Statistical inference for extreme risk measures : Implication for the insurance of natural disasters

Meryem Bousebata

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L' UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 25 mai 2016

Présentée par

Meryem BOUSEBATA

Thèse dirigée par **Stéphane GIRARD**, Directeur de recherche, Inria et co-dirigée par **Geoffroy ENJOLRAS**, Professeur des universités, Université Grenoble Alpes

préparée au sein du **Laboratoire Jean Kuntzmann**
dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Statistical inference for extreme risk measures: Implication for the insurance of natural disasters.

Inférence statistique pour les mesures de risques extrêmes : Implication pour l'assurance des catastrophes naturelles.

Thèse soutenue publiquement le **30 Mars 2022**,
devant le jury composé de :

Mme Veronique MAUME-DESCHAMPS
Professeur des universités, Université Lyon 1, Rapportrice, Présidente
Mr Alexandre GOHIN
Directeur de recherche, INRAE Rennes, Rapporteur
Mr Gilles STUPFLER
Maître de conférences, ENSAI Rennes, Examinateur
Mr Philippe MADIÈS
Professeur des universités, Université Grenoble Alpes, Examinateur
Mr Julyan ARBEL
Chargé de recherche, Inria Grenoble Rhône-Alpes, Invité
Mr Stéphane GIRARD
Directeur de recherche, Inria Grenoble Rhône-Alpes, Directeur de thèse
Mr Geoffroy ENJOLRAS
Professeur des universités, Université Grenoble Alpes, Co-Directeur de thèse

Acknowledgements

Cette thèse n'aurait pas été la même sans le soutien de nombreuses personnes que je tiens à remercier ici.

Tout d'abord, je tiens à exprimer ma profonde gratitude à mes directeurs de thèse Stéphane Girard et Geoffroy Enjolras pour la qualité de l'encadrement et la disponibilité tout au long de cette thèse. Merci Stéphane de m'avoir tant appris, j'ai beaucoup apprécié votre rigueur scientifique et votre clairvoyance, ainsi que vos qualités humaines, en particulier votre bienveillance et votre sens de l'humour. Merci Geoffroy pour vos conseils, votre expertise et vos connaissances en économie agricole ont été un appui précieux pour la réalisation de ce travail. Je vous suis profondément reconnaissante, à tous les deux, pour les encouragements, le soutien continu et la patience dont vous avez fait preuve pendant ces années de collaboration.

Je tiens à remercier mes rapporteurs, Véronique Maume-Deschamps et Alexandre Gohin, ainsi que les autres membres de mon jury, Gilles Stupfler, Philippe Madiès et Julyan Arbel. Je suis honorée de votre participation et de l'intérêt que vous avez porté à mon travail. Je remercie particulièrement Gilles et Alexandre pour les suggestions et remarques détaillées sur mon manuscrit. Je remercie également Véronique pour les conseils et le temps consacré au suivi de mon sujet de thèse. Je suis sincèrement reconnaissante à tous les membres du jury de s'être rendus disponibles pour la soutenance de cette thèse, pour toutes les critiques constructives et pour avoir suscité des discussions fructueuses.

Ma gratitude va au Cross Disciplinary Project Risk (CDP Risk) de l'Université de Grenoble Alpes qui a contribué au financement de cette thèse. Je suis ravie d'avoir eu l'occasion de participer à tant de discussions stimulantes et de séminaires qui ont enrichi mes connaissances et mon expérience. Merci à tous les membres du projet et en particulier aux riskuers avec qui j'ai partagé de très bons moments.

Je tiens à remercier tous les membres de l'équipe Mistis, rebaptisée par la suite Statify. Ce

fut un réel plaisir de partager l'environnement de travail avec les chercheurs permanents, Florence, Julyan, Jean Baptiste, Sophie, la famille Mistis avec qui j'ai eu la chance de commencer cette thèse, je pense à Brice, Karina, Thibaud, Fei, Marta, Clément, Alexis, Antoine, Steeven, Fatima, Michal, Virgilio, Veronica..., et les Statifiers, Dasha, Minti, Pascal, Lucrezia, Theo, Hana, Jacopo, Louise, Yuchen, Pierre... Une mention spéciale au groupe des fantastiques, Benoit, Masha, Alexandre avec qui j'ai partagé cette aventure de thèse du début à la fin. Merci pour les moments conviviaux, pour l'ambiance et l'atmosphère de travail très agréable grâce à la bonne humeur et la simplicité de chacun. Faire partie de cette équipe a été une expérience scientifiquement stimulante et humainement enrichissante.

Par ailleurs, je suis également reconnaissante pour les amitiés qui se sont nouées à l'Inria et à l'UGA, mais aussi pour toutes les belles personnes que j'ai rencontrées depuis mon arrivée à Grenoble. Merci Nour pour ta gentillesse et ta sérénité, Djalila pour m'avoir chaleureusement accueilli à Grenoble, Benoît pour nos discussions mathématiques et philosophiques très stimulantes, Sami le "Sniper" pour toutes les parties où tu nous as mis KO, Masha pour notre magnifique colocation et ton énergie nucléaire, Minti pour les aventures en montagne, les fêtes et les tapages nocturnes, Dasha pour ta douceur et ta volonté sans faille, Jonathan pour tes mots d'encouragement et ta présence rassurante, Théo pour tes magnifiques pizzas. Je pense aussi à Karina, Poornima, Alexandre, Qi, Charles K, Rouba, Amela, Hassan, Anouar, Tim, Michaël, Grégoire pour tous les bons moments passés en leur compagnie. Enfin Brice, merci pour avoir été là pour moi en permanence et pour ton empathie plus que remarquable. Je garde un souvenir ému de toutes nos soirées bouffe, jeux, discussions passionnantes sur tout et rien, et surtout de nos belles escapades en montagne, en commençant par la bastille et la cascade de l'Oursière... en arrivant au lac de Crozet.

Un énorme "Big Up" à tous mes amis vivant au loin avec qui j'ai passé des moments irremplaçable de gaieté et d'insouciance, je pense à Emna, Karima, Abdoulahy, Gloria, Joya, Sami, Zineb, Elena, Marie-lucie, Daphné, Lamia... merci pour votre gentillesse, votre bonne humeur et votre compréhension malgré nos séjours très courts ou peu fréquents ensemble. Un merci particulier à Emna, Karima et Abdoulahy qui, en plus sont venus en pleine semaine pour assister à ma soutenance, m'ont apporté un soutien si précieux durant les derniers mois de ma thèse. Enfin, j'adresse un grand merci à Nazim de m'avoir accompagné pendant la rédaction de mon manuscrit. Merci pour ta patience et ton soutien tant pour ce projet que pour le reste.

Finalement, je voudrais remercier infiniment ma famille; mes parents, mes sœurs et frères, mes beaux-frères et belles-sœurs, mes nièces et neveux, pour leur amour inconditionnel et leur soutien constant tout au long de ma thèse. Je les remercie du fond du cœur de m'avoir

toujours poussé à donner le meilleur de moi-même, à viser haut et à rester intègre. Je remercie particulièrement ma mère pour son soutien affectueux, Nadia et Hicham pour leur tendresse et leur accompagnement à tous les niveaux, Lyes pour sa gentillesse et ses conseils pertinents sur mon parcours universitaire, Nahid pour son aide précieuse, sa clairvoyance et son érudition qui forcent mon admiration. Votre confiance et votre soutien m'ont été indispensables pour arriver là où je suis aujourd'hui.

Enfin, je dédie cette thèse à mon défunt frère Anouar qui nous a quitté trop tôt et qui aurait tant aimé être présent le jour de ma soutenance. Merci pour tout, ta simplicité, ta gentillesse, ta joie de vivre contagieuse... Je suis tellement fière de la personne que tu as été et j'espère te rendre fier.

"Great things are done by a series of small things brought together"

Vincent Van Gogh

Abstract

This thesis takes place in extreme value statistics and agricultural insurance frameworks. For the first line of research, the extreme quantile of a response variable $Y \in \mathbb{R}$ can often be linked to a vector of covariates $X \in \mathbb{R}^p$. When p is large compared to the sample size n, the conditional distribution of Y given X becomes difficult to estimate, especially when dealing with extreme values. The first contribution of this thesis is to propose a new approach, called Extreme-PLS, for dimension reduction in conditional extreme values settings. This approach consists in reducing the dimension of X by maximizing the covariance between a linear combination of coordinates X and Y given large values of Y. We establish the asymptotic normality of the Extreme-PLS estimator under a single-index model. The second contribution provides a Bayesian extension to the Extreme-PLS method to address data scarcity problems in distribution tails. This approach allows to identify the direction of dimension reduction by introducing a prior information on it. It provides a Bayesian framework for computing the posterior distribution of the direction, where the likelihood function is obtained from a von Mises-Fisher distribution adapted to hyperballs. Three prior distributions are considered: conjugate, hierarchical and sparse priors. Finally, the performance of both approaches is evaluated on simulated data, and an application on French farm income data is provided as an illustration.

Regarding the second line of research, climate disruption and market deregulation have increased and impacted agricultural production. Farmers' incomes are faced with two main types of risk related to price and yield volatility. Protection against these risks fall within a good risk management and thus farmers' insurance coverage. The third contribution of this thesis concerns the study and modelling of the dependence structure between crop yield and price risks using copulas. We also use conditional copulas to take into account the effect of other covariates such as crop insurance purchase, claims and weather factors. The last contribution focuses on considering the natural hedge mechanism, i.e. the negative dependence between yields and prices, in a revenue insurance scheme. We analyse its effect on the value of the actuarially fair premium on an example of revenue insurance contract pricing. The results show that a natural hedge is likely to reduce insurance premiums in France. All studies focus on French farm income data in the cereal (maize and wheat) and wine sectors.

Résumé

Cette thèse s'inscrit dans le cadre de la statistique des valeurs extrêmes et de l'assurance agricole. Pour le premier axe de recherche, le quantile extrême d'une variable réponse $Y \in \mathbb{R}$ peut souvent être lié à un vecteur de covariables $X \in \mathbb{R}^p$. Lorsque p est grand comparé à la taille de l'échantillon n, la distribution conditionnelle de Y étant donné X devient difficile à estimer, surtout lorsqu'on a affaire à des valeurs extrêmes. La première contribution de cette thèse est de proposer une nouvelle approche, appelée Extreme-PLS, pour la réduction de la dimension dans le cadre des valeurs extrêmes conditionnelles. Cette approche consiste à réduire la dimension de X en maximisant la covariance entre une combinaison linéaire des composants de X et de Y étant donné de grandes valeurs de Y. Nous établissons la normalité asymptotique de l'estimateur Extreme-PLS sous un modèle à indice unique. La deuxième contribution est une extension bayésienne de la méthode Extreme-PLS pour traiter les problèmes de rareté des données dans les queues de distribution. Cette approche permet d'identifier la direction de la réduction de la dimension en introduisant des informations a priori sur celle-ci. Elle fournit un cadre bayésien pour calculer la distribution postérieure de la direction, où la fonction de vraisemblance est obtenue à partir d'une distribution de von Mises-Fisher adaptée aux hyper boules. Trois distributions a priori sont considérées : loi conjuguée, hiérarchique et sparse. Enfin, la performance des deux approches est évaluée sur des données simulées, et une application sur des données de revenus agricoles françaises est fournie à titre d'illustration.

En ce qui concerne le deuxième axe de recherche, le dérèglement climatique et la dérégulation des marchés ont augmenté et impacté la production agricole. Les revenus des agriculteurs sont confrontés à deux principaux types de risques liés à la volatilité des prix et des rendements. La protection contre ces risques relève d'une bonne gestion des risques et donc de la couverture d'assurance des agriculteurs. La troisième contribution de cette thèse concerne l'étude et la modélisation de la structure de dépendance entre les risques de rendement et de prix par les copules. Nous utilisons également les copules conditionnelles pour prendre en compte l'effet d'autres covariables telles que l'achat d'assurance récolte, les sinistres et les facteurs météorologiques. La dernière contribution porte sur la prise en compte du mécanisme de couverture naturelle, c-à-d la dépendance négative entre les rendements et les prix, dans un système d'assurance des revenus. Nous analysons son effet sur la valeur de la prime actuariellement juste sur un exemple de tarification de contrat d'assurance revenu. Les résultats montrent qu'une couverture naturelle est susceptible de réduire les primes d'assurance en France. L'ensemble des études se concentrent sur les données du revenu agricole français dans les secteurs céréaliers (maïs et blé) et viticoles.

Contents

Introduction

PhD Context

This PhD is taking place within the cross-disciplinary project CDP Risk@UGA framework. This project is in line with the Sendai framework for disaster risk reduction 2015-2030, which encourages states to prevent better and anticipate disaster risks. The PhD was carried out in the Statify team, a joint team of Inria Grenoble Rhône-Alpes and LJK (Laboratoire Jean Kuntzmann), in collaboration with CERAG (Center for applied management studies and research) at Grenoble Alpes University (UGA).
This thesis was supported by the Agence Nationale de la Recherche (ANR) in the framework of the Programme d'Investissements d'Avenir (ANR-15-IDEX-02). It was also supported by the Chair Stress Test, Risk Management and Financial Steering, led by the French Ecole Polytechnique and its Foundation, and sponsored by BNP Paribas.

Farm income risks

In recent years, climate and financial market risks have increased considerably. The agricultural sector is particularly concerned by this evolution since farmers' income is exposed to the two main risks of **yields** and **prices**. The risk of poor yields is mainly due to **natural perils** such as drought, hail, pests, diseases and agricultural techniques used by farmers (Moschini and Hennessy, 2001; Ullah et al., 2016). In contrast, price risk is linked to the inherent **volatility of financial markets** (Ortmann et al., 1992; Ullah et al., 2016).

The global food price crisis of 2007-2008 led to an unexpected rise and volatility in food prices. One of the main causes of this crisis was the production risk linked to severe droughts. The latter led to a spike in prices, which was exacerbated by some governments imposing export restrictions (Headey, 2011). International prices for maize and wheat roughly doubled (see Figure 1). Thus, farmers were faced with production risks due to drought and price

spikes over a short period. On the other hand, the impacts of climate change on agriculture are constantly increasing. In 2016, cereal crops in France greatly suffered due to adverse weather and severe drought. These unfavourable weather conditions led to significant yield declines. In particular, wheat production suffered the most extreme yield loss in over half a century (Ben-Ari et al., 2018). In 2021, cereal prices sharply increased (see Figure 1) despite the good volumes of cereals harvested in France. This is due to climatic disruption on the American continent (drought in Canada), strong international demand and limited harvests by the main exporters (Russia, United States and Canada).

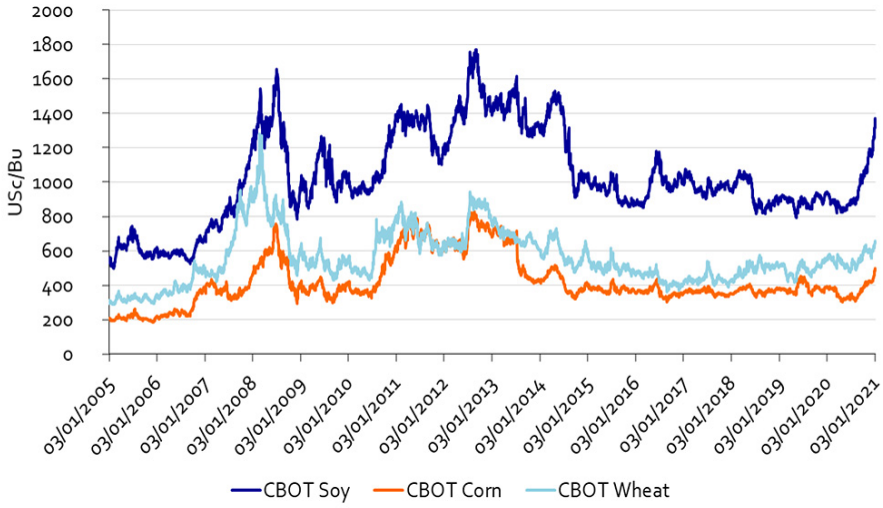


Figure 1 Prices of Chicago Board of Trade (CBOT) Corn, Soy and Wheat since 2005. Source: Bloomberg, Rabobank 2021.

Insurance is one of the main risk management tools to cope with these risks. The various agricultural insurance schemes include yield, price and revenue (Kang, 2007). Yield insurance covers yield losses of a given crop due to any natural risk (Skees et al., 1997; Kang, 2007), while price insurance protects against financial market price risks (e.g. livestock price insurance in the United States) (Kang, 2007). **Revenue insurance** provides a joint price and yield coverage that guarantees farmers a minimum level of income (Hennessy et al., 1997). The latter takes into account the potential dependence between prices and yields, particularly the "natural hedge" effect provided by the inverse relationship. This negative dependence is extremely important as it moderates revenue variability and influences the demand for risk management instruments (Finger, 2012). It is also of practical importance in designing insurance contracts (Finger, 2012; Ramsey et al., 2019). So far, revenue insurance is only available to farmers in countries such as the United States and Canada (Diaz-Caneja et al., 2008). In contrast, no such products are available for European farmers (Meuwissen et al., 2003). Insurance schemes in the European Union have focused mainly on yield coverage. However, insuring yield risk individually offers lower coverage than pooling both risks, including price,

into a single insurance policy. As for France, it has not yet promoted revenue insurance despite being the largest agricultural producer in the European Union. Indeed, the current reform of climate risk management tools in agriculture (bill no. 4758, early January 2022) concerns the adoption of a universal compensation scheme for climate risks, but still does not consider the revenue insurance scheme, which covers market risks as well. In addition, existing insurance policies, such as multi-peril crop insurance, do not provide sufficient coverage despite the important subsidies (Lidsky et al., 2017). Therefore, it seems relevant to consider a revenue insurance program to protect farmers against extreme weather conditions and financial markets volatility.

Dependence modelling in the design of revenue insurance

One of the main difficulties in implementing revenue insurance is modelling the dependence structure between risks. Indeed, premium rates for revenue insurance must be calculated by considering the joint distribution of yields and prices and the natural hedge they imply. This dependence is often measured using the Pearson correlation coefficient (Embrechts et al., 2002; Coble et al., 2000). It is widely known that this correlation is only appropriate for measuring linear and monotonic relationships. Other non-parametric measures of association such as Kendall and Spearman rank correlations are alternatives to overcome these limitations (Ramsey et al., 2019). However, these correlations are also limited as they characterise dependence over the entire risk distribution. In contrast, many risk management questions focus on the behaviour of the distribution tails. In addition, the design of the revenue insurance contract requires the calculation of the premium rates and the probability of loss. The latter depends on the joint probability distribution of yields and prices. Therefore, the joint distribution and associated dependence must be calculated using flexible models that can describe properties such as heavy-tails, extreme co-movement and tail dependence.

Recent research has proposed to use copula functions as a flexible tool to study the dependence between yields and prices. The copulas, introduced by Sklar (1959), combine the marginal distributions of the variables to form the joint (or multivariate) distribution with the associated dependence structure. This is important as the range of parametric families of multivariate distributions available in the statistical literature is not rich enough. Copulas have widespread applications in risk management, insurance and financial economics. However, their use has been a recent contribution to the agricultural economics literature (Goodwin and Hungerford, 2014; Bozic et al., 2014; Fousekis and Grigoriadis, 2017). Only a few works on its application have been carried out in the context of the revenue insurance design (Zhu et al., 2008; Ahmed and Serra, 2015; Duarte and Ozaki, 2019). Thus, it would be interesting

to evaluate how such a programme could be implemented in France: first by using copulas to **model the dependence** between risks and then by calculating an **actuarially fair insurance premium**.

Conditional extreme values in high dimensions

Extreme value theory is a branch of statistics used to model rare or extreme events with a very low probability of occurrence. It provides methods to quantify these events and their consequences in a statistical way. Unlike classical statistics, which focus on the average behaviour of distributions (law of large numbers, central limit theorem, etc.), it is concerned with the behaviour of the distribution tails. Extreme value theory has been widely used in hydrology (Anderson and Meerschaert, 1998; El Methni et al., 2012), reliability (Ditlevsen, 1994), environmental sciences, such as meteorology (Coles et al., 2003; Gardes and Girard, 2010) and climatology (Katz, 1999). It also plays an active role in insurance and finance (Embrechts et al., 2013). Applications also exist in the field of agricultural science (Chuangchid et al., 2013; van Oordt et al., 2021; Morgan et al., 2012; Mitchell et al., 2020). These applications usually require the estimation of extreme quantiles, which lies in the tails of distributions.

The extreme quantile of a variable of interest $Y \in \mathbb{R}$ is often linked to a vector of covariates $X \in \mathbb{R}^p$. The goal is being to describe how extreme values of Y may depend on X. In agricultural risk management, one motivating example is to model the lowest crop yields depending on a wide range of factors, including agricultural inputs and financial and meteorological variables. Thus, the estimation of extreme conditional quantiles and conditional tail index is an important issue in such applications. The conditional tail index drives the heaviness of the conditional distribution tail of Y. The existing literature on their estimation can be divided into three categories: parametric (Smith, 1989; Davison and Smith, 1990), semi-parametric (Hall and Tajvidi, 2000; Ahmad et al., 2019; Davison and Ramesh, 2000) and fully non-parametric (Gardes and Girard, 2008b; Daouia et al., 2011; Goegebeur et al., 2014; Daouia et al., 2013).

In practice, with the increasing volume of stored data, we often face the problem of high dimensional covariates, i.e. the dimension p of X is very large. In this situation, estimating the conditional distribution of Y given X becomes difficult if the sample size is small compared to p. This is referred to as the **curse of dimensionality**. This phenomenon leads to an exploding variance of the estimators, thus impeding the inference. Therefore it becomes necessary to combine **dimension reduction** methods with **extreme value theory**. In the literature, only a few recent works have been dedicated to the combination of these two lines

of work (Gardes, 2018; Xu et al., 2020; Drees and Sabourin, 2021; Aghbalou et al., 2021). However, these methods require some assumptions, such as conditional independence and linear conditional expectation. The latter linearity condition is satisfied as soon as X is elliptically distributed. However, in practice, the vector X is not necessarily expected to follow an elliptical distribution, especially in a high dimensional setting.

High dimensionality raises important problems in the analysis of extreme values. On the one hand, extreme conditional quantiles and classical estimators become inefficient. On the other hand, the quality of the estimate is further degraded in extreme value analysis, as the number of observations in the distribution tails is low. Indeed, the scarcity of extreme events restricts available data. Thus, the introduction of **Bayesian inference** is of great interest when dealing with problems with a **small amount of data**. Meaningful prior information provided by experts could improve the quality of inference (Coles and Powell, 1996). Up to our knowledge, there is no existing work that adopts the Bayesian approach to dimension reduction in the regression context and conditional extremes.

French farm income database

All the analyses conducted in this thesis use a survey of French farmers belonging to the Farm Accountancy Data Network (FADN). The data are accounted for each year from a representative sample of farms with commercial size. This large database provides useful information such as the balance sheet, income statement, farm expenses, chemical inputs, characteristics of the farm operator, and the farm structure. We combine this database with climate and weather information from *Météo France* weather stations, matched at the regional level.

Outline of the thesis and contributions

This thesis is structured around five main chapters.

- **Chapter 1** presents the state of the art. It recalls some basic concepts and theoretical results that are useful for the rest of this thesis. First, it provides an overview of agricultural risks types, management tools, and modelling techniques. Secondly, it introduces the copula model, a necessary tool in the context of farm risks dependence modelling. Some classical dimension reduction methods to overcome the curse of dimensionality are discussed. Then, we present an overview of the Bayesian inference. The foundations of these two approaches will be useful to develop new extreme analysis methods. Finally, some results of extreme value theory are recalled: asymptotic behaviour of extreme

values, domains of attraction and quantiles estimation. We also review the literature on extreme conditional quantiles estimation, as well as on dimension reduction and Bayesian inference in a conditional extreme setting.

In the context of extreme value statistics and agricultural insurance outlined above, the thesis addresses the issues raised by providing four contributions presented in the following chapters:

- In **Chapter 2**, we develop a new model, called Extreme-PLS, for dimension reduction in regression and adapted to distribution tails. This approach combines the Partial Least Squares dimension reduction method and the extreme value analysis. It is developed in the context of a single-index non linear inverse regression model and heavy-tailed distributions. More precisely, this approach aims to estimate the dimension reduction direction by maximising the covariance between a linear combination of covariates X and Y given Y exceeds a high threshold y. The considered model requires neither a linear conditional expectation nor an assumption of conditional independence. Then, we establish the asymptotic normality of our estimator. An iterative procedure to adapt the approach to the multiple-index situation is also given. We evaluate the performance of the Extreme-PLS on simulated data, and we show that it performs better than the proposed estimator of Xu et al. (2020) in some situations. A statistical analysis of French farm income data is also provided to analyse the smallest cereal yields considering other factors (pesticides, fertilisers, farm expenses, weather, etc).

- In **Chapter 3**, we propose a Bayesian formulation of the previous Extreme-PLS model to identify the direction of the dimension reduction and to introduce prior information on it. The application of Bayesian inference in this model allows overcoming data scarcity problems. The proposed approach provides a Bayesian framework for calculating the posterior distribution of the direction, where the likelihood function is obtained from a von Mises-Fisher distribution adapted to hyperballs. Some criteria for choosing a prior distribution are discussed, such as incorporating the sparsity of the directions in a Bayesian lasso prior. We illustrate the proposed method performance with simulations and show that it is particularly efficient for small amounts of data. Finally, the model is applied to analyse the smallest cereal yields on the same dataset as in the previous chapter.

- In **Chapter 4**, we apply the statistical tool of copulas to model the joint distribution of yields and prices and the associated dependence structure. We focus on cereal and wine production in France, a country that has not yet implemented a revenue insurance program. Various parametric copula models are investigated to model yields and

prices risks. Goodness-of-fit tests are also performed to select the most suitable copula. We show that the dependence is relatively high and can be described by the Frank copula. We also model the dependence structure given other factors such as insurance and meteorological variables using conditional copulas. We find that extreme weather conditions strongly affect French cereal and wine income. The analysis shows that existing insurance contracts do not cover wheat and maize crops sufficiently as their prices follow world market trends. Finally, these results highlight some implications for the development of revenue insurance contracts to better hedge cereal farmers.

