, and Y. Vard. A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics-Simulation and Computation*, 23(2):441–453, 1994.

S. Kotz and S. Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

P. Kynčlová, P. Filzmoser, and K. Hron. Modeling compositional time series with vector autoregressive models. *Journal of Forecasting*, 34(4):303–314, 2015.

K. Lange and J. S. Sinsheimer. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2(2):175–198, 1993.

K. L. Lange, R. J. Little, and J. M. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

T. J. Leininger, A. E. Gelfand, J. M. Allen, and J. A. Silander. Spatial regression modeling for compositional data with many zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):314–334, 2013.

J. LeSage and R. K. Pace. *Introduction to spatial econometrics*. Chapman and Hall/CRC, 2009.

J. B. Lewis and D. A. Linzer. Estimating regression models in which the dependent variable is based on estimates. *Political analysis*, 13(4):345–364, 2005.

C. Liu. Ml estimation of the multivariatetdistribution and the em algorithm. *Journal of Multivariate Analysis*, 63(2):296–312, 1997.

C. Liu and D. B. Rubin. Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*, pages 19–39, 1995.

E. Mansley and U. Demšar. Space matters: Geographic variability of electoral turnout determinants in the 2012 london mayoral election. *Electoral Studies*, 40:322–334, 2015.

R. A. Maronna. Robust m-estimators of multivariate location and scatter. *The annals of statistics*, pages 51–67, 1976.

A. B. T. Martins, W. H. Bonat, and P. J. Ribeiro. Likelihood analysis for a class of spatial geostatistical compositional models. *Spatial Statistics*, 17:121–130, 2016.

A. J. McNeil, R. Frey, P. Embrechts, et al. *Quantitative risk management: Concepts, techniques and tools*, volume 3. Princeton university press Princeton, 2005.

M. C. Mert, P. Filzmoser, G. Endel, and I. Wilbacher. Compositional data analysis in epidemiology. *Statistical methods in medical research*, 27(6):1878–1891, 2018.

J. Morais. *Impact of media investments on brands' market shares: a compositional data analysis approach.* PhD thesis, Toulouse School of Economics (TSE), 2017.

J. Morais, C. Thomas-Agnan, and M. Simionic. Interpreting the impact of explanatory variables in compositional models. *TSE Working Paper*, 2017.

T. Nguyen and T. Laurent. Coda methods and the multivariate student distribution with an application to political economy. 2019.

T. H. A. Nguyen, T. Laurent, C. Thomas-Agnan, and A. Ruiz-Gazen. Analyzing the impacts of socio-economic factors on french departmental elections with coda methods. *TSE Working paper*, 2018.

T. H. A. Nguyen, A. Ruiz-Gazen, C. Thomas-Agnan, and T. Laurent. Multivariate student versus multivariate gaussian regression models with application to finance. *Journal of Risk and Financial Management*, 12(1):28, 2019.

K. d. Overmars, G. De Koning, and A. Veldkamp. Spatial autocorrelation in multi-scale land use models. *Ecological modelling*, 164(2-3):257–270, 2003.

J. Palarea-Albaladejo and J. A. Martín-Fernández. zcompositions - r package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, 2015.

V. Pawlowsky and H. Burger. Spatial structure analysis of regionalized compositions. *Mathematical Geology*, 24(6):675–691, 1992.

V. Pawlowsky-Glahn and J. J. Egozcue. Spatial analysis of compositional data: a historical review. *Journal of Geochemical Exploration*, 164:28–32, 2016.

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

B. Pirzamanbein, J. Lindström, A. Poska, and M.-J. Gaillard. Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *arXiv preprint arXiv:1511.06417*, 2015.

E. Platen and R. Rendek. Empirical evidence on student-t log-returns of diversified world stock indices. *Journal of statistical theory and practice*, 2(2):233–251, 2008.

I. R. Prucha and H. H. Kelejian. The structure of simultaneous equation estimators: A generalization towards nonnormal disturbances. *Econometrica: Journal of the Econometric Society*, pages 721–736, 1984.

M. Roth. *On the multivariate t distribution*. Linköping University Electronic Press, 2012.

R. H. Rubio, J. F. C. L. Costa, and M. A. A. Bassani. A geostatistical framework for estimating compositional data avoiding bias in back-transformation. *Rem: Revista Escola de Minas*, 69(2):219–226, 2016.

E. Salazar, R. Giraldo, and E. Porcu. Spatial prediction for infinite-dimensional compositional data. *Stochastic environmental research and risk assessment*, 29(7):1737–1749, 2015.

G. A. Seber. *Multivariate observations*, volume 252. John Wiley & Sons, 2009.

R. S. Singh. Estimation of error variance in linear regression models with errors having multivariate student-t distribution with unknown degrees of freedom. *Economics Letters*, 27(1):47–53, 1988.

N. Small. Plotting squared radii. *Biometrika*, 65(3):657–658, 1978.

B. C. Sutradhar and M. M. Ali. Estimation of the parameters of a regression model with a multivariate t error variable. *Communications in Statistics-Theory and Methods*, 15 (2):429–450, 1986.

R. C. Sutter. Spatial econometric modeling of presidential voting outcomes. *Theses and Dissertations, The University of Toledo*, 2005.

H. T. Trinh and J. Morais. Impact of socioeconomic factors on nutritional diet in vietnam from 2004 to 2014: new insights using compositional data analysis. Technical report, TSE Working Papers, 2017.

K. G. Van den Boogaart and R. Tolosana-Delgado. *Analyzing compositional data with R*, volume 122. Springer, 2013.

M. Wagner and A. Zeileis. Heterogeneity and spatial dependence of regional growth in the eu: A recursive partitioning approach. *German Economic Review*, 20(1):67–82, 2019.

H. Wang, J. Gu, S. Wang, and G. Saporta. Spatial partial least squares autoregression: Algorithm and applications. *Chemometrics and Intelligent Laboratory Systems*, 184: 123–131, 2019.

T. Yoshida and M. Tsutsumi. On the effects of spatial relationships in spatial compositional multivariate models. *Letters in Spatial and Resource Sciences*, 11(1):57–70, 2018.

A. Zellner. Bayesian and non-bayesian analysis of the regression model with multivariate student-t error terms. *Journal of the American Statistical Association*, 71(354):400–405, 1976.