- Finally, in **Chapter 5** we measure the so-called natural hedge, a negative correlation between prices and yields, in the wheat, maize and wine-growing sectors using copulas. This mechanism is of great practical importance for the design of revenue insurance, as its costs and efficiency are strongly related to the degree of the natural hedge. Then, we quantify the impact of this correlation on the variability of farmers revenue and insurance premiums. Finally, we analyse its effect on the value of the actuarially fair premium, on an example of revenue insurance contract pricing. Results show that natural hedge is likely to reduce insurance premiums in France, particularly for wheat and wine production. This has direct implications for the design and pricing of revenue insurance contracts.

We finish with a conclusion and some perspectives on our research work.

Chapter 1

State of the art

Contents

Abstract

This chapter introduces the basic concepts and theoretical foundations for this thesis. First, we present in Section 1.1 the main types of risks in agriculture, the management tools for their coverage and the modelling of the underlying risks of farm income. A presentation of the French farm income dataset is also given. Section 1.2 provides an introduction to copula models and briefly discusses their estimation, goodness-of-fit tests and conditional copulas. Classical dimension reduction methods are presented in Section 1.3. Then, we provide an overview of Bayesian modelling and some inferential techniques in Section 1.4. The last Section 1.5 constitutes an introduction to the theory of extreme values in the univariate case. This part first deals with the asymptotic behaviour of extreme values of a sample. Then it proposes a characterisation of the distributions associated with the different domains of attraction. It also gives some results on the estimation of extreme quantiles. Finally, it provides the theory of univariate extreme values in the presence of a covariate, in particular the estimation of conditional extreme quantiles, and presents the notion of dimension reduction and Bayesian inference (seen in Sections 1.3 and 1.4) in an extreme setting.

Resumé

Ce chapitre introduit les concepts de base et les fondements théoriques de cette thèse. Tout d'abord, nous présentons dans la Partie 1.1 les principaux types de risques en agriculture, les outils de gestion permettant de les couvrir et la modélisation des risques sous-jacents du revenu agricole. Une présentation de la base de données sur le revenu agricole français est également fournie.La Partie 1.2 présente une introduction aux modèles de copules et aborde brièvement leur estimation, les tests de qualité d'ajustement et les copules conditionnelles. Les méthodes classiques de réduction de dimension sont présentées à la Partie 1.3. Nous fournissons ensuite un aperçu de la modélisation bayésienne et de certaines techniques inférentielles dans la Partie 1.4. La dernière Partie 1.5 constitue une introduction à la théorie des valeurs extrêmes dans le cas univarié. Cette partie traite d'abord le comportement asymptotique des valeurs extrêmes d'un échantillon. Elle propose ensuite une caractérisation des distributions associées aux différents domaines d'attraction. Elle donne également quelques résultats sur l'estimation des quantiles extrêmes. Enfin, elle fournit la théorie des valeurs extrêmes univariées en présence d'une covariable, en particulier l'estimation des quantiles extrêmes conditionnels, et présente les notions de réduction de dimension et d'inférence bayésienne (vues en Parties 1.3 et 1.4) dans un cadre extrême.

1.1 Agricultural income risks and risk management

Agriculture is a sector where farmers are constantly faced with multiple and increasing risks. These risks are increased due to a range of factors, including the globalisation of commodity trade, social and economic change, uncontrollable natural events and climate change (Duong et al., 2019). The latter involves adverse outcomes, including yields and incomes variability (Wing et al., 2021), and can also involve catastrophic outcomes at scales beyond the individual farmer, such as financial bankruptcy, food insecurity and human health problems. Therefore, farmers must simultaneously face and manage a variety of risks that can have cumulative effects (Komarek et al., 2020; Ullah et al., 2016). One example is the global food crisis of 2007–2008 caused by a surge in international cereal prices. This crisis was the result of a complex interaction of several factors, including excessive speculation on agricultural commodity futures markets and crop failures due to drought (Headey, 2011). All these factors have negatively impacted farmers who had to cope with production risk, markets risk and institutional risk due to unexpected changes in government policy. Hence, risks outcome can trigger cascading effects through another set of risks. Thus, it is essential to understand these agricultural risks and the options available to mitigate their impacts.
To this end, we start by giving the main types of risks in the agricultural sector in Paragraph 1.1.1. In Paragraph 1.1.2, we present some of the risk management tools available to farmers to cope with risks. We focus on the revenue insurance management instrument that will be presented in Paragraph 1.1.3. Finally, the modelling of risks is discussed in Paragraph 1.1.4.

1.1.1 Main types of agricultural risks

Agricultural risks can be classified into two main categories, according to (Hardaker et al., 2015), namely business and financial risks as summarised in Table 1.1. Business risk is generally defined as the uncertainty inherent in the business, regardless of how it is financed (Eidman, 1990). It includes production, market, institutional and human risks. First, the production risk arises from the unpredictable nature of weather, climate change, pests, diseases, technology, machinery efficiency, quality of inputs, fire, theft and other casualties (Anandhi et al., 2016; Moschini and Hennessy, 2001). Secondly, prices are connected to the global market and therefore unforeseen factors, anywhere in the world, such as weather conditions or government measures, can cause dramatic changes in output and input prices. Institutional risks include political risks, sovereign risks, i.e. risks caused by the actions of foreign governments such as failure to comply with a trade agreement, and contractual risks, i.e. risks inherent in transactions between trading partners and other commercial organisations.

Finally, the people who operate the farm can also be a source of risk to the farm business. Human risk includes personal health and well-being, family and work relationships and employee management.
In contrast, the financial risk category results from the way that agricultural businesses are financed (Gabriel and Baker, 1980; de Mey et al., 2016). It includes the cost and availability of capital and the ability to meet cash flow demands and absorb short-term financial shocks. The cash flow is particularly important because of various ongoing expenses such as input costs, tax payments, debt repayment and personal living expenses.

In some studies, variability in crop production and agricultural commodity prices are identified as the most important source of risk (Ortmann et al., 1992). Thus, the two main risks related to price and yield volatility will be the focus of interest in the work of this thesis. The risk of poor yields is mainly due to adverse natural events such as drought, hail, frost, rainfall, floods, landslide, insect infestation, plant diseases and also to the agricultural techniques implemented by farmers (Goodwin and Hungerford, 2014; Wang et al., 2020; Coble and Knight, 2002). The price risk is more related to the deregulation of financial markets (Johnson, 1975; Chavas, 2011), explained by the fact that most European countries have moved from market-based support to decoupled direct payments. Then, producers are exposed to high price volatilities on world commodity markets (El Benni et al., 2016). Finally, price and yield are very important sources of risk in agriculture and are among the main objectives of risk management.

Category	Type of risk	Causes
Business	Production	Weather risks, pests and diseases, technology change, yields, etc.
	Price or market	Output and input price fluctuation, market shocks, etc.
	Institutional	Political, sovereign and contractual risks.
	Human/personal	Death, illness or injury of farmers or farm workers.
Financial	Financial	Loans and credits.

Table 1.1. Summary of risks in agriculture (Hardaker et al., 2015)

1.1.2 Agricultural risk management tools

Risk management in agriculture is very crucial to provide the best coverage of the farmer's income and the security of other sectors of the economy (Ullah et al., 2016).
Farmers have several options for coping with yield variability. The first is to control or reduce risk through the use of inputs and techniques such as fertilisers (Serra et al., 2003), pesti-

cides (Eidman, 1990), or irrigation which is very effective in minimising the effects of low rainfall or drought (Foudi and Erdlenbruch, 2012). The second is to reduce production variability through crop diversification (Finocchio and Esposti, 2008), vertical integration (Whitson et al., 1976) or the application of improved technology (Kim and Chavas, 2003). Another way to manage production risk is to transfer some or all of the risk to someone else. Insurance policies are an effective mechanism for transferring risk (Rejda, 2011; Hardaker et al., 2015). It provides protection against an adverse event or unexpected loss. For example, crop insurance is a very important type of insurance that guarantees a level of production, thus eliminating the risk associated with forward pricing (Velandia et al., 2009). It provides the financial support to fulfil a commitment if the insured crop suffers a loss before harvest. Furthermore, a disability policy could also be a good risk management tool to deal with the potential disability of farmers. There are other types of insurance contracts available for farmers, including public liability, life, mortality, injuries, health, fire and theft cover for assets (Rejda, 2011; Hardaker et al., 2015).

Farmers can use futures hedging (Shapiro and Brorsen, 1988), forward contracting (Goodwin and Schroeder, 1994) and options markets (Makus et al., 1990) to manage the risk of price volatility. The use of financial hedging measures is also of great importance. This is due to the fact that financial risk affects the solvency of the firm (debt/equity ratio) or its liquidity positions (Ullah et al., 2016). To cope with this risk, some financial responses for farmers include (Eidman, 1990): holding assets for sale to meet cash demands and maintaining liquid credit reserves, among others.

On the European side, the reform of the Common Agricultural Policy (CAP) in 2013 has implemented some risk management tools for the member states. These tools include (García Azcárate et al., 2016): insurance, mutual funds, savings accounts, Ad-Hoc payments and fiscal measures. Insurance policies provide farmers with an indemnity, an amount linked to a certain loss calculation, in case of adverse events. These can be specific events in the case of single peril crop insurance (for example, hail or frost) or a number of weather events for combined risk insurance. There are other types of insurance for individual farms (Schaffnit-Chatterjee et al., 2010). One can mention yield insurance which covers yield losses of a given crop due to any type of climatic adversity. Multi-peril insurance covers the crop when production falls under a certain threshold. Whole-farm yield insurance covers all crops produced by the farm, except that the farmer is only eligible to indemnity if the overall production falls under a certain threshold. Revenue insurance combines yield and price coverage, while income insurance additionally covers production costs.

Mutual funds provide a way for a group of producers to share risk. They are financial reserves built up from participants' contributions (a fixed amount independent of the risk) that pro-

vide compensation in case of major income losses. The Income Stabilisation Tool (IST) is an important tool that provides financial support to mutual funds, which compensate farmers for income losses due to production, price and/or cost risks (Hine et al., 2016; Meuwissen et al., 2018; El Benni et al., 2016). The income loss is defined as a decrease of 30% of the expected income over the previous three years average or a five years Olympic average (excluding the highest and the lowest). This mutual fund provides compensation to farmers for a maximum of 70% of the income loss. Savings accounts are another management tool in which farmers can deposit part of their annual income into a special account that guarantees interest payments. Ad hoc payments are provisions that help farmers to rebuild their capital in the case of catastrophes. Finally, fiscal and tax measures can also lead to income stabilisation. For example, taxes can be reduced for farmers affected by climatic or market risks.

However, these risk management tools focus mainly on yield variability and neglect price stabilisation tools, such as futures and forward contracts, options, or revenue contracts that combine price and yield variations. Indeed, price risks can be hedged in futures and options markets, which are efficient markets for systemic price risks (Purcell et al., 1991). Nevertheless, the use of these markets in Europe is not yet widespread (Meuwissen et al., 2011). The main reasons are related to the existence of CAP support programmes and farmers' lack of education on futures and options markets. As for insurance, it plays a limited role in Europe regarding price risks (Meuwissen et al., 2018). Indeed, insurance is widely available for personal and production risks such as yield variability, rather than price variability. Some form of revenue insurance can be a very useful risk management tool. This type of contract is well developed in the USA and Canada (Diaz-Caneja et al., 2008), whereas it is not widely spread in Europe (Meuwissen et al., 2003). For instance, the Farm Bill in the USA comprises several insurance systems that hedge yield and income losses. Premiums are affordable for farmers because they are highly subsidised. On the other hand, the IST measure has not yet been developed in Europe except in three Member States: Hungary, the region of Castilla y Leon in Spain and Italy (Trestini et al., 2018). Moreover, this instrument could face problems in setting appropriate income triggers (Finger and El Benni, 2014), information asymmetries[1] (Meuwissen et al., 2011), production distortions (Mary et al., 2013) and the potentially huge cost of implementation.

The case of France is particularly interesting as it is the largest agricultural producer in the European Union (Enjolras and Sentis, 2011). France has set up two risk management tools: a private crop insurance system, where premiums can be subsidised up to 65%, and a national public fund for the mutualisation of health and environmental risks (called FMSE[2]). These

[1] Information asymmetry occurs when the potential insured knows more about the risk to be insured than the insurer. It can lead to the dual problems of moral hazard and adverse selection (Meuwissen et al., 2011).

[2] *Fonds de Mutualisation Sanitaire et Environnementale.*

tools have a limited effect, as French producers continue to receive European payments from the CAP of around 7 billion € per year, which is disconnected from market and weather trends (Lidsky et al., 2017). Insurance coverage remains quite low despite the important subsidies dedicated to multi-peril crop insurance and very recently the creation of a "*contrat socle*[3]". Regarding the FMSE, it remains even more limited in scope.

1.1.3 Revenue insurance contract

The design of revenue insurance seems to be a more attractive option to cover the farm income and cope with future volatility and uncertainty (Hennessy et al., 1997). In particular, it offers better protection against yield and price risks than current crop insurance policies. Although France is the largest agricultural producer in the European Union, it has not yet implemented revenue insurance. Thus, ongoing reforms are encouraging professional bodies and governments to develop such agricultural insurance products that take into account both yield and price risks, which are key determinants of revenue. They aim at increasing the attractiveness of revenue insurance compared to other risk management tools implemented for different reasons. First, public subsidies for revenue insurance seem justified because the risk covered is probably systemic, i.e. many farmers are exposed to the risk at the same time, thus allowing for public transfers (El Benni et al., 2016; Meuwissen et al., 2003). Second, correlations between prices and yields are implicitly considered by a total farm revenue insurance, which seems advantageous compared to single-crop yield or price risk management instruments (El Benni et al., 2016). Indeed, the relevance of taking risks into account jointly is manifest in that negative correlation (e.g. low yields with high prices), also called natural hedge, could lead to a natural stabilisation of the revenue and to low fair premium rates. In the case of a positive relationship in a low price market (low yields with low prices), revenue contracts would have higher premiums but will provide better coverage of the producer's revenue.

The main issue in implementing revenue insurance is the calculation of an actuarially fair premium, taking into account the dependence structure between price and yield risks. In Europe, many research articles focus on actuarial evaluations of potential revenue insurance, the resulting costs to the government or conceptual studies on issues of adverse selection and moral hazard (Meuwissen et al., 2003). However, the literature on modelling the dependence between price and yield risks within the revenue assurance scheme is sparse. One example is the study of Ahmed and Serra (2015) which evaluated the economic impacts of the implementation of agricultural revenue insurance in Spain and assessed the dependence between prices and yields. Therefore, it is relevant to model the interaction between the revenue underlying

[3]The *contrat socle* is a kind of reference crop insurance launched in France (García Azcárate et al., 2016).

risks, with implications for evaluating the implementation of revenue insurance.

1.1.4 Modelling the underlying risks of revenue insurance

Modelling the dependence between yields and prices is of great concern. It may have implications for the eventual implementation of revenue insurance that would provide joint price and yield coverage. The dependence between these risks is often measured using the Pearson product moment correlation coefficient (Embrechts et al., 2002). This correlation is widely applied as a measure of linear dependence in multivariate normal (more generally elliptical) distributions. For instance, Coble et al. (2000) have studied the dependence between these risks in the context of revenue insurance, using the Pearson correlation coefficient. However, this correlation suffers from many deficiencies (Embrechts et al., 2002). The independence of two random variables implies that they are uncorrelated, but a null correlation does not imply independence unless the distributions are multivariate normal. In addition, the variance of the variables must be finite, which causes problems when dealing with heavy-tailed distributions. Also, the correlation is not invariant under non-linear strictly increasing transformations of variables. Non-parametric measures of association such as Kendall and Spearman rank correlations are alternative measures of dependence that are invariant to monotone transformations and does not rely on an assumption of linearity (Schweizer and Wolff, 1981). The work of Ramsey et al. (2019) used various non linear measures of association, including Spearman and Kendall correlations to assess the dependence between prices and yields and to examine the sensitivity of premium rates for all maize producing counties in the US. However, these measures of correlation are limited since they characterise the dependence over the entire support of the variables. In contrast, many risk management issues focus on risk in the distribution tail. Besides, it is highly important that the joint distribution of yields and prices is formed and evaluated for actuarial purposes (e.g. pricing revenue insurance contract).

An alternative measure of dependence to overcome these limitations is the use of copula functions. It is a promising tool, which was originated by Sklar (1959), that combine the marginal distributions of variables to form the joint distribution (multivariate distribution) with the dependence structure. This is important because of the scarcity of multivariate distributions available in the statistical literature. Copulas also allow incorporating both linear and non-linear dependence and describing properties of extreme values such as heavy-tails, extreme co-movement, and tail dependence. In the literature, the copula model has been widely used for risk management, insurance, and financial economics issues. In the case of agricultural income insurance, its use to model the dependence structure between yields and prices is relatively recent. Among others, Zhu et al. (2008) assessed the dependence between

yields and prices of maize and soybean crops using copulas, to provide an efficient insurance contract for the whole farm revenue. Rusyda et al. (2021) modelled the variability of revenue risk by implementing a copula towards crop yield and price to provide an alternative method of multi-crop revenue insurance in Indonesia. Regarding the revenue insurance in Europe, only one work of Ahmed and Serra (2015), to the best of our knowledge, which used copulas to jointly model price and yield perils in the orange and apple sectors in Spain.

Regarding individual yield and price modelling, many research papers focus on modelling individual yields and prices. Samuelson (2015) models prices by a lognormal while Tejeda and Goodwin (2008) use a Burr distribution. For crop yield modelling, Ozaki et al. (2008) use the normal and beta parametric distributions and also the non-parametric kernel estimator. Other parametric techniques have been applied for estimating yield distributions, such as the gamma distribution (Gallagher, 1986), the lognormal distribution (Stokes, 2000) and the Johnson S_u family (Ramirez et al., 2003). A variety of kernel functions have also been used in the estimation (Ker and Goodwin, 2000; Ker and Coble, 2003). Classical statistical tools, including parametric and non-parametric methods, focus mainly on studying the average behaviour of distributions. Unfortunately, these tools fail to capture tail risks since the basic statistical measures of risk are based on the average. The underestimation of tail distributions can lead to inaccurate pricing when designing a revenue or a crop insurance programme and bias the calculation of indemnities. Indeed, insufficient pricing and/or inadequate coverage of the insured risk can lead to a series of undesirable effects for the insurer (underestimation of the risk), and for the government who has to intervene in the case of an agricultural crisis to save the crop insurance sector (Stokes, 2000). Thus, to overcome the limitations of classical methods in modelling tail risk, extreme value theory is a powerful statistical framework that provides many tools to study the extreme tails of distributions. Despite the increasing use of extreme value theory in several fields, its application to agricultural risk management related to prices and yields has so far been sparse in the academic literature. For extreme price risk, van Oordt et al. (2021) applied extreme value approach to estimate the size and probability of price spikes in agricultural commodities. Morgan et al. (2012) estimated three tail quantile-based risk measures for corn and soybean production in the US. Fretheim and Kristiansen (2015) applied the extreme value theory to commodity market risk from 1995 to 2013 using commodity prices of food. An analysis of the estimated shape parameters of the Generalised Extreme Value distribution (GEV) has been conducted. For extreme yield risk, the literature has been scarce. One can mention Mitchell et al. (2020) who used a Generalised Pareto Distribution (GPD) for the estimation of regional highest yields and highest yields as a function of pending on agricultural inputs. Performing extreme value analyses of yields and prices depending on other factors such as climate, production costs and agricultural inputs

would be of great interest.

This thesis tackles the first question by using the copula approach to model the joint distribution of yields and prices and the associated dependence structure, in the cereal and wine sector in France. We also focus on modelling the dependence structure depending on other factors (meteorology and risk management tools). Furthermore, the potential of implementing a revenue insurance scheme to cope with these risks will be assessed. Then, this thesis addresses the second issue by developing novel methods based on extreme value theory to model the conditional tail of price and yield distributions.

1.1.5 French farm income database

The overall work of this thesis is based on farm-level production data from the European Farm Accountancy Data Network (FADN, RICA-Agreste). Data are accounted for each year from a representative sample of farms, the size of which can be considered commercial[4]. FADN database provides significant accounting and financial information on French professional farms such as: balance sheet, income statement, crop insurance expenses, agricultural inputs and characteristics of the farm operator and the farm structure. Within the original database, we selected farms specialised in wheat, maize and quality wine-growing productions, over an observable period between 2014 and 2016. The choice of these years is explained by the fact that French cereal (wheat and maize) production in 2014 and 2015 reached a high record in a low prices market context. As for 2016, the year was characterised by a drop in harvests due to spring storms and summer drought, which led to lower yields as well. Our sample finally includes 6 334 observations of 2 041 farms. We combine the dataset with information on climate and weather from *Météo France* weather stations[5], matched at the regional level.

In the following we explain the choice of the sectors considered, namely wheat, maize and wine-growing. Then, we present the weather and the financial conditions of the global markets over the time period considered and for each crop. Finally, we detail the main explanatory variables that enter the different analyses of this thesis.

Choice of considered sectors. Wheat is the prominent cereal produced in France. It is mostly located in the West of France and around the Parisian basin. France is the first European producer and exporter of wheat and it is ranked fifth largest country in the world in terms of national wheat production (Ben-Ari et al., 2018). This is due to the very high yields, about 7.4 t/ha, compared to the world's four largest wheat producers, such as Russia and

[4]The commercial size is specified according to Regulation 79/65/EEC of 15 June 1965. The farm must be large enough to provide a main activity for the farmer and a level of income sufficient to support his or her family.

[5]It listed by the French Ministry of Environment, Ecology and Sea, see www.stats.environnement.developpement-durable.gouv.fr/Eider.

the United States, which harvest about 5 and 3 t/ha of wheat respectively. Maize is the second largest crop production in France, cultivated on more than 3 million hectares in 2016. Thanks to favourable soil and climate conditions and the performance of producers, France is also the world's largest exporter of maize seeds (Ben-Ari et al., 2018).

Wine-growing comes in second place after cereals in terms of yields. In terms of wine production, France occupies the first place worldwide along with Italy and Spain, depending on years. The production value in France amounts to over 12,4 billion euros (among which 79% for quality wines). French viticulture is a leading production mostly based on family farms. In spite of the slight decrease of wine consumption every year, prices increase regularly thanks to exports. The two main concepts related to French quality wines are the concept of "*terroir*"[6] and the controlled designation of origin system (*Appellation d'Origine Contrôlée*–AOC)[7].

Weather and market conditions. Weather conditions in autumn 2014 and summer 2015 had very contrasting effects on cereals[8]. Winter crops such as wheat had high yields, unlike autumn crops such as maize which suffered from drought and summer heat waves. However, the French wheat record harvest occurred within an abundant global context. Thus, wheat price dropped at the same time on global markets. However, the drop of the euro against the dollar supported the prices of agricultural commodities exchanged in euros. For maize, despite the decrease in production, global stocks remained high. In 2016, cereal production suffered greatly in France due to climatic conditions (bad weather in spring and drought in summer) which led to significant yield decline. Despite the poor harvests in France, cereal prices remained low, due to the abundance of world production[9].

Thanks to mild temperatures in winter and spring 2014, wine production increased by 17% for AOC wine. At the same time, production stocks at the beginning of the 2014/2015 wine year were lower than in the previous year (-10%) for all wine categories. Along with a reduced dynamic of foreign trade, prices of AOC wine felt sharply at the beginning of the year before stabilising, while they increased for other wines. Year 2015 was characterised by a slight increase in harvest levels but stable and limited availability, especially for AOC wines. Prices increased slightly compared to 2014. In 2016, several vineyards were severely affected by several weather accidents and the impact on harvests was very significant. However, in the

[6]The concept of "*terroir*" was first developed in the 14th century in the Bourgogne region of France, to identify the qualities of wines in terms of geoclimatic origin and authentic production methods (Whalen, 2009).

[7]The *Appellation d'Origine Contrôlée* is a French label that guarantees the place of origin and defines a set of production requirements to identify the quality.

[8]See the reports on the agricultural economic situation in France for 2014 https://agreste.agriculture.gouv.fr/agreste-web/disaron/BilanConj2014/detail/ and 2015 https://agreste.agriculture.gouv.fr/agreste-web/disaron/BilanConj2015/detail/ (in French).

[9]See the report on the agricultural economic situation in France for 2016 https://agreste.agriculture.gouv.fr/agreste-web/disaron/BilanConj2016/detail/ (in French).

first nine months, prices of AOC wine were dynamic (+7.5% year-on-year), and systematically above the 2015 prices.

The following table summarises the situation of local crop yields and prices fluctuations in the global market for cereals and the local market for wine.

Crop	Risk	2014	2015	2016
Wheat	Yield	+ Production rise	+ Production rise	− Production drop
	Price	− Abundant world production	− Abundant world production	− Abundant world production
Maize	Yield	+ Production rise	− Production drop	− Production drop
	Price	− Abundant world production	− Production drop but high stocks	− Abundant world production
Wine	Yield	+ Production rise	+ Production rise	− Production drop
	Price	0 Stable prices	− Prices drop	+ Controlled prices

Table 1.2. Crop yields and financial market prices situation of the cereal and wine sectors from 2014 to 2016. Increase: +, Decrease: −, Same level as the previous year: 0.

Variables description. We chose a wide range of potential factors, including agricultural inputs, insurance, production costs and meteorological variables. The list of variables involved in the work of this thesis is given in Table 1.3.

Variable	Unit	Description
Farm specialisation	-	3 types (wheat, maize, quality wine-growing)
Gross product	Euro	Gross product of the considered crop
Gross yield	Quintal or hL	Gross yield of the considered crop
Harvested acreage	ha	Cultivated area of the considered crop
Altitude	-	Altitude of the farm (3 classes)
Yield	Quintal or hL/ha	Yields divided by the acreage
Price	Euro/Quintal or hL	Gross product divided by quintals or hectolitres sold
Pesticide	Euro/ha	Pesticide price/Cultivated area
Fertilizer	Euro/ha	Fertiliser price/Cultivated area
Crop insurance	1/0	The Farm purchased or not a crop insurance policy
Premiums	Euro/ha	Crop insurance premiums/Cultivated area
Claims	1/0	The Farm received or not some crop insurance claims
Insurance claims	Euro/ha	Crop insurance claims received/Cultivated area
Other premiums	Euro/ha	Other insurance premiums/Cultivated area
Subsidy	Euro/ha	Farm subsidies/Cultivated area
Seeds and plants	Euro/ha	Seeds and plants costs/Cultivated area
Works and services	Euro/ha	Works and services costs/Cultivated area
Income taxes	Euro/ha	Taxes expenses/Cultivated area
Personal costs	Euro/ha	Personal social security costs/Cultivated area
Temperature	°C	Deviation from the average of the last five years
Precipitation	mm	Deviation from the average of the last five years
Sunshine duration	Hour	Deviation from the average of the last five years

Table 1.3. Database description

1.2 Introduction to dependence modelling and copulas

The concept of copulas was introduced in 1959 by Abe Sklar. During the financial crisis of 2007 and 2008, copulas have come to the attention of the general public due to their use in the modelling of multidimensional phenomena, mainly in the realm of quantitative risk management. They are a flexible tool for studying the dependence between several random variables, with the idea that this dependence should not contain any information from the marginal distributions of the variables. Their applications are widespread in many fields such as insurance (Diers et al., 2012), finance (Salmon and Schleicher, 2006), economics (Wali et al., 2018), hydrology (De Michele et al., 2005), biology (Emura and Michimae, 2017), etc. In agriculture sector, the risks linked to climate disruption and financial market deregulation are multidimensional and therefore require the joint modelling of several random variables. Usually, the classical families of bivariate distributions are used to meet this need, for instance, we can mention the bivariate Normal, Log-normal, Gamma and extreme-value distributions (Genest and Favre, 2007). The main limitation in such approaches is that the individual behaviour of each variable has to be characterised by the same parametric family of univariate distributions. Copula models avoid this restriction and have thus become widely used in the literature.

In Paragraph 1.2.1, we give the definition of a Copula and some basic properties. Some dependence measures are provided in Paragraph 1.2.2 and the main classes of copulas are discussed in 1.2.3. Then, we present some methods to estimate Copulas in Paragraph 1.2.4 and Goodness-of-fit tests in 1.2.5. Finally, in Paragraph 1.2.6, we present Copula conditionally on covariates.

1.2.1 Definition and basic properties

By definition, a copula is a multivariate distribution function with standard uniform univariate margins. For the sake of simplicity, let us focus on the bivariate case.

Definition 1.2.1. *A copula of dimension 2, $C(u, v)$ is a function of $[0,1]^2 \to [0,1]$ such that:*

- *(Uniform marginals)* $\forall u, v \in [0,1]$,

$$C(u,0) = 0, C(u,1) = u, C(0,v) = 0, C(1,v) = v.$$

- *(Increasing)* $\forall u_1, u_2, v_1, v_2 \in [0,1]$ *such that* $u_1 \leq u_2$ *and* $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - (C(u_1, v_2) - C(u_1, v_1)) \geq 0.$$

We consider a pair of continuous random variables X and Y marginally distributed according to $F(x) = \mathbb{P}(X \leq x)$ and $G(y) = \mathbb{P}(Y \leq y)$. Let $H(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ be their joint distribution function. Sklar's theorem (Sklar, 1959) is the basis of Copulas and is stated as follows:

Theorem 1.2.1 (Sklar). *There exists a copula* $C : [0, 1]^2 \to [0, 1]$ *such that:*

$$H(x, y) = C(F(x), G(y)), \quad \text{for all } (x, y) \in \mathbb{R}^2. \tag{1.1}$$

If F *and* G *are continuous, then* C *is unique. Otherwise,* C *is uniquely determined on Ran* $F \times$ *Ran* G *(Ran* F *= support of* F*).*
Conversely, for any univariate distribution functions F *and* G *and any copula* C*, the function* H *is a two-dimensional distribution function with marginals* F *and* G*.*

The copula C is the bivariate cumulative distribution function (cdf) of the random vector $(F(X), G(Y))$ with uniform margins on $[0, 1]$. The mappings $X \mapsto U := F(X)$ or $Y \mapsto V := G(Y)$ are usually referred to as the probability-integral transformations (to uniformity), if X and Y are continuous, and are standard tools for simulation purposes. Theorem 1.2.1 implies that for a continuous multivariate joint distribution, the margins and the dependence structure can be uniquely dissociated. Moreover, the dependence structure is represented by the copula C.

Theorem 1.2.2 (Fréchet-Hoeffding bounds). *Let* $M(u, v) = \min(u, v)$ *and* $W(u, v) = \max(u + v - 1, 0)$*. Thus for every Copula* C *and every* $(u, v) \in [0, 1]^2$*,*

$$W(u, v) \leq C(u, v) \leq M(u, v). \tag{1.2}$$

We refer to M as the Fréchet-Hoeffding upper bound, which represents the perfect positive dependence copula, and W as the Fréchet-Hoeffding lower bound, representing the perfect negative dependence (the perfect dependence property is also called co-monotonicity). In the case of independence, the copula is given by $C(u, v) = uv$ and is denoted by the symbol Π, i.e. $\Pi(u, v) = uv$. Moreover, Copulas are invariant under monotonically increasing transformations of their margins.

1.2.2 Bivariate dependence measures

Several measures of association between the components of a random pair can be considered, Kendall's Tau (Nelsen, 2007) [paragraph 5.1.1], and Spearman's Rho (Nelsen, 2007) [paragraph 5.1.2] being the most popular ones. The difference between these two coefficients and the usual "correlation coefficient" is that the second term refers usually to the linear

dependence between random variables, for instance Pearson's correlation coefficient. In contrast, these two measures are used rather to measure the "association". They are invariant to strictly increasing functions and can be interpreted as probabilities of concordance minus probabilities of discordance of two random pairs. Both of them can be written only in terms of the copula C:

$$\tau = 4\mathbb{E}(C(U,V)) - 1 = 4\int_0^1\int_0^1 C(u,v)\,dC(u,v) - 1, \tag{1.3}$$

$$\rho = 12\mathbb{E}(UV) - 3 = 12\int_0^1\int_0^1 uv\,dC(u,v) - 3. \tag{1.4}$$

These two dependency coefficients are equal to 1 when the dependency is positive and perfect, -1 when the dependency is negative and perfect, and 0 in the case of independence. Let us note that ρ coincides with the correlation coefficient between the uniform marginal distributions. Another measure of association based on concordance called *medial correlation coefficient*, was proposed by Blomqvist (Nelsen, 2007) [paragraph 5.1.4], and is given by:

$$\beta = 4C\left(\frac{1}{2},\frac{1}{2}\right) - 1. \tag{1.5}$$

This parameter quantifies the probability that X and Y would jointly exceed their median value. There is also the tail dependence coefficient to measure the degree of dependence in the upper right or lower left quadrant of a bivariate distribution. This concept is based on the study of the dependence between the extreme values. The coefficient of upper tail dependence of (X,Y) is given by:

$$\lambda_U = \lim_{u\nearrow 1}\mathbb{P}\{Y > G^{-1}(u)|X > F^{-1}(u)\} \tag{1.6}$$

$$= \lim_{u\nearrow 1}\frac{1-2u+C(u,u)}{1-u}. \tag{1.7}$$

If $\lambda_U \in (0,1]$ then X and Y are said to be asymptotically dependent in the upper tail. Otherwise, if $\lambda_U = 0$ then they are said to be asymptotically independent in the upper tail. Similarly, we define the lower tail dependence coefficient:

$$\lambda_L = \lim_{u\searrow 0}\mathbb{P}\{Y \leq G^{-1}(u)|X \leq F^{-1}(u)\} \tag{1.8}$$

$$= \lim_{u\searrow 0}\frac{C(u,u)}{u}. \tag{1.9}$$

1.2.3 Main Copula models

Numerous parametric families of copulas can be found in the literature. Elliptical and archimedian Copulas are the two most popular models. Elliptical copulas (Frahm et al., 2003) are built from elliptical distributions thanks to an uniformization of their margins. The level sets of an elliptical distribution density are ellipses whose shape is determined by a (kind of) covariance matrix. This family has a lot of properties in common with the multivariate normal distribution. Important examples in this family are the Gaussian and the Student copulas.

Definition 1.2.2 (Bivariate Gaussian Copula)**.** *The Gaussian copula is defined as:*

$$C_r^G(u,v) = \Phi_r\left(\Phi^{-1}(u), \Phi^{-1}(v)\right) \tag{1.10}$$

Φ^{-1} *and* Φ_r *refer to the inverse distribution of the standard normal distribution and the bivariate standard normal distribution with linear correlation* r*. Thus,*

$$\Phi_r\left(\Phi^{-1}(u), \Phi^{-1}(v)\right) = \frac{1}{2\pi\sqrt{1-r^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \exp\left(-\frac{x^2 - 2rxy + y^2}{2(1-r^2)}\right) dxdy.$$

Kendall's tau and Spearman's rho for a Gaussian Copula are given by:

$$\tau = \frac{2}{\pi} \arcsin r \tag{1.11}$$

$$\rho = \frac{6}{\pi} \arcsin \frac{r}{2}. \tag{1.12}$$

There is no explicit form to the distribution function of the bivariate normal distribution. Moreover, Gaussian Copulas do not catch the dependence in the distributions tails: both the lower and upper tail dependence parameters λ_U and λ_L are null.

Definition 1.2.3 (Bivariate Student Copula)**.** *The Student-*$t_{r,\nu}$ *Copula is defined as:*

$$C_{r,\nu}^t(u,v) = t_{r,\nu}\left(t_\nu^{-1}(u), t_\nu^{-1}(v)\right). \tag{1.13}$$

t_ν *is the distribution function of the univariate Student distribution with* ν *is degrees of freedom and* $t_{r,\nu}$ *is the bivariate Student distribution function, with the dependency parameter* r*, defined as:*

$$t_{r,\nu}\left(t_\nu^{-1}(u), t_\nu^{-1}(v)\right) = \frac{\Gamma(\frac{\nu+2}{2})}{2\pi\Gamma(\frac{\nu}{2})\sqrt{1-r^2}} \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \left(1 + \frac{x^2 - 2rxy + y^2}{\nu(1-r^2)}\right)^{-\frac{\nu+2}{2}} dxdy.$$

To our knowledge, there is no formula that gives the equivalent between the linear cor-

relation coefficient r and Spearman's rho for Student distributions. For Kendall's tau we have,

$$\tau = \frac{2}{\pi} \arcsin r. \tag{1.14}$$

The lower and upper tail dependence parameters are expressed as

$$\lambda_U = \lambda_L = 2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-r}{1+r}}\right) \tag{1.15}$$

In the limit, when $\nu \to \infty$, the Student Copula tends towards the Gaussian Copula. As with the Gaussian Copula, the Student Copula has no explicit form.

The family of Archimedean Copulas (Naifar, 2011) is an important class with a large variety of parametric families and a large possibility of dependence structures (Genest, 1987). They are determined by a univariate function, called the generator, whatever the dimension is.

Definition 1.2.4. *A copula C is said to be Archimedean if there exists a convex, decreasing function $\phi : [0,1] \to [0,\infty)$ such that $\phi(1) = 0$ and*

$$C(u,v) = \phi^{-1}(\phi(u) + \phi(v)). \tag{1.16}$$

A number of generators have been proposed, involving on one or two parameters and tuning the dependence strength between the marginals. Table 1.4 gives the generator ϕ and an expression for τ and ρ for the three Archimedean models: Gumbel, Frank, and Clayton (Beck, 2015) [pages 17-21]. The parameter of the Copula is called θ.

Family	Generator	Parameter	Kendall's tau	Spearman's rho
Gumbel	$\lvert\log(t)\rvert^\theta$	$\theta \geq 1$	$[0,1[$	$[0,1[$
Frank	$-\log(\frac{e^{-\theta t}-1}{e^{-\theta}-1})$	$\theta \in \mathbb{R}$	$]-1,1[$	$]-1,1[$
Clayton	$(t^{-\theta}-1)/\theta$	$\theta \geq -1$	$[-1,1[$	$]-1,1[$

Table 1.4. Three families of Archimedean Copulas with their generator, parameter space and Kendall's Tau and Spearman's rho space.

1.2.4 Estimation of copulas

Starting from a sample $(U_1,V_1),\ldots,(U_n,V_n)$ of independent observations from C, Spearman ρ and Kendall τ can be estimated by its empirical counterparts as

$$\hat{\rho} = \frac{12}{n}\sum_{i=1}^{n} U_i V_i - 3. \tag{1.17}$$

$$\hat{\tau} = \frac{4}{n}\sum_{i=1}^{n} U_i V_i - 1. \tag{1.18}$$

Three main approaches have been proposed for estimating copulas: parametric, semi-parametric and non-parametric methods (Genest and Favre, 2007). First, the parametric approach is based on the estimation of the parameters(s) θ of the copula assumed to belong to some parametric family $\{C_\theta,\ \theta \in \Theta\}$. The estimation of the parameter(s) θ can be done for instance using the maximum likelihood method or the method of moments. In the latter case, θ is estimated by minimizing a given distance between the empirical $\hat{\tau}$ and $\hat{\rho}$ computed from Equations (1.17)-(1.18) and the theoretical ones $\tau(\theta)$ and $\rho(\theta)$ calculated according to Equations (1.3) and (1.4) under the model C_θ.

Second, the semi-parametric approach does not assume that the margins F and G belong to any parametric family. They are estimated directly by the non-parametric estimator given by,

$$\hat{F}(x) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}. \tag{1.19}$$

Then, to estimate the parameter θ, there are two main strategies. The first is based on maximising a certain likelihood function, and the second is a method of moments based on dependence coefficients.

Third, to dispense with the assumption that the Copula belongs to a parametric family, one can use non-parametric approach for estimation (Deheuvels, 1981). Most of methods are based on the empirical copula defined by,

$$\hat{C}(u,v) = \hat{H}\left(\hat{F}^{-1}(u), \hat{G}^{-1}(v)\right), \tag{1.20}$$

where $\hat{H}$ is the estimator of the distribution function H, for example the empirical distribution function or an estimator constructed using a kernel. $\hat{F}^{-1}$ and $\hat{G}^{-1}$ are non-parametric estimators of the quantile function.

1.2.5 Goodness-of-fit tests

Two main techniques can be used to select the copula that fits best a dataset. First, one can use graphical diagnostics such as Rosenblatt's transformation (Hofert and Mächler, 2014). The main idea is to transform pairs of data columns towards bivariate standard uniform distributions under the null hypothesis $C \sim C_0$ where C is represented by a specific copula C_0. Then, the p-value of an independence test is computed and encoded as background color (Hofert and Mächler, 2014). This method is usually used in higher dimensions to detect the pair that violate the null hypothesis.

Second, one may use a goodness-of-fit test based on Kendall's distribution function K (also called multivariate probability integral transformation) (Genest et al., 2009) defined as,

$$K(t) = \mathbb{P}(C(U,V) \leq t) = \int_0^1 \int_0^1 \mathbb{1}_{\{C(u,v)\leq t\}} dC(u,v). \tag{1.21}$$

The theoretical Kendall distribution under the null hypothesis $(U,V) \sim C_0$ is compared to its sample version thanks to a Cramér–von Mises statistics (Genest and Favre, 2007) and the associated p-value is computed. The Cramér–von Mises statistic is given by:

$$S_n = n \int_0^1 \{K_n(w) - K_{\theta_n}(w)\}^2 dw, \tag{1.22}$$

where K_n is the empirical Kendall distribution function and K_{θ_n} denotes the parametric Kendall cumulative function under the null hypothesis with θ_n a consistent estimator of its parameter θ. The computation of p-values requires a bootstrap method to find the underlying distribution of the statistics S_n under the null hypothesis (i.e when the data are described by the copula).

1.2.6 Conditional copulas

In some cases, the dependence structure of the random pair (X,Y) may depend on an external (possible multivariate) random variable Z. Conditional copulas were introduced to tackle this issue (Gijbels et al., 2011). Similarly to Equation (1.1), one can write the joint and marginal distribution functions of (X,Y) conditionally on $Z = z$, as:

$$H_z(x,y) = \mathbb{P}(X \leq x, Y \leq y | Z = z) = C_z(F_z(x), G_z(y)) \tag{1.23}$$

where $F_z(x) = \mathbb{P}(X \leq x | Z = z)$ and $G_z(y) = P(Y \leq y | Z = z)$. In this context, C_z is referred to as a conditional copula. Starting from a set of observations $(X_1, Y_1, Z_1), ..., (X_n, Y_n, Z_n)$, a non-parametric estimator of $C_z(u,v)$ can be considered (Gijbels et al., 2011):

$$\hat{C}_z(u,v) = \sum_{i=1}^{n} w_i(z,h) \mathbb{1}_{\{\hat{F}_z(X_i)\leq u, \hat{G}_z(Y_i)\leq v\}}. \tag{1.24}$$

Here, $w_i(z,h)$ is a sequence of weights selecting the observations (X_i, Y_i) such that the associate covariate Z_i is close to the estimation point z. The range of the selected points is tuned by the parameter h called the bandwidth. The margin distributions F_z and G_z are estimated using similar smoothing techniques:

$$\hat{F}_z(x) = \sum_{i=1}^{n} w_i(z,h) \mathbb{1}_{\{X_i \leq x\}}, \ \hat{G}_z(y) = \sum_{i=1}^{n} w_i(z,h) \mathbb{1}_{\{Y_i \leq y\}}. \tag{1.25}$$

The conditional copula can be used to estimate conditional Spearman's ρ and Kendall's τ providing then association measures depending on the covariate Z. As an example, Spearman's ρ (1.4) is extended to the conditional framework as

$$\rho(z) = 12 \int_0^1 \int_0^1 C_z(u, v) du dv - 3, \tag{1.26}$$

and the associated estimator $\hat{\rho}(z)$ is obtained by plugging the estimated conditional copula (1.24) in the previous Equation (1.26).

1.3 Dimension reduction

Regression analysis is widely used to study the relationship between a response variable Y and a p-dimensional vector X of covariates starting from a n-sample. When the dimension p grows, it is well-known that the space becomes sparsely populated with data points. This issue is referred to as the "curse of dimensionality". Thus, in a large dimension setting, a dimension reduction becomes necessary to overcome the curse and to reveal the most relevant directions of the high-dimensional covariate space. In the statistical literature, there are three important classes of dimension reduction methods in regression problems. The first is the variables selection method, where only a few variables are selected that are truly related to the response variable Y. Second, the shrinkage approach consists of fitting a model with all covariates using a technique that regularises or reduces the coefficient estimates towards zero. Depending on the type of shrinkage (also called regularisation) performed, some of the coefficients can be estimated as exactly zero. The third approach is called the projection method, which assumes that Y relates to only a few linear combinations of covariates. Thus, it is possible that all covariates have an explanatory effect, but only a few linear combinations represent this effect. First, we briefly introduce in Paragraph 1.3.1 a number of approaches to variables selection. We then describe each of the approaches, namely shrinkage in Paragraph 1.3.2 and projection methods in Paragraph 1.3.3.

1.3.1 Variables selection

With variables selection, we keep only a subset of the variables that exhibit the strongest effects and eliminate the rest in the model. There are a number of different strategies for selecting these variables. Best subset selection is a classical method in statistics which dates from at least (Hocking and Leslie, 1967; Beale et al., 1967). The stepwise approach is a method where variables are selected for inclusion or elimination from the model in a sequential way (Draper and Smith, 1998). There exist many variations of this approach but the two main ones are: forward selection and backward elimination (Hastie et al., 2001, Chapter 3).

1.3.2 Shrinkage methods

The problem of dimension reduction is considered through the following regression model:

$$Y = g(\beta^t X) + \varepsilon, \tag{1.27}$$

where $\varepsilon \in \mathbb{R}$ is the noise, Y is a real random variable, $g : \mathbb{R}^r \to \mathbb{R}$ is a possibly unknown link function, $\beta \in \mathbb{R}^{p\times r}$ are unknown directions. Note that the link function g belongs to a space of infinite dimension, and in this case this approach is referred to as semi-parametric. In the literature, this regression model is called Sufficient Dimension Reduction (Cook and Ni, 2005). This model is guided by a major issue, namely the estimation of the dimension reduction space. Shrinkage methods aim at reducing the dimension of the covariate by selecting variables that are relevant for the model. The main idea of this approach is to add a penalty term to the minimisation of the sum of squared residuals, in order to shrink the small coefficients towards zero while leaving the large coefficients. One can mention Ridge regression (Hoerl and Kennard, 1970) who shrinks the regression coefficients by imposing a L_2 penalty on their size. Considering the regression model (1.27) where g is the linear function and $\beta \in \mathbb{R}^p$ $(r = 1)$, the ridge coefficients minimise the following penalised residual sum of squares:

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}, \tag{1.28}$$

where $\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2$ is the L_2 norm of β and $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage. The larger the value of λ, the greater the amount of shrinkage. The L_2 penalisation has the particularity of not cancelling the β coefficients but rather of reducing them and making them tend towards 0. This is referred to as the "shrinking" of the coefficients. The Lasso method (Tibshirani, 1996) penalises regression coefficients similarly to ridge regression (Hoerl and Kennard, 1970) but replacing the L_2 penalisation by the L_1 counterpart. The lasso estimate is defined by:

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}. \tag{1.29}$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ is the L_1 norm of β. In contrast to the L_2 penalisation, for large values of the L_1 type penalty parameter λ some coefficients of β are set exactly to 0, which allows the selection of more parsimonious models. An extension of the Lasso method, called Elastic net, has been proposed in Zou and Hastie (2005) combining the two penalisation methods presented above. The Elastic Net estimate is defined as follows:

$$\hat{\beta}_{EN} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right\}. \tag{1.30}$$

Other extensions of Lasso method have been proposed and include for instance Adaptative Lasso (Zou, 2006) that assigns different weights to different coefficients of β, Fused lasso (Tibshirani et al., 2005) which penalises both the coefficients and their successive difference and Group Lasso (Yuan and Lin, 2006) who uses an intermediate penalty between the L_1 and L_2 penalisations which selects variables at group level. Many other shrinkage and variable selection methods are discussed in Hastie et al. (2001, Chapter 3).

1.3.3 Projection methods

One of the most popular methods is Partial least squares (PLS) regression, introduced by Wold (1975), that combines characteristics of Principal component analysis (PCA) for dimension reduction and multiple regression. The development of PLS regression has been initiated within the chemometrics field, see the reference book Martens and Naes (1992). Since then, PLS has also received attention in the statistical literature. For example, Helland (1990) discusses the statistical properties of the PLS procedure under a factor analysis model, while Frank and Friedman (1993) provides a comparison between PLS and Principal component regression (PCR) from various perspectives. See also Cook et al. (2013) for a connection between PLS regression and envelopes and Chun and Keleş (2010) for a sparse version of PLS. The basic idea of PLS regression is to seek directions, *i.e.* linear combinations of X, called latent variables, coordinates having both high variance and high correlation with Y, unlike PCR method which only takes into account high variance components (Frank and Friedman, 1993; Stone and Brooks, 1990). Considering the regression model (1.27) where X is centred and $\beta \in \mathbb{R}^{p\times r}$, the PLS consists in finding latent variables $X\beta$ where $\beta = (\beta^{(1)}, \ldots, \beta^{(r)})$ are solutions of the following optimisation problem:

$$\arg \max_{\|\beta^{(\ell)}\|=1} \mathrm{Cov}(X\beta^{(\ell)}, Y), \quad \text{for} \quad \ell = 1, \ldots, r. \tag{1.31}$$

The matrix β is obtained via an iterative approach by calculating the regression of Y on the r constructed centred latent variables.

Sliced inverse regression (SIR) Li (1991) is an alternative method that takes profit of the simplicity of the inverse regression view of X against Y. It aims at replacing X by its projection onto a subspace of smaller dimension without loss of regression information. We consider a generalisation of the model (1.27) in which the noise is not necessarily additive:

$$Y = g(\beta^t X, \varepsilon), \tag{1.32}$$

where X satisfy the so-called linearity condition. We note that the noise is independent of X. Given independent observations $(y_i, x_i)_{1\leq i\leq n}$, SIR first divides the range of the y into H

disjoint slices, denoted as $I_1, \ldots, I_H$, and computes $\overline{x}_h = 1/n_h \sum_{y_i \in I_h} x_i$, where n_h is the number of y_i in I_h and $h = 1, \ldots, H$. Then SIR estimates the variance of the conditional expectation $\mathbb{E}(X|Y)$ by $\hat{M} = 1/n \sum_{h=1}^{H} n_h(\overline{x}_h - \overline{x})(\overline{x}_h - \overline{x})^t$ and $Cov(X)$ by the sample covariance matrix $\hat{\Sigma}$. The conditional expectation $\mathbb{E}(X|Y)$ is called the inverse regression curve which is contained in the dimension reduction subspace. Finally, SIR uses the R largest eigenvectors of $\hat{\Sigma}^{-1}\hat{M}$, denoted by $\hat{\eta}_\ell (\ell = 1, \ldots, R)$, to estimate the SDR directions $\beta^{(\ell)}$. Thus, we have $\hat{\beta}^{(\ell)} = \hat{\eta}_\ell \hat{\Sigma}^{-1/2}$ where $\ell = 1, ..., R$.

Despite the successful use of SIR for dimension reduction in many applications, it has some drawbacks. Indeed, only one direction can be obtained as SIR method only uses $\mathbb{E}(X|Y)$ to extract the information in the slice. This information is not sufficient for non-linear structures. A Generalisation of SIR such as SAVE (Cook and Ni, 2006), among others, proposed to add second moment information on the conditional distribution of X given Y. Various extensions of SIR have been proposed. Some of them were developed to perform dimension reduction and variable selection simultaneously. For instance, Li et al. (2005) proposed a backward subset selection approach based on Cook (2004) and Li (2007) developed the sparse SIR to obtain shrinkage estimators of the SDR directions under L_1 norm. Other extensions of SIR have been proposed such as Partial inverse regression handling the $p > n$ situation (Li et al., 2007), Kernel sliced inverse regression allowing the estimation of a nonlinear subspace (Wu, 2008), Student sliced inverse regression dealing with non-Gaussian and heavy-tailed errors (Chiancone et al., 2017), and Sliced inverse regression for multivariate response (Coudret et al., 2014).

Single-index models provide additional practical tools to overcome the curse of dimensionality, by modelling the non-linear relationship between Y and X through an unknown link function and a single linear combination of the covariates referred to as the index, see Horowitz (2009a, Chapter 2). As such, they provide a reasonable compromise between non-parametric and parametric approaches. Among the numerous works dedicated to the estimation of the index and the link function, the most popular are the average derivative estimation method in the context of kernel smoothing (Härdle and Stoker, 1989; Powell et al., 1989), and the M-estimation technique based on spline regression (Wang and Yang, 2009; Yu and Ruppert, 2002). One can also mention Kong and Xia (2012); Wu et al. (2010) who considered single-index models for the estimation of conditional quantiles.

1.4 Bayesian statistics

In statistical modelling, there are two main viewpoints, namely the frequentist approach and the Bayesian approach. In the frequentist approach, the parameters of a statistical model are

considered to be unknown fixed constants and one tries to estimate them using observed data X. In contrast, the Bayesian approach models all unknown parameters as random variables. The idea behind this approach is that instead of modelling unknown quantities by numbers, it might be interesting to model them by probability distributions. We consider the following identifiable statistical model:

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}, \tag{1.33}$$

where Θ is a parameter space. The way to formalise the Bayesian approach is to assume that the unknown parameter θ of the model $\mathcal{P}$ is random, with a distribution called *a prior distribution*. The latter reflects our prior knowledge and beliefs on the parameter θ. Then, using the observed data $X_1, ..., X_n$, the Bayesian approach consists in updating the prior distribution using the information contained in the data. Thus, we obtain a new distribution, called *a posterior distribution*, which is the update of the prior distribution once the data have been observed.

In Paragraph 1.4.1, we introduce the probabilistic inference of the theory of Bayesian modelling. Then, some sampling techniques for implementing Bayesian inference are given in Paragraph 1.4.2. We refer to the reference book Robert and Casella (2004) to present the basic elements of the Bayesian approach.

1.4.1 Bayesian inference

We consider a random variable X with values in a space E provided with a σ-algebra $\mathcal{E}$, and the statistical model $\mathcal{P}$ defined in (1.33) where $\theta \in \Theta \subset \mathbb{R}^d$ with $d \geq 1$ fixed. We suppose that all P_θ distributions have a density f_θ with respect to a positive σ-finite measure μ on E such that $dP_\theta = f_\theta d\mu$. Also, suppose that prior knowledge about θ can be formulated and expressed by a probability distribution Π, with a density function π with respect to a positive σ-finite measure ν on Θ such that $d\Pi = \pi d\nu$. Thus the distribution of observations of X is conditional on θ and denoted by $f_\theta(x)$.

Definition 1.4.1 (Bayesian model)**.** *A Bayesian model is defined, for a given random variable, by a likelihood and a prior distribution:*

$$\theta \sim \Pi \tag{1.34}$$

$$X|\theta \sim P_\theta. \tag{1.35}$$

From a Bayesian model, we can calculate the posterior distribution on θ, which is defined on Θ and denoted by $\Pi[.|X]$.

Definition 1.4.2 (Posterior distribution)**.** *Using Bayes theorem, the posterior distribution has*

a density with respect to ν given by:

$$f_{\theta|X=x}(\theta) = \frac{f_\theta(x)\pi(\theta)}{\int_\Theta f_\theta(x)\pi(\theta)d\theta}. \tag{1.36}$$

The function $\pi(\theta)f_\theta(x)$ is the joint density of (X, θ) with respect to $\mu \bigotimes \nu$, and the integration of the latter (denominator) is the marginal distribution density of X.
In the case of a statistical sampling experiment where $x = (x_1, \ldots, x_n)$ are realisations of $X = (X_1, \ldots, X_n)$ and $P_\theta = \bigotimes_{i=1}^n P_\theta$, the posterior distribution has a density with respect to ν given by:

$$f_{\theta|X_1=x_1,\ldots,X_n=x_n}(\theta) = \frac{\prod_{i=1}^n f_\theta(x_i)\pi(\theta)}{\int_\Theta \prod_{i=1}^n f_\theta(x_i)\pi(\theta)d\theta}, \tag{1.37}$$

where $\prod_{i=1}^n f_\theta(x_i)\pi(\theta)$ is the joint density of (X, θ) with respect to $\mu^{\bigotimes n} \bigotimes \nu$. The posterior distribution can be rewritten as:

$$f_{\theta|X_1=x_1,\ldots,X_n=x_n}(\theta) \propto \prod_{i=1}^n f_\theta(x_i)\pi(\theta). \tag{1.38}$$

The latter is the product of the likelihood and the prior density.
Note that most of the time μ and ν are taken to be equal to the Lebesgue measure on $\mathbb{R}$, or to a discrete measure such as counting measure for integers.
Several aspects of the posterior distribution could be of interest and are presented in the following definition.

Definition 1.4.3. *Let X be a random variable, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, Π a prior distribution of θ, and $\Pi[.|X]$ the corresponding posterior. We define the following quantities, if they exist:*

- ***Posterior mean:***

$$\overline{\theta}(X) = \int \theta d\Pi[\theta|X]. \tag{1.39}$$

- ***Posterior mode:*** *which is a point $\hat{\theta}_m(X)$ where the maximum of the posterior density is reached. It is given by:*

$$\hat{\theta}_m(X) = \arg\max_{\theta\in\Theta} f_{\theta|X}(\theta). \tag{1.40}$$

- ***Posterior variance:*** *if $\Theta \subset \mathbb{R}^d$,*

$$v(X) = \int (\theta - \overline{\theta}(X))(\theta - \overline{\theta}(X))^t d\Pi[\theta|X]. \tag{1.41}$$

- ***Posterior quantiles:*** *let $\Theta \subset \mathbb{R}^d$ and $F_{\theta|X}(.)$ be the distribution function of the posterior $\Pi[.|X]$, which admits a reciprocal function $F^{-1}_{\theta|X}$. The posterior quantiles are*

defined as, for all $t \in [0,1]$,

$$q_{\theta|X}(t) = F^{-1}_{\theta|X}(t). \tag{1.42}$$

- ***Posterior median:***

$$\hat{\theta}_{med}(X) = q_{\theta|X}(1/2). \tag{1.43}$$

The choice of a prior distribution is a crucial step in Bayesian statistics. The different possible choices can be motivated by different points of view. The first one is the choice based on past experience, expertise of specialists in a certain field or statistician's intuition. For example, if we are in financial agriculture field, we will rely on the knowledge of agricultural experts and farmers to determine a prior distribution. If there are several distinct experts opinions, they can be weighted using a hierarchical model.
The second choice is based on the feasibility of the calculations. One can mention the notion of conjugate distributions.

Definition 1.4.4. *A family* $\mathcal{F}$ *of a prior distributions is said to be conjugate with respect to the model* $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, *if for any* $\Pi \in \mathcal{F}$, *the associated posterior distribution* $\Pi[.|X]$ *also belongs to* $\mathcal{F}$.

The advantage of conjugate distributions is to simplify the calculations. Another interest is that the parameters of the a prior distribution are simply updated using the data and thus the interpretation is much easier.
The last choice is based on uninformative distributions. This is the case where we have little information about θ, and therefore we can not bring new information that could bias the estimation. One can choose a prior distribution said to be uninformative. Intuitively, if one does not want an informative prior, one might think of considering the uniform distribution on Θ. One can also mention Jeffreys prior distribution (Jeffreys, 1946) which is based on Fisher information and is invariant by the change of parametrisation.

1.4.2 Bayesian computational methods

Bayesian statistics often require potentially heavy or unfeasible calculations if simple examples are not used. It is often necessary to use numerical resolution methods which allow to obtain numerical approximations in reasonable time. Let consider the posterior distribution presented in (1.37) and denote by I the renormalisation constant:

$$I = \int_\Theta \prod_{i=1}^{n} f_\theta(x_i)\pi(\theta)d\theta.$$

In practice, the calculation of this integral is potentially problematic, especially when Θ is high dimensional. The first idea is based on Monte-Carlo method which allows to approximate numerically this integral. More generally, we aim to approximate:

$$J = \mathbb{E}[h(\theta)] = \int_{\Theta} h(\theta) f(\theta) d\theta,$$

where $h : \mathbb{R}^d \to \mathbb{R}$ and f is a density of θ (known) and $\int |h| f < \infty$. Assume that $\theta_1, ..., \theta_N$ are i.i.d with the distribution density f. The law of large numbers gives:

$$\hat{J}_N = \frac{1}{N} \sum_{i=1}^{N} h(\theta_i) \to \int_{\Theta} h(\theta) f(\theta) d\theta = J, \tag{1.44}$$

almost surely. Moreover, if $\int h^2 f < \infty$, the central limit theorem gives the asymptotic normality of the estimator. Three sampling methods are presented here: Acceptance-rejection algorithm, Importance sampling, Markov Chain Monte Carlo (MCMC).

1.4.2.1 Acceptance-rejection

The standard acceptance-rejection sampling is an important and commonly used general simulation method. The idea of this method is to generate sampling values from the target density $f_{\theta|X}(\theta)$ by using a proposal density $\rho(\theta)$ (referred to as candidate also) with a support S_ρ. Thus we can sample from $\rho(\theta)$ directly and then "accept" the samples with the probability $f_{\theta|X}(\theta)/K\rho(\theta)$ where $K \geq 1$ is a constant. For this method to be efficient, K must be carefully selected, because the expected number of iterations required to accept a candidate is given by $1/K$. Thus, the rejection method is optimized by setting:

$$K = \sup_{\theta \in S_f} \frac{f_{\theta|X}(\theta)}{\rho(\theta)},$$

where S_f is the support of f such that $S_f \subset S_\rho$. We can therefore propose the following acceptance-rejection algorithm:

Algorithm 1: Acceptance-rejection algorithm

1. Simulate θ according to the candidate distribution $\rho(\theta)$;
2. Simulate u according to the uniform distribution $U[0,1]$ independently;
3. Test:
   if $u \leq \frac{f(\theta|X)}{K\rho(\theta)}$ **then** accept θ ;
   otherwise reject θ and go back to step 1 and iterate until the desired number of realisations is obtained.

1.4.2.2 Importance sampling

Importance sampling is another simulation technique based on so-called importance functions. It is useful when the area of interest may be in a region with a low probability of occurrence. Indeed, if K is much larger than 1, acceptance-rejection method will generate many candidates from ρ and eventually reject most of them. Importance sampling overcomes this problem by sampling from the candidate distribution ρ and then weights these samples again by $f_{\theta|X}(\theta)/\rho(\theta)$. Given a sample $(\theta^1, \ldots, \theta^N)$ from the candidate density $\rho(\theta)$, the importance weights is defined as:

$$q_\ell = \frac{f_{\theta^\ell|X}(\theta^\ell)}{\rho(\theta^\ell)},$$

where $\ell = 1, \ldots, N$. Since the target $f_{\theta^\ell|X}(\theta^\ell)$ is known only up to a normalising constant, we can normalize q_ℓ such as,

$$\widetilde{q}_\ell = \frac{q_\ell}{\sum_{j=1}^N q_j}.$$

Then we reweight the sample $(\theta^1, \ldots, \theta^N)$ by $(\widetilde{q}_1, \ldots, \widetilde{q}_N)$. Finally, we can propose the following importance-sampling algorithm for $\ell \in \{1, \ldots, N\}$:

Algorithm 2: Importance sampling algorithm

1. Simulate θ^ℓ according to the candidate distribution $\rho(\theta)$;
2. For each θ^ℓ compute the importance sampling weight $q_\ell = \frac{f_{\theta^\ell|X}(\theta^\ell)}{\rho(\theta^\ell)}$ and then the normalised sampling weight $\widetilde{q}_\ell = \frac{q_\ell}{\sum_{j=1}^N q_j}$;
3. Reweight each of the candidates θ^ℓ generated from $\rho(\theta)$ by $\widetilde{q}_\ell$.

1.4.2.3 MCMC

The aim of MCMC methods is to approximate a distribution or an integral using a Markov chain with invariant measure. Suppose one wants to either simulate according to a distribution of density f, or to evaluate an integral of the type $J = \int h(t)f(t)dt$. One need only construct a chain (Z_N) of stationary density f. Thus, the distribution of (Z_N), when N is large, will be close to the density distribution f, while the mean $\frac{1}{N}\sum_{i=1}^{N} h(Z_i)$ will approach the integral J. There exist many MCMC techniques of which two are described here: Metropolis Hastings and Gibbs. There is many literature about the theory behind MCMC techniques and on their applications. Introductions are provided by Besag et al. (1995); Gamerman and Lopes (2006); Besag (2001).

Metropolis-Hastings algorithm. The Metropolis-Hastings method comes from work by Metropolis et al. (1953); Hastings (1970). Let $q(.|x)$ be a collection of conditional densities that we know how to simulate. For example, one can consider the distribution $\mathcal{N}(x, v)$ for $q(.|x)$ with v a positive constant. In this case, the algorithm is called Random Walk Metropolis-Hastings. Let $f_{\theta|X}$ be the posterior density that we aim to simulate.

Algorithm 3: Metropolis-Hastings algorithm

Suppose that $Z_1, ..., Z_i$ have been generated. Z_{i+1} is generated as follows:

1. Generate $Y \sim q(.|Z_i)$;
2. Let $r = r(Z_i, Y)$ such that
$$r(z, y) = \min\left(\frac{f_{\theta|X}(y)q(z|y)}{f_{\theta|X}(z)q(y|z)}, 1\right);$$
3. Set
$$Z_{i+1} = \begin{cases} Y & \text{with probability } r \\ Z_i & \text{with probability } 1 - r. \end{cases}$$

By construction (Z_N) is a homogeneous Markov chain, since $\mathcal{L}(Z_{i+1}|Z_i = x)$ depends only on z and not on i.

Gibbs algorithm. The Gibbs sampling was used by (Geman and Geman, 1984) for models with the Gibbs distribution and was extended to the general form (Gelfand and Smith, 1990). This algorithm is particularly useful in hierarchical models. Thus, if one knows how to simulate according to the distributions $\mathcal{L}(\theta|\alpha, Z)$ and $\mathcal{L}(\alpha|\theta, Z)$, the Gibbs algorithm allows to simulate according to an approximation of $\mathcal{L}(\theta, \alpha|Z)$. For the sake of simplification, we

suppose that we want to simulate according to the pair (Z, Y) distribution, in a framework where it is easy to simulate according to the conditional distributions $\mathcal{L}(Z|Y)$ and $\mathcal{L}(Y|Z)$.

Algorithm 4: Gibbs algorithm

Suppose that $(Z_1, Y_1), ..., (Z_i, Y_i)$ have been generated. (Z_{i+1}, Y_{i+1}) are generated as follows:

1. Generate $Z_{i+1} \sim \mathcal{L}(Z|Y = Y_i)$;
2. Generate $Y_{i+1} \sim \mathcal{L}(Y|Z = Z_i)$.

The sequence $(Z_i, Y_i)_{i\geq 1}$ is a homogeneous Markov chain, where $\mathcal{L}((Z, Y))$ is a stationary distribution. Thus for a large N, the distribution of (Z_N, Y_N) will be a good approximation of the distribution of (Z, Y).

1.5 Extreme Value Theory

Most of classical statistical approaches focus on the study of the average behaviour of observed phenomena using some probabilistic tools such as the well-known central limit theorem. However, these approaches do not address the behaviour of rare or extreme events found in the tails of probability distributions. Hence, the extreme value theory comes to study these events in order to understand and characterise their behaviour. This theory consists in solving one of the following two problems: (i) first assess the probability of observing the occurrence of an event whose amplitude is greater than a given sample value, (ii) second determine the amplitude of the event that is exceeded with a low probability. The first problem concerns calculating a low probability associated with an extreme event. The second problem is about quantifying the value of an extreme event (called quantile), i.e. an event whose probability of occurrence is low by definition. The main results of the extreme value theory are based on the theorem of Fisher and Tippett (1928) and Gnedenko (1943) on the convergence in distribution of the maximum value of a sequence of independent and identically distributed random variables, then on the result of Pickands (1975) on the convergence in distribution of excesses above a threshold.

We start by studying the asymptotic behaviour of extreme values of a sample, in Paragraph 1.5.1, by first looking at the law of the maximum of a sample then by considering the excesses above a given threshold. Then, we give the characterisations of the domains of attraction, in Paragraph 1.5.2, as well as the definition and some properties of functions with regular variations. In paragraph 1.5.3, we recall the different methods of estimation of extreme quantiles. Finally, we consider the presence of a covariate and recall different

approaches for conditional extreme quantile estimation, in last Paragraph 1.5.4. We also review the literature on dimension reduction and Bayesian statistics in the extreme regression context in this last paragraph.

1.5.1 Asymptotic behaviour of extreme values of a distribution

We consider n real random variables $Y_1, \ldots, Y_n$ independent and identically distributed (i.i.d) with a distribution function $F(y) = \mathbb{P}(Y \leq y)$ and a survival function $\bar{F}(y) = 1 - F(y)$. Let consider $Y_{1,n} \leq \cdots \leq Y_{n,n}$ the order statistics of the sample. Two order statistics are particularly interesting for the study of extreme events: the minimum and the maximum denoted respectively by $Y_{1,n} = \min(Y_1, \ldots, Y_n)$ and $Y_{n,n} = \max(Y_1, \ldots, Y_n)$. In the following, we focus on the study of the maximum, since the results for the minimum can be deduced directly from the results for the maximum by considering the opposite series $-Y_1, \ldots, -Y_n$, according to the following equality:

$$Y_{1,n} = -\max(-Y_1, \ldots, -Y_n).$$

The distribution function of the maximum $Y_{n,n}$ is given by:

$$\begin{aligned} F_{Y_{n,n}}(y) &:= \mathbb{P}(Y_{n,n} \leq y) \\ &= \mathbb{P}(Y_1 \leq y, ..., Y_n \leq y) \\ &= \prod_{i=1}^{n} \mathbb{P}(Y_i \leq y) \\ &= F^n(y). \end{aligned} \tag{1.45}$$

This result is not useful in practice, since the distribution function F is unknown. However, from (1.45), we can conclude about the form of the limit distribution of $Y_{n,n}$ when $n \to \infty$:

$$\lim_{n\to\infty} F_{Y_{n,n}}(y) = \lim_{n\to\infty} F^n(y) = \mathbb{1}_{\{y \geq y_F\}} = \begin{cases} 1 & \text{if} \quad y \geq y_F \\ 0 & \text{if} \quad y < y_F \end{cases} \tag{1.46}$$

where $y_F = \sup\{y \in \mathbb{R}, F(y) < 1\}$ is the endpoint of F. The result (1.46) shows that the distribution of the maximum $Y_{n,n}$ is degenerated since it reduces to a Dirac mass at y_F. This result provides limited information on the behaviour of $Y_{n,n}$. Therefore, to find a non-degenerated limit distribution, we need to transform $Y_{n,n}$. The simplest transformation that one can imagine is a normalisation operation. The variable $Y_{n,n}$ is adjusted with a scale parameter a_n, assumed positive, and a location parameter b_n. In the next pragraph, we state the theorem that gives the form of the limit distibution of $(Y_{n,n} - b_n)/a_n$.

1.5.1.1 Asymptotic behaviour of maximum (GEV)

The works of Fisher and Tippett (1928) and Gnedenko (1943) give the following theorem which is fundamental in extreme value theory as it establishes the asymptotic distribution of the normalised maximum.

Theorem 1.5.1 (Fisher–Tippett–Gnedenko)**.** *Let $(Y_n)_{n\geq 1}$ be a sequence of i.i.d random variables with a distribution function F. Suppose there exist sequences $(a_n)_{n\geq 1} > 0$, $(b_n)_{n\geq 1} \in \mathbb{R}$ and a non-degenerate distribution G_γ such that $\forall y \in \mathbb{R}$:*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{Y_{n,n} - b_n}{a_n} \leq y\right) = \lim_{n\to\infty} F^n(a_n y + b_n) = G_\gamma(y),$$

Then, up to location and scale parameters,

$$G_\gamma(y) = \begin{cases} \exp\left(-(1+\gamma y)_+^{-1/\gamma}\right) & \text{if } \gamma \neq 0 \\ \exp\left(-\exp\left(-y\right)\right) & \text{if } \gamma = 0, \end{cases} \tag{1.47}$$

where $\gamma \in \mathbb{R}$ and $z_+ = \max(0, z)$.

The above theorem is an equivalent of the central limit theorem for the maximum of n i.i.d. random variables. It states that the asymptotic behaviour of the renormalized maximum $Y_{n,n}$ is governed by a single distribution G_γ called the Generalized Extreme Value distribution (GEV). The GEV distribution is based on a shape parameter γ, called the extreme-value index or the tail-index, which characterises the behaviour of the tail of the distribution of F. We distinguish three different forms for the law of G_γ, according to the sign of γ, which are called domains of attraction:

- if $\gamma > 0$, F is said to belong to the **Fréchet** domain of attraction (Fréchet, 1927), denoted by $\mathcal{D}$(Fréchet). This domain of attraction includes distributions with heavy tails such as Pareto, Student, Burr, Fréchet, whose survival function decreases as a power function.
- if $\gamma = 0$, F is said to belong to the **Gumbel** domain of attraction (Gumbel, 1958), denoted by $\mathcal{D}$(Gumbel). This domain of attraction contains distributions with light tails such as Gaussian, Exponential, Gamma, Weibull, whose survival function decreases exponentially fast.
- if $\gamma < 0$, F is said to belong to the **Weibull** domain of attraction (Weibull, 1951), denoted by $\mathcal{D}$(Weibull). This domain of attraction includes distributions with short tails such as Beta, ReverseBurr, Uniform, which have a finite endpoint.

Several examples of distributions per domain of attraction are available in Embrechts et al. (2013) page 145.

Considering only the sample maximum to model the behaviour of the extreme values leads to a loss of the information contained in the other large values of the sample. The method of excesses over a high threshold is an alternative to the GEV approach based on the large values of the sample.

1.5.1.2 Asymptotic behaviour of excesses over threshold (POT)

The excesses over a high threshold approach, called also Peaks-Over-Threshold (POT), consists in keeping only the observations that exceed a certain threshold. We assume a sample of random variables i.i.d $Y_1, \ldots, Y_n$ and define a fixed threshold $u (u < y_F)$. Let us consider N_u observations $Y_{i_1}, \ldots, Y_{i_{N_u}}$ exceeding the threshold u and define the excesses over the threshold u by $Z_j := Y_{i_j} - u$ with $j = 1, \ldots, N_u$. We denote by F_u the distribution function of the excess Z above the threshold u defined as follows, for $y \geq 0$:

$$F_u(y) \quad = \quad \mathbb{P}(Z \leq y | Y > u) = \mathbb{P}(Y - u \leq y | Y > u) = \frac{F(u+y) - F(u)}{1 - F(u)},$$

or equivalently for the survival function:

$$\overline{F}_u(y) = \mathbb{P}(Z > y | Y > u) = 1 - F_u(y) = \frac{\overline{F}(u+y)}{\overline{F}(u)}. \tag{1.48}$$

The works of Balkema and de Haan (1974) and Pickands (1975) give the following theorem that establishes the existence of a limit distribution for excesses when the threshold u is close to the endpoint y_F.

Theorem 1.5.2 (Pickands–Balkema–de Haan)**.** *The distribution function F belongs to the maximum domain of attraction G_γ if and only if there exist $\sigma > 0$ and $\gamma \in \mathbb{R}$ such that:*

$$\lim_{u \to y_F} \sup_{y \in (0, y_F - u)} |F_u(y) - H_{\gamma,\sigma}(y)| = 0,$$

where $H_{\gamma,\sigma}$ is the Generalized Pareto Distribution (GPD) defined as follows:

$$H_{\gamma,\sigma}(y) = \begin{cases} 1 - \left(1 + \gamma \frac{y}{\sigma}\right)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma}\right) & \text{if } \gamma = 0, \end{cases} \tag{1.49}$$

with $y \in \mathbb{R}^+$ and $0 \leq y \leq -\frac{\sigma}{\gamma}$ if $\gamma < 0$.

The above Theorem presents an equivalent to Theorem 1.5.1 and establishes the convergence in distribution of excesses over a high threshold to a GPD $H_{\gamma,\sigma}(y)$.

Remark 1.5.1.

1. *When $\gamma = 0$, $H_{\gamma,\sigma}(y)$ corresponds to the Exponential distribution with parameter $1/\sigma$.*
2. *When $\gamma = -1$, $H_{\gamma,\sigma}(y)$ corresponds to the Uniform distribution on $[0, \sigma]$.*
3. *The tail-index γ is identical for both distributions GEV and GPD.*

In the following paragraph, we give the conditions for F to belong to one of the three previously defined domains of attraction.

1.5.2 Characterisation of domains of attraction

First, we recall the notions of slowly varying functions and regularly varying functions which will be useful for the characterisation of the three domains of attraction. Several detailed results on functions with regular variations are given in Bingham et al. (1987).

1.5.2.1 Slowly varying functions

Definition 1.5.1. *A Lebesgue measurable function $\ell > 0$ is slowly varying at infinity, denoted by $\ell \in RV_0$, if $\forall \lambda > 0$:*

$$\frac{\ell(\lambda y)}{\ell(y)} \to 1 \quad as \quad y \to \infty.$$

We can characterise slowly varying functions more precisely using the following representation of Karamata (Bingham et al., 1987)[Theorem 1.3.1].

Theorem 1.5.3 (Karamata's representation)**.** *The function ℓ is slowly varying if and only if it may be written in the form*

$$\ell(y) = c(y) \exp\left\{\int_a^y \frac{\varepsilon(u)}{u} du\right\}, \forall y \geq a > 0,$$

where $\varepsilon(.)$, $c(.)$ are measurable and $c(y) \to c_0 \in (0, +\infty)$, $\varepsilon(y) \to 0$ as $y \to \infty$.

If the function c is constant, the function ℓ is said to be normalized and it is differentiable with derivative ℓ' for all $y > 0$,

$$\ell'(y) = \frac{\varepsilon(y)\ell(y)}{y}.$$

In particular, we have $y\ell'(y)/\ell(y) = \varepsilon(y) \to 0$ as $y \to \infty$.
Using the representation Theorem 1.5.3, one can mention some specific examples of functions with slow variations at infinity: for instance, the logarithm function $\ell(y) = \log y$, the iterates $\ell(y) = \log\log y$, or functions of the form $\ell(y) = \exp(\log(y)^d)$ with $0 < d < 1$.

Finally, we recall some elementary properties of slowly varying functions, see Bingham et al. (1987)[Proposition 1.3.6].

Proposition 1.5.1.

1. *If* $\ell \in RV_0$, *then* $\log \ell(y)/\log(y) \to 0$ *as* $y \to \infty$.
2. *If* $\ell \in RV_0$, *then* $\forall \alpha \in \mathbb{R}$, $(\ell(y))^\alpha \in RV_0$.
3. *If* $\ell \in RV_0$ *and* $\alpha > 0$, *then*

$$y^\alpha \ell(y) \to \infty \quad \text{and} \quad y^{-\alpha}\ell(y) \to 0 \quad \text{as} \quad y \to \infty.$$

4. *If* ℓ_1 *and* ℓ_2 *vary slowly, then* $\ell_1 + \ell_2 \in RV_0$ *and* $\ell_1\ell_2 \in RV_0$. *Moreover, if* $\ell_2(y) \to \infty$ *as* $y \to \infty$, *then* $\ell_1 \circ \ell_2 \in RV_0$.

1.5.2.2 Regularly varying functions

Regularly varying functions characterise functions that behave like a power function at infinity. Let us recall the definition of a regularly varying function (Bingham et al., 1987)[Page 18].

Definition 1.5.2. *A Lebesgue measurable function* $f > 0$ *is called regularly varying with index* $\rho \in \mathbb{R}$ *at infinity, denoted by* $f \in RV_\rho$, *if* $\forall \lambda > 0$*:*

$$\frac{f(\lambda y)}{f(y)} \to \lambda^\rho \quad \text{as} \quad y \to \infty. \tag{1.50}$$

We note that if $\rho = 0$, then $f \in RV_0$, i.e f is a slowly varying function at infinity (see Paragraph 1.5.2.1). Besides, any regularly varying function of index $\rho \in \mathbb{R}$ can be written as $f(y) = y^\rho \ell(y)$, with $\ell \in RV_0$. Some examples of regularly varying functions with index ρ are: $f(y) = y^\rho$ and $f(y) = y^\rho(\log(y))^\gamma$, $f(y) = y^\rho(\log\log(y))^\gamma$, for $y > 1$, $\rho \in \mathbb{R}$ and $\gamma \in \mathbb{R}$. An important result is given in the following theorem, stating that convergence in (1.50) is uniform on each interval of $(0, \infty)$ (see de Haan and Ferreira (2007)[Theorem B.1.4]).

Theorem 1.5.4. *If* $f \in RV_\rho$, *then for all* $0 < a < b < \infty$*:*

$$\frac{f(\lambda y)}{f(y)} \to \lambda^\rho \quad \text{as} \quad y \to \infty\ , \quad \text{uniformly in } \lambda \in [a, b]. \tag{1.51}$$

Some properties of regularly varying functions are presented in the following proposition (we refer to de Haan and Ferreira (2007)[Proposition B.1.9]).

Proposition 1.5.2.

1. *If $f \in RV_\rho$, then $\log f(y)/\log(y) \to \rho$ as $y \to \infty$. This implies*

$$\lim_{y\to+\infty} f(y) = \begin{cases} 0 & \text{if} \quad \rho < 0 \\ +\infty & \text{if} \quad \rho > 0. \end{cases}$$

2. *If $f \in RV_\rho$ and $\alpha \in \mathbb{R}$, then $(f(y))^\alpha \in RV_{\alpha\rho}$.*

3. *If $f_1 \in RV_{\rho_1}$ and $f_2 \in RV_{\rho_2}$, then $f_1 + f_2 \in RV_{\max(\rho_1,\rho_2)}$. Moreover, if $f_2(y) \to \infty$ as $y \to \infty$, then $f_1 \circ f_2 \in RV_{\rho_1\rho_2}$.*

4. *Suppose $F(y) = \int_0^y f(t)dt$. If $F \in RV_\rho$ with $\rho \neq 0$ and f is monotone at infinity, then $f \in RV_{\rho-1}$.*

5. *If $f \in RV_\rho$, with $\rho \geq -1$, is integrable on finite intervals of $\mathbb{R}^+$, then $\int_0^y f(t)dt \in RV_{\rho+1}$. Besides, if $f \in RV_\rho$, with $\rho < -1$, then $\int_y^\infty f(t)dt \in RV_{\rho+1}$.*

6. *(Potter Bounds) Suppose $f \in RV_\rho$ and $\delta_1 > 0$, $\delta_2 > 0$. Then, there exists $t_0 > 0$ such that for $t \geq t_0$, $ty \geq t_0$,*

$$(1-\delta_1)y^\rho \min(y^{\delta_2}, y^{-\delta_2}) < \frac{f(ty)}{f(t)} < (1+\delta_1)y^\rho \max(y^{\delta_2}, y^{-\delta_2}).$$

Statements 4 and 5 of Proposition 1.5.2 characterise the derivative and the primitive of a regularly varying function. They indicate that the derivative or primitive of a regularly varying function of index ρ is usually also a regularly varying function, but of index $\rho - 1$ or $\rho + 1$. The following proposition concerns the limit of $yf'(y)/f(y)$ with $f \in RV_\rho$.

Proposition 1.5.3. *Let $F(y) = \int_0^y f(t)dt$, with derivative $f(y)$. If $F \in RV_\rho$, $\rho \in \mathbb{R}$, and f is monotone at infinity, then $yf(y)/F(y) \to \rho$ as $y \to \infty$.*

Potter bounds in the statement 6 of Proposition 1.5.2 can be used to prove that regularly varying functions preserve equivalences:

Proposition 1.5.4. *If $f \in RV_\rho$, with $\rho \in \mathbb{R}$, $u_n \to \infty$ and $v_n \sim u_n$ as $n \to \infty$, then $f(v_n) \sim f(u_n)$ as $n \to \infty$.*

Let us recall that the generalised inverse of an increasing function f is defined as:

$$f^{\leftarrow}(y) = \inf\{t, f(t) \geq y\}, \tag{1.52}$$

and it coincides with the classical inverse f^{-1} when f is continuous and strictly increasing.

Proposition 1.5.5.

1. *If* $f \in RV_\rho$ *with* $\rho > 0$, *then* $f^{\leftarrow}(\cdot) \in RV_{1/\rho}$.
2. *If* $f \in RV_\rho$ *with* $\rho < 0$, *then* $(1/f)^{\leftarrow}(\cdot) \in RV_{-1/\rho}$.

1.5.2.3 Fréchet domain of attraction

We recall that the Fréchet domain of attraction contains heavy-tailed distributions, i.e. distributions whose survival function decreases as a power function.

Theorem 1.5.5. *The distribution function* F *is in the Fréchet domain of attraction,* $F \in \mathcal{D}$*(Fréchet), with index* $\gamma > 0$, *if and only if the endpoint* $y_F = \infty$ *and* $\overline{F} \in RV_{-1/\gamma}$, *i.e* $\forall y > 0$*:*

$$\frac{\overline{F}(ty)}{\overline{F}(t)} \to y^{-1/\gamma}, \quad as \quad t \to \infty.$$

Using Proposition 1.5.5, $F \in \mathcal{D}$(Fréchet) can be rewritten as:

$$F(y) = 1 - y^{-1/\gamma}\ell(y) \quad \text{with} \quad \ell \in RV_0. \tag{1.53}$$

The normalisation sequences a_n and b_n are given by:

$$a_n = F^{\leftarrow}(1 - 1/n) \quad , \quad b_n = 0.$$

Note that all distribution functions belonging to the Fréchet domain of attraction have an infinite endpoint. Besides, a characterisation of the quantile function in the Fréchet attraction domain can be obtained by: $q(\alpha) := F^{\leftarrow}(1-\alpha)$ with $\alpha \in (0,1)$. Using Proposition 1.5.5, we can show that expression (1.53) is equivalent to:

$$q(\alpha) = \alpha^{-\gamma}\ell(\alpha^{-1}), \quad \text{with} \quad \ell \in RV_0. \tag{1.54}$$

Finally, the Fréchet domain of attraction only contains heavy-tailed distributions such as Pareto, Student, Burr, Fréchet, Log-gamma, Cauchy, etc. It is used in several applications including meteorology (Gardes and Girard, 2010; El Methni et al., 2014), hydrology (Anderson and Meerschaert, 1998; El Methni et al., 2012) and insurance (Matthys et al., 2004; Ahmad et al., 2019).

1.5.2.4 Weibull domain of attraction

It is recalled that distributions belonging to the Weibull domain of attraction have a short tail with finite endpoint y_F.

Theorem 1.5.6. *The distribution function F is in the Weibull domain of attraction, $F \in \mathcal{D}(\text{Weibull})$, with extreme-value index $\gamma < 0$, if and only if the endpoint $y_F < \infty$ and the distribution function F_* defined by:*

$$F_*(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ F(y_F - 1/y) & \text{if } y > 0, \end{cases} \tag{1.55}$$

belongs to the Fréchet domain of attraction with index $-\gamma > 0$, i.e

$$\forall y > 0, \quad \lim_{t\to\infty} \frac{\overline{F}\left(y_F - 1/(ty)\right)}{\overline{F}\left(y_F - 1/t\right)} = \lim_{t\to\infty} \frac{\overline{F}_*(ty)}{\overline{F}_*(t)} = y^{1/\gamma}.$$

Thus, a distribution function $F \in \mathcal{D}(\text{Weibull})$ can be written as, for $y \leq y_F$:

$$F(y) = 1 - (y_F - y)^{-1/\gamma} \ell((y_F - y)^{-1}), \tag{1.56}$$

where $\ell \in RV_0$. Besides, a possible choice of normalizing sequences (a_n) and (b_n) is:

$$a_n = y_F - F^{\leftarrow}(1 - 1/n) \quad , \quad b_n = y_F.$$

In addition, a characterisation of the associated quantile function is given by:

$$q(\alpha) = y_F - \alpha^{-\gamma} \ell(1/\alpha), \quad \text{with} \quad \ell \in RV_0 \quad \text{and} \quad \alpha \in (0, 1). \tag{1.57}$$

Finally, the Weibull domain of attraction includes short tails distributions as for example Uniform, Beta, ReverseBurr. It is used in many applications such as the estimation of the maximal life span of humans (Aarssen and de Haan, 1994) or in hydrology (Durrans, 1996). This domain of attraction has also been considered by Falk (1995); Hall and Park (2002); Girard et al. (2012) for the estimation of the endpoint y_F.

1.5.2.5 Gumbel domain of attraction

The Gumbel domain of attraction contains light-tailed distributions, i.e distributions whose survival functions decrease exponentially fast. This domain covers a wide range of distributions which are difficult to characterise in a simple way.

Theorem 1.5.7. *The distribution function F is in the Gumbel domain of attraction, $F \in \mathcal{D}(\text{Gumbel})$, with index $\gamma = 0$, if and only if there exists $t < y_F \leq \infty$ such that:*

$$\overline{F}(y) = c(y) \exp\left\{ - \int_t^{y_F} \frac{g(t)}{a(t)} dt \right\}, \quad t < y < y_F, \tag{1.58}$$

where c, g are measurable and a, c and g are three functions verifying $a'(y) \to 0$, $c(y) \to c > 0$ and $g(y) \to 1$ as $y \to y_F$.

A possible choice of the normalizing constants is:

$$a_n = q(1/n) = F^{\leftarrow}(1 - 1/n) \quad , \quad b_n = \frac{1}{\overline{F}(a_n)} \int_{a_n}^{y_F} \overline{F}(t)dt.$$

It is difficult to characterise the quantile function of distributions belonging to the Gumbel attraction domain. The result of Theorem 1.5.7 gives a precise characterisation of $\overline{F}$ (see (Resnick, 1987)[Proposition 1.4]), but it is difficult to implement for quantiles which reflects the complexity of this domain. However, we can find in the literature several characterisations proposed for subfamilies of distributions derived from the latter such as the family of distributions with a Weibull tail (Galambos, 1977; Girard, 2004; Gardes et al., 2011; Gardes and Girard, 2013). This family includes several usual distributions, including the Gaussian, Exponential, Gamma and Weibull.

Definition 1.5.3. *The distribution function F is said to have a Weibull tail if there exists $\beta > 0$ such that $\forall y > 0$,*

$$\frac{-\log \overline{F}(ty)}{-\log \overline{F}(t)} \to y^{1/\beta} \text{ as } t \to \infty.$$

The shape parameter $\beta > 0$, called Weibull tail-index, controls the decay of the tail. Thus, the survival function of a Weibull tail distribution can be given by:

$$\overline{F}(y) = \exp\left(-y^{1/\beta}\ell(y)\right) \quad \text{with} \quad \ell \in RV_0. \tag{1.59}$$

A characterization of the associated quantile function can also be obtained:

$$q(\alpha) = (-\log(\alpha))^{\beta}\ell(-\log(\alpha)) \quad \text{with} \quad \ell \in RV_0 \quad \text{and} \quad \alpha \in (0,1). \tag{1.60}$$

This family of distributions is used in many applications such as insurance (Beirlant and Teugels, 1992), hydrology (Gumbel, 1941; Diebolt et al., 2008), Bayesian neural network models (Vladimirova et al., 2020).

1.5.2.6 General characterisation of maximum domains of attraction

A common characterisation of the three previous domains of attraction is given in the following theorem (see (de Haan and Ferreira, 2007)[Theorem 1.1.6]).

Theorem 1.5.8. *For $\gamma \in \mathbb{R}$ the following statements are equivalent:*

1. *There exist* $(a_n)_{n\geq 1} > 0$ *and* $(b_n)_{n\geq 1} \in \mathbb{R}$ *such that for* $y > 0$ *with* $1 + \gamma y > 0$*:*

$$F^n(a_n y + b_n) \to G_\gamma(y) = \exp\left(-(1+\gamma y)^{-1/\gamma}\right), \text{ as } n \to \infty.$$

2. *There exists a positive function* a *such that for* $y > 0$*:*

$$\lim_{t\to\infty} \frac{q(1/(ty)) - q(1/t)}{a(t)} = \begin{cases} \frac{y^\gamma - 1}{\gamma} & \text{if } \gamma \neq 0 \\ \log y & \text{if } \gamma = 0. \end{cases} \tag{1.61}$$

3. *There exists a positive function* f *such that for* $y > 0$ *with* $1 + \gamma y > 0$*:*

$$\lim_{t\uparrow y_F} \frac{\overline{F}(t + y f(t))}{\overline{F}(t)} = \begin{cases} (1+\gamma y)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ e^{-y} & \text{if } \gamma = 0. \end{cases} \tag{1.62}$$

Once we have seen how to model the behaviour of the extreme values of a sample through the GEV and GPD approaches, now we focus on the estimation of extreme quantiles.

1.5.3 Estimation of extreme quantiles

We begin by defining the notion of quantiles and extending it to extreme quantiles before discussing their estimation.

Definition 1.5.4. *The quantile* q *of order* α *associated with the distribution function* F *is defined by:*

$$q(\alpha) := F^{\leftarrow}(1-\alpha) = \inf\{y,\ F(y) \geq 1-\alpha\} \quad \text{with} \quad \alpha \in (0,1), \tag{1.63}$$

where $F^{\leftarrow}$ *represents the generalized inverse of* F*.*

We refer to an extreme quantile if we replace its order α by a sequence $\alpha_n \to 0$ when $n \to \infty$.

Definition 1.5.5. *The extreme quantile* q *of order* α_n *associated with the distribution function* F *is defined by:*

$$q(\alpha_n) := F^{\leftarrow}(1-\alpha_n) \quad \text{with} \quad \alpha_n \to 0 \quad \text{when} \quad n \to \infty. \tag{1.64}$$

In the context of insurance and finance, an extreme quantile can be interpreted as the "Value-at-Risk" of an extreme loss (McNeil et al., 2005), or as the return level in hydrology associated with a climatic event (Jagger and Elsner, 2006; Coles et al., 2003).
Statistically speaking, problem (ii), posed at the very beginning of the Paragraph 1.5, relates

to the estimation of a small probability, or equivalently in hydrology to a return period[10]. As for problem (i), it refers to the estimation of an extreme quantile or return level[11] in hydrology. Now, one can ask what is the probability that the extreme quantile is greater than the sample maximum?

Since the random variables are i.i.d and F is continuous, then when $\alpha_n \to 0$, this probability is written as:

$$\begin{aligned}
\mathbb{P}(Y_{n,n} < q(\alpha_n)) &= \mathbb{P}\left(\bigcap_{i=1}^{n}\{Y_i \leq q(\alpha_n)\}\right) \\
&= \prod_{i=1}^{n}\mathbb{P}(Y_i \leq q(\alpha_n)) \\
&= F^n(q(\alpha_n)) \\
&= (1-\alpha_n)^n \\
&= \exp(n\log(1-\alpha_n)) \\
&= \exp(-n\alpha_n(1+o(1)).
\end{aligned}$$

Thus, we see that the probability that the extreme quantile is greater than the sample maximum depends on the asymptotic behaviour of $n\alpha_n$. Therefore, one must distinguish between two cases depending on the speed of convergence of $\alpha_n \to 0$.

In the first case, when $n\alpha_n \to \infty$, then $\mathbb{P}(Y_{n,n} < q(\alpha_n)) \to 0$. The quantile to be estimated is with high probability within the sample. We are in the case where α_n converges slowly to 0. The estimation of the extreme quantile can be achieved thanks to an interpolation within the sample. The $\lfloor n\alpha_n \rfloor$th largest observation can be the estimator of this quantile, i.e $\hat{q}(\alpha_n) = Y_{n-\lfloor n\alpha_n \rfloor,n}$. This estimator is obtained by inverting the empirical distribution function:

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{Y_i \leq y\}},$$

and this latter is asymptotically Gaussian (see de Haan and Ferreira (2007)[Theorem 2.2.1]).

In the second case, when $n\alpha_n \to 0$, then $\mathbb{P}(X_{n,n} < q(\alpha_n)) \to 1$. The quantile to be estimated is with high probability outside the sample. The inversion of the empirical distribution function can not be used to estimate $q(\alpha_n)$ since $\hat{F}_n(y) = 1$ for $y \geq Y_{n,n}$. In the case where α_n converges quickly to 0, this requires extrapolating out of the sample to estimate $q(\alpha_n)$.

In extreme value theory, there exist three different approaches to estimate the extreme quantile: GEV, GPD and semi-parametric methods.

[10] The return period function is given by: $T(y) = 1/\overline{F}(y)$.

[11] The return level function can be defined as the inverse of the return period: $y(T) = F^{\leftarrow}(1-1/T)$. It represents the level that is exceeded with return period T.

1.5.3.1 GEV approach

To estimate the quantile $q(\alpha_n)$ with GEV approach, we use Theorem 1.5.1. One has the following approximation

$$\mathbb{P}\left(Y_{n,n} \leq a_n y + b_n\right) = F^n(a_n y + b_n) \simeq G_\gamma(y), \tag{1.65}$$

which can be rewritten as:

$$n \log\left(1 - \overline{F}(a_n y + b_n)\right) \simeq \log G_\gamma(y).$$

when $n \to \infty$. Since $\overline{F}(a_n y + b_n) \to 0$, as $a_n y + b_n \to y_F$ when $n \to \infty$, one can use the Taylor expansion of $\log(1+y)$ to first order and a change of variable. Thus, we have

$$\overline{F}(y) \simeq -\frac{1}{n} \log G_\gamma\left(\frac{y - b_n}{a_n}\right).$$

Using the expression for G_γ given in (1.47), we obtain

$$\overline{F}(y) \simeq \begin{cases} \dfrac{1}{n}\left(1 + \gamma \dfrac{y - b_n}{a_n}\right)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ \dfrac{1}{n} \exp\left(-\dfrac{y - b_n}{a_n}\right) & \text{if } \gamma = 0, \end{cases} \tag{1.66}$$

an approximation of the distribution tail of F. Then, we can approximate the quantile $q(\alpha_n)$ by inverting the equation (1.66) and we get:

$$q(\alpha_n) \simeq \begin{cases} b_n + \frac{a_n}{\gamma}\left(\left(\frac{1}{n\alpha_n}\right)^\gamma - 1\right) & \text{if } \gamma \neq 0 \\ b_n - a_n \log(n\alpha_n) & \text{if } \gamma = 0. \end{cases} \tag{1.67}$$

From the above approximation, we can obtain an estimator of the extreme quantile, defined as follows:

$$\hat{q}_{GEV}(\alpha_n) = \begin{cases} \hat{b}_n + \frac{\hat{a}_n}{\hat{\gamma}_n}\left(\left(\frac{1}{n\alpha_n}\right)^{\hat{\gamma}_n} - 1\right) & \text{if } \gamma \neq 0 \\ \hat{b}_n - \hat{a}_n \log(n\alpha_n) & \text{if } \gamma = 0, \end{cases} \tag{1.68}$$

where $\hat{a}_n$, $\hat{b}_n$ and $\hat{\gamma}_n$ are respectively estimators of the unknown parameters a_n, b_n and γ. In order to estimate these parameters, Gumbel (1958) proposes the block maxima approach. The idea of the latter is to divide the sample into m disjoint blocks of approximately equal size. For each of these blocks, the largest observation is considered. Thus, we obtain a sample of block maxima whose siqtribution can be approximated by a GEV distribution (see equation (1.47)). Hence, the parameters a_n, b_n and γ can be estimated by several methods proposed in the literature. One can mention two well-known methods, the first is maximum

likelihood (Prescott and Walden, 1980, 1983), and the second is weighted moments (Hosking et al., 1985). The method of weighted moments is more popular because it gives better results than the method of maximum likelihood on small samples. In addition, weighted moment estimators have an analytical expression, simple to calculate.

1.5.3.2 GPD approach

The excesses over threshold approach consists of approximating the distribution of excesses by a GPD distribution (see Theorem 1.5.2). Using equation (1.48), for all $z \geq 0$, we have

$$\overline{F}(z + u_n) = \overline{F}(u_n)\overline{F}_{u_n}(z). \tag{1.69}$$

We perform a change of variable $y = z + u_n$ and we get:

$$\overline{F}(y) = \overline{F}(u_n)\overline{F}_{u_n}(y - u_n). \tag{1.70}$$

Using Theorem 1.5.2, for a large threshold u_n such that $\overline{F}(u_n) = \mathbb{P}(Y > u_n) = p_n$, where p_n is the probability that Y exceeds u_n, we obtain an approximation of the tail survival function:

$$\overline{F}(y) \simeq \begin{cases} p_n \left(1 + \gamma \dfrac{y - u_n}{\sigma_n}\right)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ p_n \exp\left(-\dfrac{y - u_n}{\sigma_n}\right) & \text{if } \gamma = 0. \end{cases} \tag{1.71}$$

Then, we approximate the quantile $q(\alpha_n)$ by inverting equation (1.71) and we get:

$$q(\alpha_n) \simeq \begin{cases} u_n + \frac{\sigma_n}{\gamma}\left[\left(\frac{p_n}{\alpha_n}\right)^{\gamma} - 1\right] & \text{if } \gamma \neq 0 \\ u_n + \sigma_n \log\left(\dfrac{p_n}{\alpha_n}\right) & \text{if } \gamma = 0. \end{cases} \tag{1.72}$$

Thus, we obtain an estimator for the extreme quantile $q(\alpha_n)$ based on the GPD approach and defined by:

$$\hat{q}_{GPD}(\alpha_n) = \hat{u}_n + \frac{\hat{\sigma}_n}{\hat{\gamma}_n}\left(\left(\frac{p_n}{\alpha_n}\right)^{\hat{\gamma}_n} - 1\right), \tag{1.73}$$

where $\hat{u}_n$, $\hat{\sigma}_n$ and $\hat{\gamma}_n$ are respectively estimators of the unknown parameters u_n, σ_n and γ_n. The threshold u_n is a quantile which is within the sample and can be estimated by inversion of the empirical survival function. In particular, if we set $p_n = k_n/n$, where k_n is the number of excesses, then $u_n = F^{\leftarrow}(1 - p_n)$ can be estimated by $\hat{u}_n = Y_{n-k_n+1,n}$. The parameters $\hat{\sigma}_n$ and $\hat{\gamma}_n$ can be estimated by the maximum likelihood (Smith, 1987; Davison and Smith, 1990) and the method of moments or weighted moments (Hosking and Wallis, 1987).

1.5.3.3 Semi-parametric approach

The semi-parametric approach is based on the characterisation of the considered domain of attraction in order to propose estimators of the extreme quantiles. If we consider the Fréchet domain of attraction, $F \in \mathcal{D}(\text{Fréchet})$ (see Paragraph 1.5.2), the quantile function is given by (see equation (1.54)), for some $\gamma > 0$:

$$q(\alpha_n) = \alpha_n^{-\gamma}\ell(\alpha_n^{-1}), \quad \text{with} \quad \ell \in RV_0, \tag{1.74}$$

$$q(\beta_n) = \beta_n^{-\gamma}\ell(\beta_n^{-1}). \tag{1.75}$$

From this expression, Weissman proposed a semi-parametric estimator of the extreme quantile (Weissman, 1978), by dividing (1.75) by (1.74). We get, for a small β_n and $\alpha_n \leq \beta_n$:

$$q(\alpha_n) \simeq q(\beta_n)\left(\frac{\beta_n}{\alpha_n}\right)^{\gamma}. \tag{1.76}$$

Thus, we obtain the Weissman estimator defined by:

$$\hat{q}_W(\alpha_n) = \hat{q}(\beta_n)\left(\frac{\beta_n}{\alpha_n}\right)^{\hat{\gamma}_n}. \tag{1.77}$$

For more details on the properties of Weissman estimator, see Embrechts et al. (2013); Weissman (1978). Weissman proposes to estimate $q(\beta_n)$ by its empirical equivalent $\hat{q}(\beta_n) = Y_{n-\lfloor n\beta_n \rfloor+1,n}$. There exist several semi-parametric methods dedicated to the estimation of the tail-index γ in the literature, the best known being the Hill estimator. This estimator was introduced by Hill (1975) to estimate in a non-parametric way the tail parameter of distributions belonging to the Fréchet domain of attraction. The Hill estimator is defined by:

$$\hat{\gamma}_n^H = \frac{1}{k_n - 1}\sum_{i=1}^{k_n-1} \log\left(Y_{n-i+1,n}\right) - \log\left(Y_{n-k_n+1,n}\right), \tag{1.78}$$

where $1 < k_n < n$. The method for constructing the Hill estimator is given in de Haan and Ferreira (2007)[Page 69]. Properties of Hill's estimator are the subject of many works, such as Mason (1982); Deheuvels et al. (1988) on weak and strong consistency, or the asymptotic normality (Haeusler and Teugels, 1985; de Haan and Resnick, 1998; Smith, 1987).
For distributions belonging to the Gumbel domain of attraction, $F \in \mathcal{D}(\text{Gumbel})$, the semi-parametric estimation of extreme quantiles is based on subfamilies of distributions, in particular those with a Weibull tail (Beirlant et al., 1996). According to the expression of the quantile function given in equation (1.60), for some $\beta > 0$, we have:

$$q(\alpha_n) = (-\log(\alpha_n))^{\beta}\ell(-\log(\alpha_n)) \quad \text{with} \quad \ell \in RV_0. \tag{1.79}$$

We take the logarithm and divide (1.79) by $\log(\log(1/\alpha_n))$ and we get:

$$\frac{\log q(\alpha_n)}{\log(\log(1/\alpha_n))} = \beta + \frac{\log \ell(\log(1/\alpha_n))}{\log(\log(1/\alpha_n))}.$$

According to the statement 1 of Proposition 1.5.1, $\log \ell(\log(1/\alpha_n))/(\log(\log(1/\alpha_n)) \underset{n\to\infty}{\longrightarrow} 0$, thus we have, when $n \to \infty$:

$$\log q(\alpha_n) \simeq \beta \log(\log(1/\alpha_n)). \tag{1.80}$$

Consider a sequence (k_n), such that $k_n \to \infty$ and $k_n/n \to 0$ when $n \to \infty$ and let $p_n = k_n/n$, and we write:

$$\log(q(p_n)) \simeq \beta \log(\log(1/p_n)). \tag{1.81}$$

Subtracting (1.81) from (1.80) and by switching to exponential, we obtain:

$$q(\alpha_n) \simeq q(p_n) \left(\frac{\log(1/\alpha_n)}{\log(1/p_n)}\right)^{\beta}. \tag{1.82}$$

Hence, (Beirlant et al., 1996) propose the following extreme quantile estimator:

$$\hat{q}(\alpha_n) = \hat{q}(p_n) \left(\frac{\log(1/\alpha_n)}{\log(1/p_n)}\right)^{\hat{\beta}_n}. \tag{1.83}$$

To estimate $q(p_n)$, we use its empirical equivalent $\hat{q}(p_n) = Y_{n-k_n+1,n}$. For the Weibull tail coefficient β, there exists many estimators dedicated to the estimation of this coefficient such as the approach based on the logarithms of the excesses above a threshold (Beirlant et al., 1995, 2006; Broniatowski, 1993) or the approach based on spacing between logarithms (log-spacings) (Beirlant et al., 1996; Gardes and Girard, 2011; Girard, 2004; Gardes and Girard, 2008a). For example, Girard (2004) proposed an estimator for β given by:

$$\hat{\beta}_n^G = \frac{1}{\sum_{i=1}^{k_n} [\log(\log(n/i)) - \log(\log(n/k_n))]} \sum_{i=1}^{k_n} [\log(Y_{n-i+1,n}) - \log(Y_{n-k_n,n})]. \tag{1.84}$$

The method of construction of this estimator is detailed in (Girard, 2004).
If we are restricted to the Weibull domain of attraction, equation (1.57) gives a characterisation of the quantile function for $F \in \mathcal{D}(\text{Weibull})$.

1.5.4 Estimation of conditional extreme quantiles

In this section, we focus on the case where the random variable of interest Y depends on a explanatory vector X, called covariate. The goal is to describe how tail characteristics such as extreme quantiles of Y may depend on X. In contrast to classical regression, the estimation of

extreme conditional quantiles has been little studied. In practice, there is a growing interest in many applications for the estimation of conditional extremes quantiles in the regression context. One can mention, for example, the study of the influence of meteorological parameters (temperature, humidity, sunshine, etc), agricultural inputs (pesticides, fertilisers, etc) and risk management (insurance premiums, subsidies, etc) on the low values of crop yields. Further motivations in other fields include the study of extreme temperatures as a function of several topological parameters (Ferrez et al., 2011), the modelling of the conditional quantiles of hedge fund returns using a set of risk factors (Meligkotsidou et al., 2009), or the analysis of extreme earthquakes as a function of location (Pisarenko and Sornette, 2003), to name a few. Such situations require the estimation of conditional extreme quantiles in order to quantify the relationship between Y and X.

Definition 1.5.6. *The conditional extreme quantile of order α_n associated with the conditional distribution function $F(\cdot|x)$ is defined by:*

$$q(\alpha_n|x) := F^{\leftarrow}(1-\alpha_n|x) = \inf\{y,\ F(y|x) \geq \alpha_n\} \quad \textit{with}\ \ 1-\alpha_n \to 0\ \ \textit{as}\ \ n \to \infty. \tag{1.85}$$

To estimate the conditional extreme quantiles in the presence of a covariate X, it is necessary to use estimators of the parameters of the GEV and GPD distributions adapted to the conditional case. In the literature, estimation of conditional extreme quantiles has been studied from various points of view: parametric, non-parametric and semi-parametric approaches. In the framework of these approaches, three cases are distinguished according to the nature of the covariate X: the "fixed design" setting where X is non-random, the "random design" setting where X is random and the "functional covariates" setting when X is functional.

1.5.4.1 Parametric approach

This approach model extreme conditional quantiles by fitting parametric distributions, see for instance Chernozhukov (2005) who deals with extreme quantiles in the linear regression model and derives their asymptotic behaviour under several error distributions. The work of Smith (1989) proposes to model the maxima by an extreme values distribution whose parameters are functions of the covariate and the estimation is performed by maximum likelihood or least squares. Then, Davison and Smith (1990) propose to model exceedances above a high threshold by a GPD distribution whose parameters are also functions of the covariate which are estimated by maximum likelihood. Other parametric models are proposed in Chavez-Demoulin and Davison (2005); Wang and Li (2013) using extreme-value techniques to model exceedances above a high threshold.

1.5.4.2 Semi-parametric approach

Semi-parametric approaches have also been considered for modelling trends in extreme values. One can mention Beirlant and Goegebeur (2003) who derived a semi-parametric approach by transforming the data and then using them in an exponential regression model, where the parameters are estimated by the maximum likelihood method. Hall and Tajvidi (2000) proposed the non-parametric estimation of the time trend when fitting parametric models to the extreme values of a weakly dependent time series. A semi-parametric approach to modeling trends in extremes values, based on local polynomial fitting of the Generalized extreme-value distribution, is introduced in (Davison and Ramesh, 2000). Another Local polynomial fitting of extreme-value models has been developed in (Beirlant and Goegebeur, 2004), where the regression is based on a Pareto-type conditional distribution of Y. Ahmad et al. (2020) also adopts a semi-parametric approach by proposing a location-dispersion regression model for heavy-tailed distributions.

1.5.4.3 Non-parametric approach

The non-parametric estimation of conditional extreme quantiles has been the subject of several works. Thus, we will distinguish the existing works in the literature according to the three categories defined above, namely fixed design, random design and functional covariates setting.

Estimation in a fixed design setting. Non-parametric estimation of conditional extreme quantiles has been first introduced in Davison and Ramesh (2000), where the authors use a local polynomial modeling of the extreme observations. Spline estimators of conditional extreme quantiles are used in Chavez-Demoulin and Davison (2005) through a penalised maximum likelihood method. These results are extended to multidimensional covariates case in Beirlant and Goegebeur (2004) where the asymptotic properties of the local polynomial estimators are established. Gardes and Girard (2010) proposed a nearest neighbour technique for conditional extreme quantiles estimators. The authors consider independent copies $(Y_i, x_i)_{1\leq i\leq n}$ of the random pair $(Y, x) \in \mathbb{R}^d$, where the conditional distribution function of Y is heavy-tailed and x is a deterministic covariate defined on a metric space E associated to a distance d. For a given $t \in E$, they propose to estimate the conditional extreme quantiles $q(\alpha_{n,t}, t)$ when $\alpha_{n,t} \to 0$ as $n \to \infty$. Let $(m_{n,t})$ be a sequence such that $1 < m_{n,t} < n$ and let $\{x_1^*, ..., x_{m_{n,t}}^*\}$ be the $m_{n,t}$ nearest covariates of t (with respect to the distance d). The associated observations taken from Y_i are denoted by $\{Z_i(t), i = 1, \ldots, m_{n,t}\}$. Denoting the corresponding order statistics by $Z_{1,m_{n,t}} \leq ... \leq Z_{m_{n,t},m_{n,t}}$, the conditional extreme quantiles

estimator is given by:

$$\hat{q}^W(\alpha_{n,t},t) \quad = \quad Z_{m_{n,t}-k_{n,t}+1,m_{n,t}}(t)\left(\frac{k_{n,t}}{m_{n,t}\alpha_{n,t}}\right)^{\hat{\gamma}_n(t)}, \tag{1.86}$$

where $(k_{n,t})$ is a sequence such that $1 < k_{n,t} < m_{n,t}$, and $\hat{\gamma}_n(t)$ is an estimator of the conditional tail-index. The estimator proposed above is in the same spirit as the quantile estimator proposed by Weissman (see (1.77)). Moreover, the authors proposed to estimate the conditional tail-index by the Hill estimator adapted to the conditional framework. Gardes and Girard (2008b) proposed also a tail-index estimator using a moving window approach and extended the estimators proposed in Beirlant et al. (2002) to the conditional context.

Estimation in a random design setting. One can cite Daouia et al. (2011); Goegebeur et al. (2014) who studied the estimation of extreme quantiles under a conditional heavy-tail model, later extended in Daouia et al. (2013) to conditional distributions belonging to any maximum domain of attraction. For instance, the authors in Daouia et al. (2011) propose a kernel estimation of the extreme quantiles in the presence of a finite dimension covariate. Considering $(Y_i, X_i)_{1\leq i\leq n}$ independent copies of the random pair $(Y, X) \in \mathbb{R}\times\mathbb{R}^p$, where Y is heavy-tailed with a conditional survival function $\overline{F}(y|x)$, they propose the following estimator:

$$\hat{\overline{F}}_n(y|x) = \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h})\mathbb{1}_{\{Y_i>y\}}}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})},$$

where K is a probability density on $\mathbb{R}^d$, called kernel, and $h = h_n$ is a non-random sequence (called window-width) such that $h_n \to 0$ as $n \to \infty$. Then, to estimate the conditional extreme quantiles, they propose a Weissman estimator (see (1.77)) adapted to the conditional case given by:

$$\hat{q}_n^W(\beta_n|x) = \hat{q}_n(\alpha_n|x)\left(\frac{\alpha_n}{\beta_n}\right)^{\hat{\gamma}_n(x)}, \tag{1.87}$$

where $\hat{q}_n(\alpha_n|x) = \hat{F}_n^{\leftarrow}(1-\alpha_n|x) = \inf\left\{y,\ \hat{F}_n(y|x) \geq 1-\alpha_n\right\}$ is the intermediate conditional extreme quantile estimator and $\hat{\gamma}_n(x)$ is an estimator of the tail-index. A Hill estimator $\hat{\gamma}_n^H$ adapted to the conditional framework was also proposed to estimate the tail-index. Gardes and Stupfler (2014) also considered the estimation of the conditional tail-index by proposing a smoothed local Hill estimator adapted to the presence of a finite dimension covariate. One can mention also Goegebeur et al. (2014) who introduced a family of non-parametric estimators of the conditional tail-index in the presence of a random covariate.

Estimation in a functional covariates setting. In the case of functional covariate of infinite dimension, Gardes and Girard (2012) have proposed a functional kernel estimators of conditional extreme quantiles based on a functional Weissman estimator. They also proposed functional versions of the Hill and Pickands estimators for the tail-index. Gardes et al. (2010) have addressed conditional extreme quantiles estimators using the moving window approach. They proposed an adaptation of Weissman estimator in the case where covariate information is functional.

1.5.4.4 Dimension reduction in the extreme values framework

Nowadays, large volumes of data are stored and high-dimensional covariates problems are encountered. In such situations, inference on the tail distribution of Y given X becomes difficult since the space is sparsely populated with data points. Indeed, only a few points can be considered to estimate the quantiles and classical estimators become inefficient, unless the sample size is very large. This is the so-called "curse of dimensionality". Thus, a dimension reduction becomes necessary to overcome the latter and to reveal the most relevant directions of the high-dimensional covariate space. In practice, there is a growing interest in many applications to combine dimension reduction methods with conditional extremes quantiles. However, there is a limited literature on the combination of these two lines of work.
One can mention Gardes (2018) who established a dimension reduction method suited to the case where the tail distribution of Y depends on the projection of the covariates on a lower dimensional subspace. A classical approach is to assume the existence of a $p \times q$ full rank matrix B, with $q < p$, such that X and Y are independent conditionally on $B^t X$. As mentioned in Gardes (2018), this assumption can be strong when we are concerned with the tail of the conditional distribution. Thus, the author introduces a notion of tail conditional independence. In other words, he assumes the existence of a Tail Dimension Reduction subspace[12] spanned by B, such that the tail of the conditional distribution of Y given X can be approximated by the tail of the conditional distribution of Y given $B^t X$. Then, a new kernel estimator of conditional extreme quantiles is proposed. In the same vein, Aghbalou et al. (2021) also address dimensionality reduction for quantile regression by considering a low-dimensional orthogonal projection to explain the tail distribution of Y. The authors developed a model based on sliced inverse regression (SIR) method, using the notion of tail conditional independence in order the Extreme Sufficient Dimension Reduction space[13].
In Xu et al. (2020) a semi-parametric approach is introduced for the estimation of extreme

[12] Tail Dimension Reduction subspace is an adaptation of the Dimension Reduction subspace introduced by Li (1991) to extreme case.
[13] Extreme Sufficient Dimension Reduction space is tail version of classical Sufficient Dimension Reduction space (Cook and Ni, 2005).

conditional quantiles $q(\alpha|x)$ basing on a tail single-index model. Considering independent copies $(Y_i, X_i)_{1\leq i\leq n}$ of the random pair $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$, where the conditional distribution function of Y is heavy-tailed, they estimate the dimension reduction direction β under the global single-index quantile regression model and the conditional mean linearity assumption, which is satisfied when X is elliptically distributed. Indeed, β is estimated through fitting a misspecified linear quantile regression model, i.e. by solving the following optimisation problem:

$$\arg\min_{u,\beta} \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(Y_i - u - X_i^t\beta), \tag{1.88}$$

where $\rho_\alpha(r) = r(\alpha - \mathbb{1}_{\{r<0\}})$ is the quantile check loss function (Koenker et al., 2005) and u is the intercept. Secondly, they estimate the conditional quantile at intermediate quantile levels $q(\alpha|X^t\beta)$ by applying a local linear quantile regression and then estimate the tail-index through a Hill-type estimator. Then, they use extrapolation, from the intermediate quantile level to the extreme tail, to estimate the extreme conditional quantile by adapting Weissman's estimator. Finally, another approach proposed by Drees and Sabourin (2021); Cooley and Thibaud (2019) consists in adapting the (unsupervised) dimension reduction method PCA to an extreme setting.

1.5.4.5 Bayesian inference in the extreme values framework

The issues with extreme value methods are characterised by the scarcity of extreme events which limits the available data, and the requirement to model where the data are most sparse. The use of Bayesian inference would allow to deal with problems with a very small amount of data and to incorporate any available information or beliefs as prior information. Thus, if an expert can provide meaningful prior information on the data, the quality of inference could be improved in extreme values problems. Examples in the literature include: Coles and Tawn (1990) integrated expert knowledge formulated into prior information as the basis for a Bayesian analysis of extreme rainfall, Walshaw (2000) proposed a mixture model for extreme wind speeds which incorporates the prior distribution, Smith and Goodman (2000); Bottolo et al. (2003) discussed the choice of prior and posterior evaluation for hierarchical modelling of extreme values in insurance problems, Smith (1998) compared the predictive inference aspects of Bayesian and frequentist approaches and Engelund and Rackwitz (1992) used Bayesian approach to estimate parameters of a three extreme value distributions.
There exist other works in the literature relating the topics of extreme value modelling and Bayesian inference, which can be split into three categories. The first category includes the Peak-Over-Threshold model (which corresponds to GPD distributions). One can mention

for example: Pickands (1994) considered Bayesian estimation of extreme quantiles in the GPD model with uninformative independent priors for parameters, de Zea Bermudez and Turkman (2003) proposed using a vague proper prior for GPD parameters, de Zea Bermudez et al. (2001) used a Bayesian predictive approach to the peaks over threshold method through a hierarchical Bayesian model, Behrens et al. (2004) proposed an elicitation technique to obtain a prior distribution on GPD parameters, Diebolt et al. (2005) proposed a quasi-conjugate Bayesian inference approach for estimating GPD parameters, distribution tails and extreme quantiles, Castellanos and Cabras (2007) considered a default Bayesian method for GPD parameters estimation when prior information is not available using Jeffreys's prior. The second category covers Block Maxima (which corresponds to a GEV distribution). For instance, Coles et al. (2001) (Chapter 9) used GEV model for annual maximum sea levels and placed independent normal priors on the parameters of the model, Rostami and Adam (2013) considered the estimation of GEV parameters, Coles and Tawn (1996) constructed the prior for the GEV model, Northrop and Attalides (2016) used improper priors for the GEV model (respectively for GPD model) parameters, including Jeffreys prior, the maximal data information (MDI) prior and independent uniform priors. Finally, for the Poisson Process category, one can mention Sharkey and Tawn (2017) who presented a Bayesian approach for estimation of the Poisson process model parameters.

Chapter 2

Extreme Partial Least-Squares

Contents

Abstract

In a conditional extremes setting, where the extreme values of the response variable $Y \in \mathbb{R}$ depend on the covariate vector $X \in \mathbb{R}^p$, the conditional distribution of Y given X is difficult to estimate when p is very large compared to the sample size. This problem is particularly accentuated in the case of extreme value analysis where the number of largest observations, considered as representative of the tail of the distribution, is small. Thus, dimensionality reduction is a key step in the extreme value analysis framework. In this chapter, we propose a new approach, called Extreme-PLS, which combines the partial least squares dimension reduction method and extreme value analysis, in the context of a nonlinear inverse regression model. This chapter is presented as an article submitted for publication (Bousebata et al., 2021). After introducing the background of the study (Section 2.1), the Extreme-PLS approach is presented in the context of a single-index non linear inverse regression model and heavy-tailed distributions (Section 2.2). The approach estimates the reduction dimension direction, by maximizing the covariance between a linear combination of X and Y given large values of Y. The associated empirical estimator is exhibited in Section 2.3 and its asymptotic normality is established. An iterative procedure to adapt the approach to the multiple-index situation is presented in Section 2.4. We illustrate the performance of the approach by applying it to simulated data in Section 2.5. In particular, we show that it performs well in high dimension, whether or not linearity or independence assumptions are satisfied. It is also shown that the Extreme-PLS estimator is more efficient than the proposed estimator of (Xu et al., 2020) in some situations. Then, we perform an application on real data of French farm income, in Section 2.6, to analyse the lowest cereal yields given different factors. A discussion is provided in Section 2.7 and proofs are postponed to the Appendix.

Resumé

Dans un contexte des extrêmes conditionnelles, où les valeurs extrêmes de la variable de réponse $Y \in \mathbb{R}$ dépendent d'un vecteur de covariables $X \in \mathbb{R}^p$, la distribution conditionnelle de $Y \in \mathbb{R}$ étant donné $X \in \mathbb{R}^p$ est difficile à estimer lorsque p est très grand par rapport à la taille de l'échantillon. Ce problème est particulièrement accentué dans le cas de l'analyse des valeurs extrêmes où le nombre d'observations les plus grandes, considérées comme représentatives de la queue de la distribution, est faible. Ainsi, la réduction de la dimensionnalité est une étape essentielle dans le cadre de l'analyse des valeurs extrêmes. Dans ce chapitre, nous proposons une nouvelle approche, appelée Extreme-PLS, qui combine la méthode de réduction de dimension "partial least squares" et l'analyse des valeurs extrêmes, dans le contexte d'un modèle de régression inverse non linéaire. Ce chapitre est présenté comme un article soumis pour publication (Bousebata et al., 2021). Après avoir introduit le contexte de l'étude (Partie 2.1), l'approche Extreme-PLS est présentée dans le contexte d'un modèle de régression inverse non linéaire à indice unique et de distributions à queue lourde (Partie 2.2). L'approche estime la direction de la réduction de dimension, en maximisant la covariance entre une combinaison linéaire de X et Y étant donné de grandes valeurs de Y. L'estimateur empirique associé est présenté à la Partie 2.3 et sa normalité asymptotique est établie. Une procédure itérative permettant d'adapter l'approche à la situation des indices multiples est ensuite présentée à la Partie 2.4. Nous illustrons les performances de l'approche en l'appliquant à des données simulées dans la Partie 2.5. En particulier, nous montrons qu'elle est performante en haute dimension, que des hypothèses de linéarité ou d'indépendance soient satisfaites ou non. Il a également été montré que l'estimateur Extreme-PLS est plus efficace que l'estimateur proposé de (Xu et al., 2020) dans certaines situations. Ensuite, nous réalisons une application sur des données réelles du revenu agricole français, dans la Partie 2.6, pour analyser les rendements céréaliers les plus faibles en fonction de différents facteurs. Une discussion est fournie dans la Partie 2.7 et les preuves sont reportées en Annexe.

2.1 Introduction

One of the main goals of statistical analysis is to seek a relationship between a response variable Y and a p-dimensional vector X of covariates starting from a n-sample. A common way to describe the possible link is to use the regression function $\mathbb{E}(Y|X)$. However, in many situations, the entire conditional distribution of Y given X may be of interest, rather than the only central part. This has led to the development of conditional quantiles, or regression quantiles, as an alternative to the conditional mean. Quantile regression was introduced by (Koenker and Bassett Jr, 1978) in a parametric framework. Since then, non-parametric regression methods have been considered in the literature. Among others, (Bhattacharya and Gangopadhyay, 1990) studied kernel and nearest neighbour estimators of conditional quantiles while (He et al., 1998) focused on spline methods.

A complementary way to investigate the relationship between Y and X is to focus on conditional extremes. The goal is then to describe how tail characteristics such as extreme quantiles of Y may depend on the explanatory vector X. One motivating example in agricultural risk management is the study of the influence of meteorological parameters (temperature, humidity,...), agricultural inputs (pesticides, fertilisers,...) and risk management tools (insurance premiums, subsidies,...) on low values of crop yields (see Section 2.6). Other motivations can be found in finance (Meligkotsidou et al., 2009), climatology (Jagger and Elsner, 2009), hydrology (Gardes and Girard, 2010) and environment (Smith, 1989), to name a few. In these applications, the estimation of extreme conditional quantiles is a crucial issue that has been studied from various points of view. One first approach relies on the fit of a parametric model, see for instance (Chernozhukov, 2005) who deals with extreme quantiles in the linear regression model and derives their asymptotic behaviour under several error distributions. Other parametric models are proposed in (Chavez-Demoulin and Davison, 2005; Davison and Smith, 1990; Smith, 1989) using extreme-value techniques to model exceedances above a high threshold. A second line of work relies on non-parametric approaches that can be split into three main categories: fixed design, random design, and functional covariates. Fixed design methods aim at estimating conditional extreme quantiles depending on a non-random covariate, see (Gardes and Girard, 2010) for a nearest neighbors technique. In the random design setting, one can cite (Daouia et al., 2011; Goegebeur et al., 2014) who studied the estimation of extreme quantiles under a conditional heavy-tail model, later extended in (Daouia et al., 2013) to conditional distributions belonging to any maximum domain of attraction. Finally, see (Gardes and Girard, 2012) for the functional covariate situation. Semi-parametric approaches have also been considered for trend modelling in extreme values. Local polynomial fitting of extreme-value models is investigated in (Beirlant and Goegebeur, 2004; Davison

and Ramesh, 2000), a non-parametric estimation of the temporal trend is combined with parametric models for extreme values in (Hall and Tajvidi, 2000), and a location-dispersion regression model for heavy-tailed distributions is introduced in (Ahmad et al., 2020).

In a high dimensional context, *i.e.* when the dimension p of X is large, the above mentioned estimation methods may suffer from the so-called "curse of dimensionality". This phenomenon results in an exploding variance of the estimators, thus impeding the inference in practice. A number of statistical approaches to dimension reduction were introduced to circumvent this issue and to reveal the most relevant directions in the high-dimensional covariate space. One of the most popular ones is Partial least squares (PLS), introduced by (Wold, 1975), that combines characteristics of Principal component analysis (PCA) for dimension reduction and multiple regression. The development of PLS has been initiated within the chemometrics field, see the reference book (Martens and Naes, 1992). Since then, PLS has also received attention in the statistical literature. For example, (Helland, 1990) discusses the statistical properties of the PLS procedure under a factor analysis model, while (Frank and Friedman, 1993) provides a comparison between PLS and Principal component regression (PCR) from various perspectives. See also (Cook et al., 2013) for a connection between PLS approach and envelopes and (Chun and Keleş, 2010) for a sparse version of PLS. The basic idea of PLS is to seek directions, *i.e.* linear combinations of X coordinates having both high variance and high correlation with Y, unlike PCR method which only takes into account high variance components (Frank and Friedman, 1993; Stone and Brooks, 1990). Sliced inverse regression (SIR) (Li, 1991) is an alternative method that takes advantage of the simplicity of the inverse regression view of X against Y. It aims at replacing X by its projection onto a subspace of smaller dimension without loss of regression information. Many extensions of SIR have been proposed (Girard et al., 2022) such as Partial inverse regression handling the $p > n$ situation (Li et al., 2007), Kernel sliced inverse regression allowing the estimation of a nonlinear subspace (Wu, 2008), Student sliced inverse regression dealing with heavy-tailed errors (Chiancone et al., 2017), Sliced inverse regression for multivariate response (Coudret et al., 2014), among others. Single-index models provide additional practical tools to overcome the curse of dimensionality, by modelling the non-linear relationship between Y and X through an unknown link function and a single linear combination of the covariates referred to as the index, see (Horowitz, 2009b, Chapter 2). As such, they provide a reasonable compromise between non-parametric and parametric approaches. Among the numerous works dedicated to the estimation of the index and the link function, the most popular ones are the average derivative estimation method in the context of kernel smoothing (Härdle and Stoker, 1989; Powell et al., 1989), and the M-estimation technique based on spline regression (Wang and Yang, 2009; Yu and Ruppert, 2002). One can also mention (Kong and Xia, 2012; Wu

et al., 2010) who considered single-index models for the estimation of conditional quantiles. In (Naik and Tsai, 2000), it is proved that PLS provides a consistent estimator of the direction when Y given X follows a single-index model and when $n \to \infty$ with a fixed dimension p, under the additional assumption of independence between noise and covariates. From the practical point of view, it is also shown that PLS can perform better than SIR even though the link function is non-linear. This result is extended in (Chun and Keleş, 2010) to the multiple-index situation and when both $n \to \infty$ and $p \to \infty$ with $p/n \to 0$. It is also shown that, in contrast, the PLS estimator is no more consistent in the case where $p/n \to k > 0$, and a sparse version of PLS is introduced to avoid this vexing effect. More recently, (Cook and Forzani, 2018) tempered this conclusion by exhibiting some designs under which single-index PLS is consistent in the linear regression situation, when both $n \to \infty$ and $p \to \infty$, regardless of the alignment between n and p.

Finally, the curse of dimensionality may also be tackled using shrinkage methods which aim at reducing the complexity of the inference by variable selection. As an example, Lasso method (Tibshirani, 1996) penalizes regression coefficients similarly to ridge regression (Hoerl and Kennard, 1970) but replacing the L_2 penalization by the L_1 counterpart. Some extensions include Fused lasso (Tibshirani et al., 2005) and Elastic net (Zou and Hastie, 2005) to deal with the case where p is larger than n. Many other shrinkage and variable selection methods are discussed in (Hastie et al., 2001, Chapter 3).

Dimension reduction dedicated to conditional extremes is limited in the literature, and only a few recent works have been devoted to it. One can mention (Gardes, 2018) where a dimension reduction framework adapted to conditional tail distributions is developed assuming that, when Y is large, X and Y are independent conditionally on a linear combination of the covariates. Another approach (Cooley and Thibaud, 2019; Drees and Sabourin, 2021) consists in adapting the (unsupervised) dimension reduction method PCA to the extreme setting. In (Xu et al., 2020), a semi-parametric approach is introduced for the estimation of extreme conditional quantiles based on a tail single-index model. The authors propose to estimate the dimension reduction direction β using local linear quantile regression. The method is developed under the tail single-index model and a conditional mean linearity assumption, which is satisfied, for instance, when X is elliptically distributed (the method is described in further details in Section 2.5).

We introduce a new approach, referred to as extreme-PLS (EPLS), for dimension reduction in an extreme conditional setting. The underlying idea is to look for linear combinations of covariates that best explain the extreme values of Y. More precisely, we first propose a single-index approach to find a direction $\hat{\beta}$ maximizing the covariance between $\beta^t X$ and Y given Y exceeds a high threshold y. An iterative procedure is then exhibited to adapt

the method to the multiple-index situation. In practice, $\hat{\beta}$ allows to quantify the effect of the covariates on the extreme values of Y in a simple and interpretable way. Plotting Y against the projection $\hat{\beta}^t X$ provides a visual interpretation of conditional extremes. Moreover, working on the pair $(\hat{\beta}^t X, Y)$ should yield improved results for most estimators dealing with conditional extreme values thanks to the dimension reduction achieved in the projection step. From the theoretical point of view, the asymptotic properties of the EPLS estimator are established under an inverse single-index model and a heavy tail assumption, without recourse to linearity as in (Xu et al., 2020) nor independence assumptions as in (Gardes, 2018). It appears on simulated data that the EPLS estimator provides promising results in high-dimensional settings and outperforms the estimator proposed in (Xu et al., 2020) in a wide range of situations.

The paper is organized as follows. In Section 2.2, the EPLS approach is introduced in the framework of a single-index model and heavy-tailed distributions. Some preliminary properties are stated in order to justify the above heuristics from a theoretical point of view. The associated estimator is exhibited in Section 2.3 and its asymptotic distribution is established under mild assumptions. This approach is extended to the multiple-index setting in Section 2.4. The performances of the method are investigated through a simulation study in Section 2.5. EPLS approach is then applied in Section 2.6 to assess the influence of various parameters on cereal yields collected on French farms. A small discussion is provided in Section 2.7 and proofs are postponed to the Appendix. A Supplementary material is also provided to complete the simulation study. Data and R code are available at `https://github.com/meryembst/EPLS`.

2.2 Single-index EPLS approach

Let Y be a real random response variable and X a p-dimensional random covariate. We denote by $w(y)$ the unit vector maximizing the covariance between $w^t X$ and Y given that Y exceeds a large threshold y:

$$w(y) = \arg \max_{\|w\|=1} \operatorname{cov}(w^t X, Y | Y \geq y). \tag{2.1}$$

This linear optimization problem under a quadratic constraint benefits from a closed-form solution obtained with Lagrange multipliers method and given in the next Proposition. For all $y \in \mathbb{R}$, introduce $\bar{F}(y) = \mathbb{P}(Y \geq y)$ the survival function of Y and consider the three tail-moments, whenever they exist, $m_Y(y) = \mathbb{E}(Y \mathbb{1}_{\{Y \geq y\}}) \in \mathbb{R}$, $m_X(y) = \mathbb{E}(X \mathbb{1}_{\{Y \geq y\}}) \in \mathbb{R}^p$, $m_{XY}(y) = \mathbb{E}(XY \mathbb{1}_{\{Y \geq y\}}) \in \mathbb{R}^p$.

Proposition 2.2.1. *Suppose that* $\mathbb{E}(\|X\|\mathbb{1}_{\{Y\geq y\}}) < \infty$, $\mathbb{E}(|Y|\mathbb{1}_{\{Y\geq y\}}) < \infty$ *and* $\mathbb{E}(\|XY\|\mathbb{1}_{\{Y\geq y\}}) < \infty$ *for all* $y \in \mathbb{R}$. *Then, the unique solution of the optimization problem (2.1) is given for all* $y \in \mathbb{R}$ *by:*

$$w(y) = v(y)/\|v(y)\| \text{ where } v(y) = \bar{F}(y)m_{XY}(y) - m_X(y)m_Y(y). \tag{2.2}$$

Let us note that solution (2.2) is invariant with respect to the scaling and location of X. In the following, we aim at investigating the behaviour of $w(y)$ for large thresholds, *i.e.* as $y \to \infty$. To this end, consider the following single-index non linear inverse regression model:

($\mathbf{M}_1$) $X = g(Y)\beta + \varepsilon$, where X and ε are p-dimensional random vectors, Y is a real random variable, $g : \mathbb{R} \to \mathbb{R}$ is an unknown link function, $\beta \in \mathbb{R}^p$ is an unknown unit vector.

Let us highlight that no independence assumption is made on the pair (X, ε). However, in the particular case where ε is centered and independent of Y, we recover the classical PLS framework and it is easily shown that $w(y) = \pm\beta$ for all $y \in \mathbb{R}$. Model ($\mathbf{M}_1$) is referred to as an inverse regression model since the covariates are written as functions of the response variable. Similar models were used to establish the theoretical properties of SIR, see for instance (Bernard-Michel et al., 2009; Cook, 2007). Under model ($\mathbf{M}_1$), when Y is large, provided the distribution tail of ε is negligible, one has $X \simeq g(Y)\beta$ leading to the approximate single-index forward model $Y \simeq g^{-1}(\beta^t X)$. Our first goal is to establish the convergence of $w(y)$ towards β, as $y \to \infty$, without resorting to a linear conditional expectation assumption as in (Xu et al., 2020) nor a conditional independence assumption as in (Gardes, 2018; Saracco, 1997). In contrast, additional assumptions on the link function g and the distribution tail of Y are considered:

($\mathbf{A}_1$) Y is a real random variable with density function f regularly varying at infinity with index $-1/\gamma - 1$, $\gamma \in (0, 1)$ *i.e.* for all $t > 0$,

$$\lim_{y\to\infty} \frac{f(ty)}{f(y)} = t^{-\frac{1}{\gamma}-1}.$$

This property is denoted for short by $f \in RV_{-1/\gamma-1}$.

($\mathbf{A}_2$) $g \in RV_c$ with $c > 0$.

($\mathbf{A}_3$) There exists $q > 1/(\gamma c)$ such that $\mathbb{E}(\|\varepsilon\|^q) < \infty$.

Let us note that ($\mathbf{A}_1$) implies that $\bar{F} \in RV_{-1/\gamma}$ in view of Karamata's theorem (Bingham et al., 1987, Theorem 1.5.8). In other words, ($\mathbf{A}_1$) entails that Y has a right heavy-tail. This is equivalent to assuming that the distribution of Y is in the Fréchet maximum domain of

attraction, with tail-index $\gamma > 0$, see (de Haan and Ferreira, 2007, Theorem 1.2.1). This domain of attraction includes for Pareto, Burr and Student distributions, see (Beirlant et al., 2004) for further examples of heavy-tailed distributions. The restriction to $\gamma < 1$ ensures that $\mathbb{E}(|Y|\mathbb{1}_{\{Y \geq y\}})$ exists for all $y \in \mathbb{R}$. Assumption $(\mathbf{A}_2)$ means that the link function ultimately behaves like a power function. Finally, $(\mathbf{A}_3)$ can be interpreted as an assumption on the tails of $\|\varepsilon\|$. It is satisfied, for instance, by distributions with exponential-like tails such as Gaussian, Gamma or Weibull distributions. More specifically, $\mathbb{E}(\|\varepsilon\|^q) < \infty$ implies that the tail-index, say $\gamma_{\|\varepsilon\|}$, associated with $\|\varepsilon\|$ is such that $\gamma_{\|\varepsilon\|} < 1/q$. Besides, it can readily be shown that $g(Y)$ is heavy-tailed with tail-index $\gamma_{g(Y)} := c\gamma$. Condition $(\mathbf{A}_3)$ thus imposes that $\gamma_{g(Y)} > \gamma_{\|\varepsilon\|}$, meaning that $g(Y)$ has an heavier right tail than $\|\varepsilon\|$. In model $(\mathbf{M}_1)$, the tail behavior of $|\beta^t X|$ and $\|X\|$ is thus driven by $g(Y)$ rather than by $|\beta^t \varepsilon|$, *i.e.* $\gamma_{\|X\|} = \gamma_{g(Y)}$, which is the desired property.

In order to assess the convergence of $w(y)$ to β as $y \to \infty$, we let

$$\Delta(w(y), \beta) := 1 - \cos^2(w(y), \beta) = 1 - (w(y)^t \beta)^2. \tag{2.3}$$

A value close to 1 implies a low colinearity ($w(y)$ is almost orthogonal to β) while a value close to 0 means a high colinearity.

Proposition 2.2.2. *Assume* $(\mathbf{M}_1)$, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ *and* $(\mathbf{A}_3)$ *hold with* $\gamma(c+1) < 1$. *Then,*

$$\Delta(w(y), \beta) = O\left\{\left(\frac{1}{g(y)\bar{F}^{1/q}(y)}\right)^2\right\} \to 0 \quad \textit{and} \quad \|w(y) - \beta\| = O\left(\frac{1}{g(y)\bar{F}^{1/q}(y)}\right) \to 0,$$

as $y \to \infty$.

It should be noted that, since $\|w(y) - \beta\| \to 0$ as $y \to \infty$, the EPLS axis has asymptotically the same direction as β (without sign issue). Besides, in view of assumptions $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$, the function $y \mapsto g(y)\bar{F}^{1/q}(y)$ is regularly varying with index $c - 1/(q\gamma) > 0$ from $(\mathbf{A}_3)$. Unsurprisingly, the above convergence rates are large when c is large (*i.e.* the link function is rapidly increasing), q is large (*i.e.* the noise ε is small) or/and γ is large (*i.e.* the tail of Y is heavy). The inference from data distributed from model $(\mathbf{M}_1)$ is addressed in the following section.

2.3 Single-index EPLS: Estimators and main properties

Let (X_i, Y_i), $1 \leq i \leq n$ be independent and identically distributed random variables from model $(\mathbf{M}_1)$ and let $y_n \to \infty$ as the sample size n tends to infinity. The solution (2.2) is

estimated by its empirical counterpart introducing

$$\hat{v}(y_n) = \hat{\bar{F}}(y_n)\hat{m}_{XY}(y_n) - \hat{m}_X(y_n)\hat{m}_Y(y_n), \tag{2.4}$$

with $\hat{\bar{F}}$ the empirical survival function and

$$\hat{m}_{XY}(y_n) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \mathbb{1}_{\{Y_i \geq y_n\}}, \hat{m}_Y(y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\{Y_i \geq y_n\}}, \hat{m}_X(y_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}_{\{Y_i \geq y_n\}}.$$

For all $j = 1, \ldots, p$, let us denote by $X_{.,j}$ the jth coordinate of X. It readily follows that the jth coordinates of $m_X(y_n)$ and $m_{XY}(y_n)$ are respectively given by $m_{X_{.,j}}(y_n)$ and $m_{X_{.,j}Y}(y_n)$.

We first provide a tool establishing the joint asymptotic normality of $\hat{\bar{F}}(y_n)$, $\hat{m}_{XY}(y_n)$, $\hat{m}_Y(y_n)$ and $\hat{m}_X(y_n)$ when $y_n \to \infty$ and $n\bar{F}(y_n) \to \infty$. This latter condition ensures that the rate of convergence, which is driven by the number of effective observations $n\bar{F}(y_n)$ used in the estimators, tends to infinity as the sample size increases.

Proposition 2.3.1. *Assume* $(\mathbf{M}_1)$, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ *and* $(\mathbf{A}_3)$ *hold with* $2\gamma(c+1) < 1$. *Let us denote by* d *the number of non-zero* β_j *coefficients in* $\beta \in \mathbb{R}^p$, *and assume for the sake of simplicity that* $\beta_j \neq 0$ *for all* $j = 1, \ldots, d$ *and* $\beta_{d+1} = \cdots = \beta_p = 0$. *Let* $y_n \to \infty$ *such that* $n\bar{F}(y_n) \to \infty$ *and introduce the* $\mathbb{R}^{2(d+1)}-$ *random vector*

$$\Lambda_n := \left\{ \left(\frac{\hat{\bar{F}}(y_n)}{\bar{F}(y_n)} - 1 \right), \left(\frac{\hat{m}_Y(y_n)}{m_Y(y_n)} - 1 \right), \left(\frac{\hat{m}_{X_{.,j}}(y_n)}{m_{X_{.,j}}(y_n)} - 1 \right)_{1 \leq j \leq d}, \left(\frac{\hat{m}_{X_{.,j}Y}(y_n)}{m_{X_{.,j}Y}(y_n)} - 1 \right)_{1 \leq j \leq d} \right\}.$$

Then,

$$\sqrt{n\bar{F}(y_n)}\Lambda_n \xrightarrow{d} \mathcal{N}(0, B),$$

where B *is the* $2(d+1) \times 2(d+1)$ *covariance matrix defined by*

$$B = \left(\begin{array}{c|c|ccc|ccc} 1 & 1 & 1 & \ldots & 1 & 1 & \ldots & 1 \\ \hline 1 & b_{22} & b_{23} & \ldots & b_{23} & b_{24} & \ldots & b_{24} \\ \hline 1 & b_{23} & b_{33} & \ldots & b_{33} & b_{34} & \ldots & b_{34} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & b_{23} & b_{33} & \ldots & b_{33} & b_{34} & \ldots & b_{34} \\ \hline 1 & b_{24} & b_{34} & \ldots & b_{34} & b_{44} & \ldots & b_{44} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & b_{24} & b_{34} & \ldots & b_{34} & b_{44} & \ldots & b_{44} \end{array} \right)$$

and where

$$b_{22} = \frac{(1-\gamma)^2}{1-2\gamma}, \qquad b_{23} = \frac{(1-\gamma)(1-\gamma c)}{1-\gamma(c+1)},$$
$$b_{33} = \frac{(1-\gamma c)^2}{1-2\gamma c}, \qquad b_{24} = \frac{(1-\gamma(c+1))(1-\gamma)}{1-\gamma(c+2)},$$
$$b_{44} = \frac{(1-\gamma(c+1))^2}{1-2\gamma(c+1)}, \qquad b_{34} = \frac{(1-\gamma c)(1-\gamma(c+1))}{1-\gamma(2c+1)}.$$

Let us remark that the above result only provides the (joint) asymptotic distribution of $\hat{m}_{XY}(y_n)$ and $\hat{m}_X(y_n)$ in the directions associated with non-zero β_j. It is however sufficient to establish the asymptotic normality of $\hat{v}(y_n)$ centered on $v(y_n)$, the direction provided by the EPLS criterion, see Proposition 2.2.1.

Proposition 2.3.2. *Assume* $(\mathbf{M}_1)$, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ *and* $(\mathbf{A}_3)$ *hold with* $2\gamma(c+1) < 1$. *Let* $y_n \to \infty$ *such that* $n\bar{F}(y_n) \to \infty$. *Then,*

$$\sqrt{n\bar{F}(y_n)}\left(\frac{\hat{v}(y_n) - v(y_n)}{\|v(y_n)\|}\right) \xrightarrow{d} \xi\beta,$$

where $\xi \sim \mathcal{N}(0, \lambda(c,\gamma))$ *and*

$$\lambda(c,\gamma) = a_1^2(3+b_{44}) + a_2^2(2b_{23}+b_{22}+b_{33}) - 2a_1a_2(2+b_{24}+b_{34}), \tag{2.5}$$

with $a_1 = (1-\gamma)(1-\gamma c)/(\gamma^2 c)$ *and* $a_2 = (1-\gamma(c+1))/(\gamma^2 c)$.

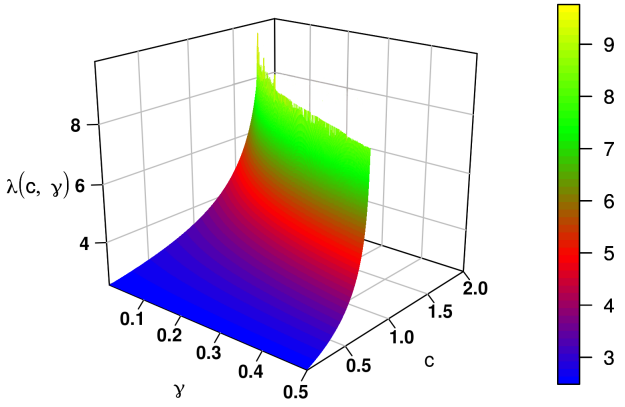


Figure 1 Asymptotic variance $(c,\gamma) \in [1/2,2] \times [0,1/2] \mapsto \lambda(c,\gamma)$ given in Proposition 2.3.2, Equation (2.5), on a logarithmic scale.

The asymptotic variance $\lambda(c,\gamma)$ is plotted on Figure 1 as a function of $(c,\gamma) \in [1/2,2] \times [0,1/2]$ and under the constraint $2\gamma(c+1) < 1$. This condition imposes an upper bound on the tail-index of $\|X\|$: $\gamma_{\|X\|} < c/(2(c+1))$, see Section 2.2. Similarly, asymptotic properties of usual dimension-reduction methods are established under the assumption that $\mathbb{E}(\|X\|^4) < \infty$ which

implies $\gamma_{\|X\|} < 1/4$, see (Saracco, 1997) in the SIR case. The latter bound is the strongest one when $c > 1$.

It appears from Proposition 2.3.2 that the asymptotic distribution of $\hat{v}(y_n)$ is Gaussian and degenerated in every direction orthogonal to β. Combining the above result with Proposition 2.2.2 provides an asymptotic normality result for $\hat{v}(y_n)$ centered on the true direction β.

Theorem 2.3.1. *Assume* $(\mathbf{M}_1)$, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ *and* $(\mathbf{A}_3)$ *hold with* $2\gamma(c+1) < 1$. *Let* $y_n \to \infty$ *such that* $n\bar{F}(y_n) \to \infty$ *and* $n\bar{F}(y_n)^{1-2/q}/g^2(y_n) \to 0$ *as* $n \to \infty$. *Then,*

$$\sqrt{n\bar{F}(y_n)}\left(\frac{\hat{v}(y_n)}{\|v(y_n)\|} - \beta\right) \xrightarrow{d} \xi\beta,$$

with $\xi \sim \mathcal{N}(0, \lambda(c,\gamma))$ *and where* $\lambda(c,\gamma)$ *is defined in (2.5).*

Assumption $n\bar{F}(y_n) \to \infty$ ensures that the variance of the estimator tends to zero while condition $n\bar{F}(y_n)^{1-2/q}/g^2(y_n) \to 0$ entails that the bias (bounded above by $1/(g(y_n)\bar{F}^{1/q}(y_n))$, see Proposition 2.2.2) is asymptotically small compared to the standard deviation $1/\sqrt{n\bar{F}(y_n)}$. Choosing $y_n = \bar{F}^{-1}(\tau_n)$ with $\tau_n = n^{-\nu}$, these conditions are fulfilled provided that $\nu \in \left(\frac{q}{(2\gamma c+1)q-2}, 1\right)$ since both functions g and $\bar{F}$ are assumed to be regularly-varying. Let us also highlight that the above interval is not empty under $(\mathbf{A}_3)$. Finally, Theorem 2.3.1 shows that the estimated direction $\hat{v}(y_n)$ is asymptotically aligned with the true direction β.

2.4 Extension to several directions

The single-index model $(\mathbf{M}_1)$ can be extended to a multi-index setting by considering, for some $K \geq 1$:

$(\mathbf{M}_K)$ $X = \sum_{\ell=1}^{K} g_\ell(Y)\beta^{(\ell)} + \varepsilon$, where X and ε are p-dimensional random vectors, Y is a real random variable, $g_\ell : \mathbb{R} \to \mathbb{R}$ are unknown link functions, $\beta^{(\ell)} \in \mathbb{R}^p$ are unknown orthogonal unit vectors.

Denoting by $\mathcal{Y}$ a set of candidate values for y_n, the following iterative procedure is considered to estimate $\beta^{(1)}, \ldots, \beta^{(K)}$:

1. Initialization: Set $R_i^{(0)} := X_i$ for all $i = 1, \ldots, n$.
2. For $\ell \in \{1, \ldots, p\}$,
   - Estimation of the ℓth direction for all $y_n \in \mathcal{Y}$:

$$\hat{v}^{(\ell)}(y_n) = \hat{\bar{F}}(y_n)\hat{m}_{R^{(\ell-1)}Y}(y_n) - \hat{m}_{R^{(\ell-1)}}(y_n)\hat{m}_Y(y_n).$$

- Computation of the threshold maximizing the conditional correlation:

$$\begin{aligned} y^{(\ell)} &= \arg\max_{y_n \in \mathcal{Y}} \rho\left(\left(R^{(\ell-1)}\right)^t \hat{v}^{(\ell)}(y_n), Y | Y \geq y_n\right) \\ &= \arg\max_{y_n \in \mathcal{Y}} \frac{\operatorname{cov}\left(\left(R^{(\ell-1)}\right)^t \hat{v}^{(\ell)}(y_n), Y | Y \geq y_n\right)}{\sigma\left(\left(R^{(\ell-1)}\right)^t \hat{v}^{(\ell)}(y_n) | Y \geq y_n\right) \sigma\left(Y | Y \geq y_n\right)}, \end{aligned} \tag{2.6}$$

  and recording of the optimal value: $\Xi_\ell = \rho\left(\left(R^{(\ell-1)}\right)^t \hat{v}^{(\ell)}(y^{(\ell)}), Y | Y \geq y^{(\ell)}\right)$.

- Update of the residuals: for all $i = 1, \ldots, n$:

$$R_i^{(\ell)} := R_i^{(\ell-1)} - \frac{\hat{v}^{(\ell)}(y^{(\ell)}) \left(\hat{v}^{(\ell)}(y^{(\ell)})\right)^t}{\|\hat{v}^{(\ell)}(y^{(\ell)})\|^2} R_i^{(\ell-1)}.$$

The idea is the following. At the first iteration, $\hat{v}^{(1)}(y_n)$, $y_n \in \mathcal{Y}$, corresponds to the estimator associated with the single-index model computed by (2.4). Then, the threshold $y^{(1)}$ maximizing w.r.t. $y_n \in \mathcal{Y}$ the correlation between the projected covariate $X^t \hat{v}^{(1)}(y_n)$ and the response variable Y given $Y \geq y_n$ is computed and the maximum correlation $\Xi^{(1)}$ is recorded. From model $(\mathbf{M}_K)$, in view of the orthogonality of the directions, one has $X^t \hat{v}^{(1)}(y^{(1)}) \simeq \|\hat{v}^{(1)}(y^{(1)})\| X^t \beta^{(1)} \simeq \|\hat{v}^{(1)}(y^{(1)})\| g_1(Y)$ and thus the residual

$$R^{(1)} := X - \frac{\hat{v}^{(1)}(y^{(1)}) \left(\hat{v}^{(1)}(y^{(1)})\right)^t}{\|\hat{v}^{(1)}(y^{(1)})\|^2} X \simeq X - g_1(Y)\beta^{(1)} = \sum_{\ell=2}^{K} g_\ell(Y)\beta^{(\ell)} + \varepsilon$$

approximately satisfies the same inverse regression model with $K-1$ directions. It is then natural to iterate the process and estimate $\beta^{(2)}$ from (2.4) computed on the residuals $R^{(1)}$. Moreover, since these residuals are by construction orthogonal to $\hat{v}^{(1)}(y^{(1)})$, one necessarily has $\hat{v}^{(2)}(y^{(2)}) \perp \hat{v}^{(1)}(y^{(1)})$. Thanks to the above orthogonality property, the estimated number of directions can be upper bounded by p. We refer to (Helland, 1990) for a similar result on the original PLS method. The estimation of the number K of directions can be achieved by a visual inspection of the scree plot $\ell \in \{1, \ldots, p\} \mapsto \Xi(\ell)$. The estimated $\hat{K}$ is defined as an elbow in the above graph, which is detected using Cattell's method (Cattell, 1966), see Figure 4 for an illustration on the real data experiment (Section 2.6).

2.5 Validation on simulations

Let us consider a sample of size $n = 1000$ and dimension p from model $(\mathbf{M}_1)$ with a power link function $g(t) = t^c$, $t > 0$, $c \in \{1/4, 1/2, 1, 3/2\}$. The behavior of the EPLS estimator $\hat{v}(y_n)$ is illustrated on this inverse regression model and compared to the estimator

introduced in (Xu et al., 2020) referred to as SIMEXQ (single-index model extreme quantile) in the sequel. SIMEXQ method is an extension of the global single-index quantile regression model developed in (Zhu et al., 2012) where β is estimated by the slope obtained by fitting a misspecified linear quantile regression model to the data. In the SIMEXQ methodology, it is shown that β can be similarly estimated under the weaker assumption of a tail single-index model and a conditional mean linearity assumption. In practice, it is sufficient to narrow the fit of the misspecified linear quantile regression model to the exceedances.

Two heavy-tailed distributions are selected for the response variable Y:

- a Pareto distribution with survival function $\bar{F}(y) = (y/2)^{-5}$, $y \geq 2$,
- a Student t_5 distribution with 5 degrees of freedom.

Let us stress that, in both cases, the tail-index of Y is $\gamma = 1/5$ and does not depend on the covariate. Two dimensions of the covariate are considered:

- $p = 3$ with $\beta = (1, 1, 0)^t/\sqrt{2}$,
- $p = 30$ with $\beta = (1, \ldots, 1, 0, \ldots, 0)^t/\sqrt{15}$.

Each component $\varepsilon^{(j)}$, $j = 1, \ldots, p$ of the error ε is simulated from the $\mathcal{N}(0, \sigma^2)$ distribution and depending on Y using a copula. Two copula models are investigated:

- the Frank copula defined for all $(u_1, u_2) \in [0, 1]^2$ by

$$C_\theta(u_1, u_2) = -\frac{1}{\theta} \log\left(1 + \frac{(\mathrm{e}^{-\theta u_1} - 1)(\mathrm{e}^{-\theta u_2} - 1)}{\mathrm{e}^{-\theta} - 1}\right),$$

  where $\theta \in \mathbb{R}$ is a parameter tuning the dependence between the margins. Frank copula is an Archimedean copula, see (Nelsen, 2007, Section 4.2), able to model the full range of dependence: $\theta \to -\infty$ yields the counter-monotonicity copula, $\theta \to +\infty$ yields the co-monotonicity copula while $\theta = 0$ corresponds to independence. Here, we choose $\theta \in \{0, 10, 20\}$ corresponding to the association measure Kendall's $\tau \in \{0, 0.67, 0.82\}$, see (Nelsen, 2007, Section 5.1).

- the Gaussian copula defined for all $(u_1, u_2) \in [0, 1]^2$ by

$$C_\theta(u_1, u_2) = \Phi_{R_\theta}(\Phi^{-1}(u_1) + \Phi^{-1}(u_2)) \quad \text{with } R_\theta = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix},$$

  where Φ and Φ_{R_θ} are respectively the cumulative distribution functions of the standard univariate Gaussian distribution and bivariate centered Gaussian distribution with covariance matrix R_θ, $\theta \in (-1, 1)$. Here $\theta \to -1$ yields the counter-monotonicity copula,

$\theta \to 1$ yields the co-monotonicity copula while $\theta = 0$ corresponds to independence. We choose $\theta \in \{0, 0.87, 0.96\}$ corresponding to the same Kendall's $\tau \in \{0, 0.67, 0.82\}$ as above.

The standard deviation σ is selected such that the Signal to Noise Ratio (SNR) defined as SNR:$= g(\bar{F}^{-1}(1/n))/\sigma$ is equal to 10. Note that $g(\bar{F}^{-1}(1/n))$ represents the approximate maximum value of g on a n-sample from the distribution with associated survival function $\bar{F}$. Finally, the mean proximity criterion between $\hat{v}(y_n)$ and β is computed on $N = 100$ replications as follows:

$$\mathrm{PC}(y_n) = \frac{1}{N}\sum_{r=1}^{N} \cos^2(\hat{v}(y_n)^{[r]}, \beta) = 1 - \frac{1}{N}\sum_{r=1}^{N} \Delta(\hat{v}(y_n)^{[r]}, \beta), \tag{2.7}$$

where $\Delta(\cdot,\cdot)$ is defined in (2.3) and $\hat{v}(y_n)^{[r]}$ denotes the estimator (2.4) computed on the rth replication. The closer $\mathrm{PC}(y_n)$ is to 1, the better the estimator is. The performance of EPLS and SIMEXQ methods are compared by computing (2.7) in the $4 \times 2 \times 2 \times 2 \times 3 = 96$ considered situations. To this end, denoting by $Y_{1,n} \leq Y_{2,n} \leq \cdots \leq Y_{n,n}$ the order statistics of the sample $(Y_1, \ldots, Y_n)$, the quality measure $\mathrm{PC}(Y_{n-k+1,n})$ is plotted in Figure 2 as a function of the number of exceedances $k \in \{1, \ldots, 200\}$ in the Pareto + Frank case with $p = 3$. Other situations including the Student distribution, the Gaussian copula and a larger dimension $p = 30$ are reported in Figures 6–12 of the Supplementary material.

It first appears that the performance of the EPLS estimator does not depend on the distribution of the response variable. Besides, in small dimension ($p = 3$), the EPLS method yields very accurate results (with $\mathrm{PC} \geq 0.8$) for a wide range of choices of k and whatever the exponent c and the dependence coefficient are. In contrast, the SIMEXQ method appears to be very sensitive to the distribution of Y and to the dependence strength. In this small dimension situation, EPLS outperforms SIMEXQ as soon as independence does not hold. In a high dimension setting ($p = 30$), EPLS still provides very good results (with $\mathrm{PC} \geq 0.8$) for a wide range of choices of k when $c \geq 1$ for all dependence situations. Good results ($\mathrm{PC} \geq 0.6$) can also be obtained when $c = 1/2$ for well-chosen values of k. Here again, the SIMEXQ method is not robust to dependence and is outperformed by EPLS. Finally, it appears that the choice of the number of exceedances k may be a crucial point in difficult situations (high dimension p, high dependence and c small). The selection of k using the procedure described in Section 2.4 is illustrated in the next section on a real dataset.

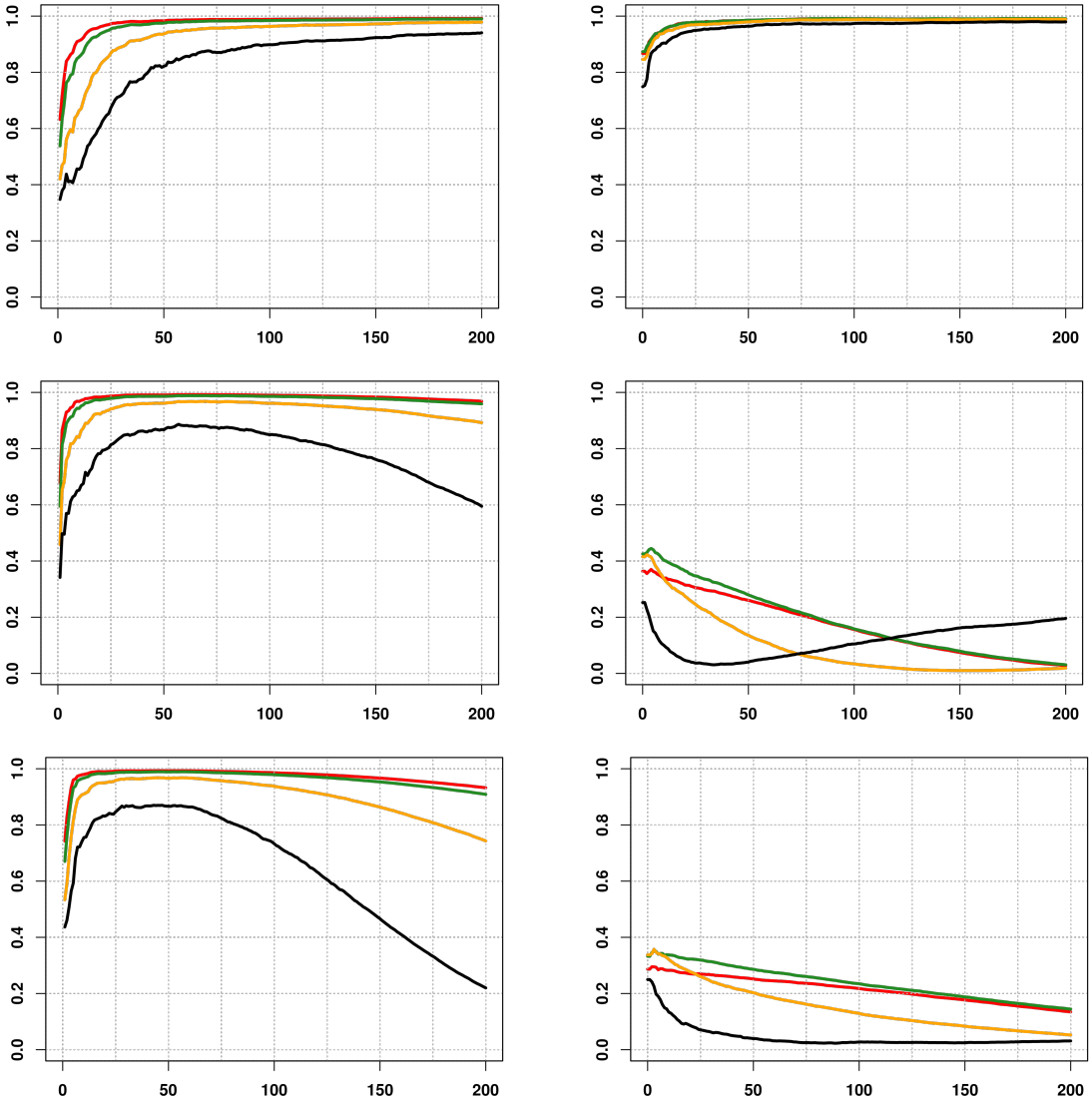


Figure 2 Finite sample behaviour of EPLS $\hat{v}(Y_{n-k+1,n})$ (left) and SIMEXQ (Xu et al., 2020) (right) estimators, on simulated data from a Pareto distribution, Frank copula, dimension $p = 3$. Horizontally: number $k \in \{1, \ldots, 200\}$ of exceedances, vertically: $\mathrm{PC}(Y_{n-k+1,n})$ quality measure. From top to bottom, Frank copula parameter $\theta \in \{0, 10, 20\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

2.6 Application to farm income modelling

Our approach is applied to data extracted from the Farm accountancy data network[1], an annual database of commercial-sized farm holdings. This dataset of size $n = 949$ contains significant accounting and financial information about French farm incomes in 2014. Our goal is to investigate the relationship between low yields and various factors.

The response variable Y is the inverse of the wheat yield (in quintals/hectare), as we focus on the analysis of low yields, and the covariate X includes 12 continuous variables: selling prices (euro/quintal), pesticides, fertilizers, crop insurance purchase, insurance claims, farm subsidies, seeds and seedlings costs, works and services purchase for crops, other insurance premiums, farm income taxes, farmer's personal social security costs (euro/hectare) and average temperature (degree Celsius). We first carry out, in Figure 3, a number of visual checks of whether the heavy-tailed assumption makes sense for these data. First, the histogram of the Y_i (top left panel) gives descriptive evidence that Y has a heavy right tail. The second step consists in drawing a Hill plot:

$$\left(k, \hat{\gamma}_k = \frac{1}{k}\sum_{i=1}^{k} Z_{i,n}\right),\ k = 1, \ldots, 150,$$

where $Z_{i,n} := \log(Y_{n-i+1,n}/Y_{n-k,n})$, $i \in \{1, \ldots, k\}$ denote the log-excesses computed from the consecutive top order statistics. The Hill statistics $\hat{\gamma}_k$ (Hill, 1975) aims at estimating the tail index γ under the semi-parametric model $(\mathbf{A}_1)$. In this situation, for small i, the $Z_{i,n}$ are approximately independent copies of an exponential random variable with mean γ, see for instance (Beirlant et al., 2004, pp.109–110), and $\hat{\gamma}_k$ thus estimates γ using the empirical mean. The resulting graph (top right panel) shows a nice stability of $\hat{\gamma}_k$ as a function of $k \in \{50, \ldots, 150\}$ pointing towards $\gamma \simeq 0.25$. To further confirm that the heavy-tailed framework is appropriate, we draw a quantile-quantile plot of the log-excesses against the quantiles of the unit exponential distribution (bottom panel of Figure 3) for $k = 150$. The relationship appearing in this plot is approximately linear, which constitutes an empirical evidence that the heavy-tail assumption on Y makes sense.

We thus set $\mathcal{Y} := \{Y_{n-k+1,n},\ k = 50, \ldots, 150\}$ and compute $\hat{v}^{(\ell)}(y_n)$ for $y_n \in \mathcal{Y}$ using the procedure described in Section 2.4. The top left panel of Figure 4 displays the conditional correlation (2.6) between the projected residuals and the high values of the response variable Y. All graphs benefit from a stable behaviour with respect to the threshold $y_n \in \mathcal{Y}$, confirming together with the previous Hill-plot and quantile-quantile plot that the associated range of number of exceedances is well-suited to the dataset. It also appears that the first index

[1] A detailed presentation of the database can be found at: http://agreste.agriculture.gouv.fr (in French).

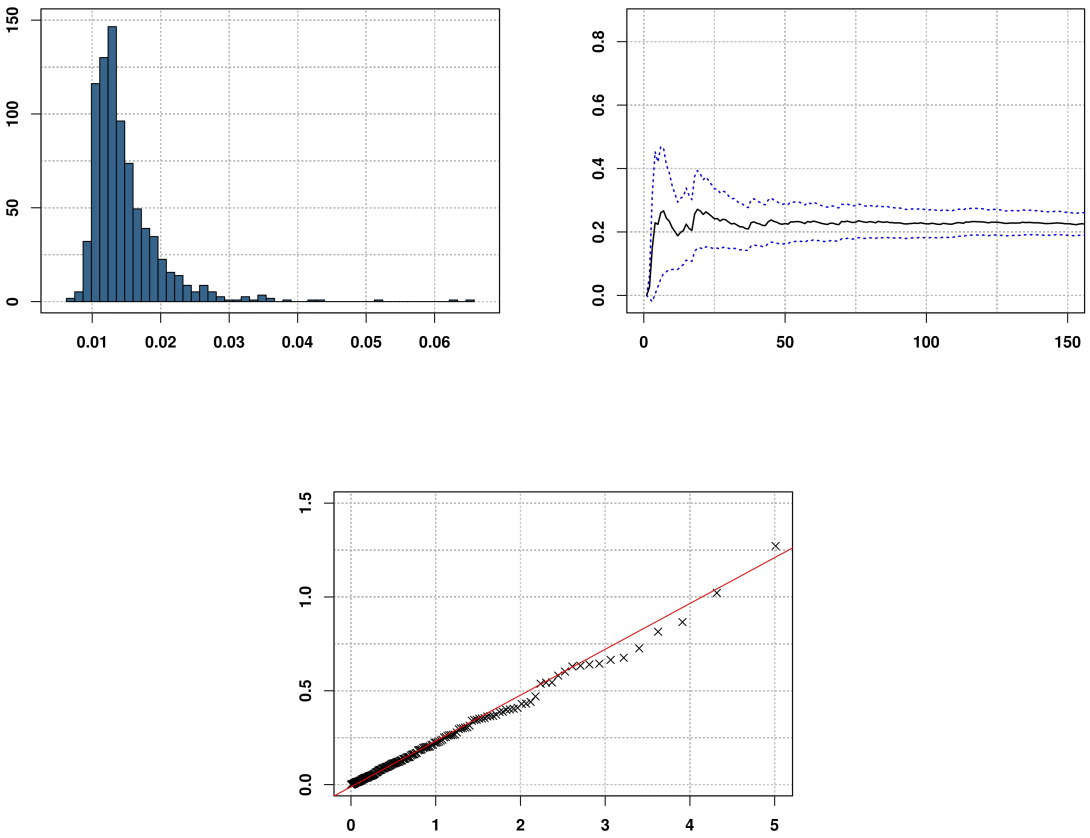


Figure 3 Farm income data. Top left panel: Histogram of the inverse yields Y_i, $i = 1, \ldots, n$. Top right panel: Hill plot (Horizontally: $k \in \{1, \ldots, 150\}$, vertically: Hill estimator $\hat{\gamma}_k$, dashed blue line: empirical 95% confidence interval). Bottom panel: quantile-quantile plot (horizontally: $\log(k/i)$, vertically: $\log(Y_{n-i+1,n}/Y_{n-k,n})$ for $i \in \{1, \ldots, k = 150\}$, red: regression line).

captures about $\Xi_1 = 59\%$ of the correlation (with $k^{(1)} = 97$, blue curve), while the second index fails to represent a significant correlation ($\Xi_2 = 13\%$, $k^{(2)} = 58$, red curve). The second direction is thus discarded in the sequel. The top right panel of Figure 4 represents the conditional correlation between the projected covariate $X^t\hat{v}^{(1)}(y_n)$ on the first direction and each coordinate $X^{(j)}$ of the covariate as a function of $y_n \in \mathcal{Y}$. One can note that small yields are mainly linked to operating costs that can be divided into two categories: agricultural inputs (fertilisers, pesticides, seeds and seedlings, works and services purchase, personal social security costs) and risk management (claims, crop insurance purchase, farm subsidies). This result could be expected since, in 2014, yields were strongly impacted by agricultural inputs, despite mild winter temperatures (which are favourable for wheat crops)[2]. Besides, the effect of crop insurance purchase could be explained by moral hazard that leads insured farmers to use fewer agricultural inputs (Smith and Goodwin, 1996).

The projected scatter plot $(Y_i, X_i^t\,\hat{v}^{(1)}(y^{(1)}))$, $i = 1, \ldots, n$ is displayed in a logarithmic scale for the visualization sake on the top panel of Figure 5 together with two estimations (linear and non-linear) of the conditional mean $\mathbb{E}(X^t\,\hat{v}^{(1)}(y^{(1)})\,|\,Y)$. A positive trend appears for large values of Y in accordance to the inverse regression model ($\mathbf{M}_1$). Let us now focus on the conditional quantiles $\hat{q}(\alpha\,|\,X^t\hat{v}^{(1)}(y^{(1)}))$ computed through a kernel estimator of the conditional survival function (Daouia et al., 2011). The results are reported in the bottom panel of Figure 5 together with the scatter plot $(X_i^t\,\hat{v}^{(1)}(y^{(1)}), Y_i)$, $i = 1, \ldots, n$. The vertical and horizontal axes are represented in a logarithmic scale. Both curves of the conditional quantiles associated with levels $\alpha = 0.15$ (blue line) and $\alpha = 0.05$ (red line) behave in a similar way. The estimated conditional quantiles of inverse yields feature an increasing shape for $\log(X^t\,\hat{v}^{(1)}(y^{(1)})) \leq 9.5$: Lowest yields are (mainly) linked to high operating costs. The interpretation of the results for $\log(X^t\,\hat{v}^{(1)}(y^{(1)})) > 9.5$ is difficult, the estimation being unreliable for large values of the covariate because of data sparsity in this area and boundary effects of kernel estimators, see for instance (Kyung-Joon and Schucany, 1998).

2.7 Discussion

We introduced a new approach EPLS for dimension reduction adapted to distribution tails. It allows to quantify the effect of covariates X on the extreme values of Y in a simple and interpretable way. The asymptotic properties of the estimated direction are established under an inverse single-index model and a heavy tail assumption but without recourse neither to linearity nor to independence assumptions. An extension to the multi-index setting is proposed together with a data-driven method for selecting the number of directions to estimate.

[2] http://agreste.agriculture.sg-ppd.maaf.ate.info/IMG/pdf/conjbilan2014.pdf (in French).

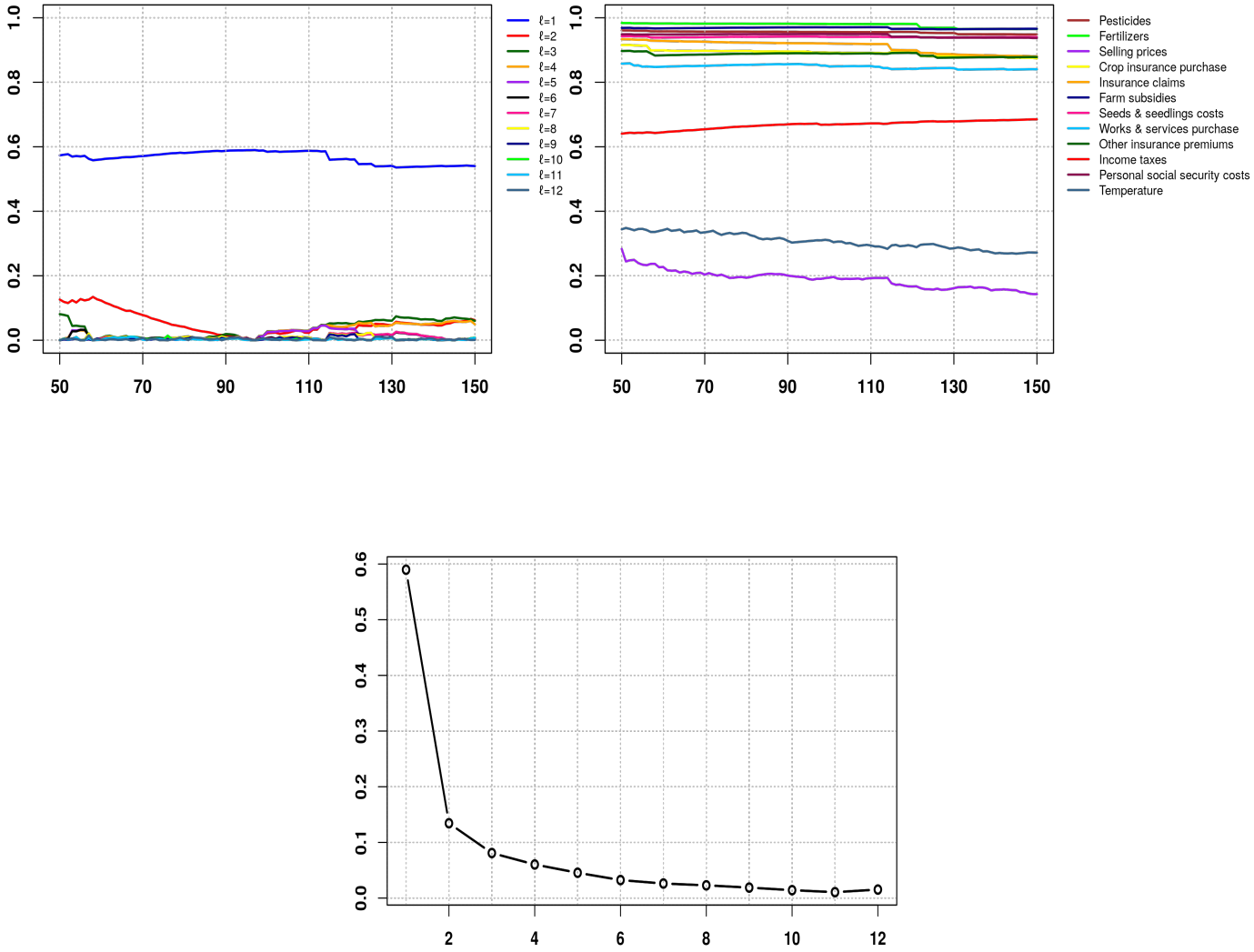


Figure 4 Farm income data. Top left panel: Graph of the estimated conditional correlation function $y_n \in \mathcal{Y} \mapsto \rho((R^{(\ell-1)})^t \hat{v}^{(\ell)}(y_n), Y | Y \geq y_n)$ for $\ell = 1, \ldots, 12$ and top right panel: Graph of the estimated conditional correlation function $y_n \in \mathcal{Y} \mapsto \rho(X^t \hat{v}^{(1)}(y_n), X^{(j)} | Y \geq y)$ for $j = 1, \ldots, 12$ (horizontally: number of exceedances k, vertically: conditional correlation estimated by its empirical counterpart using the threshold $y_n = Y_{n-k+1,n}$). Bottom panel: Scree plot of $\ell \in \{1, \ldots, 12\} \mapsto \Xi_\ell$ (horizontally: iteration ℓ, vertically: maximum correlation Ξ_ℓ.

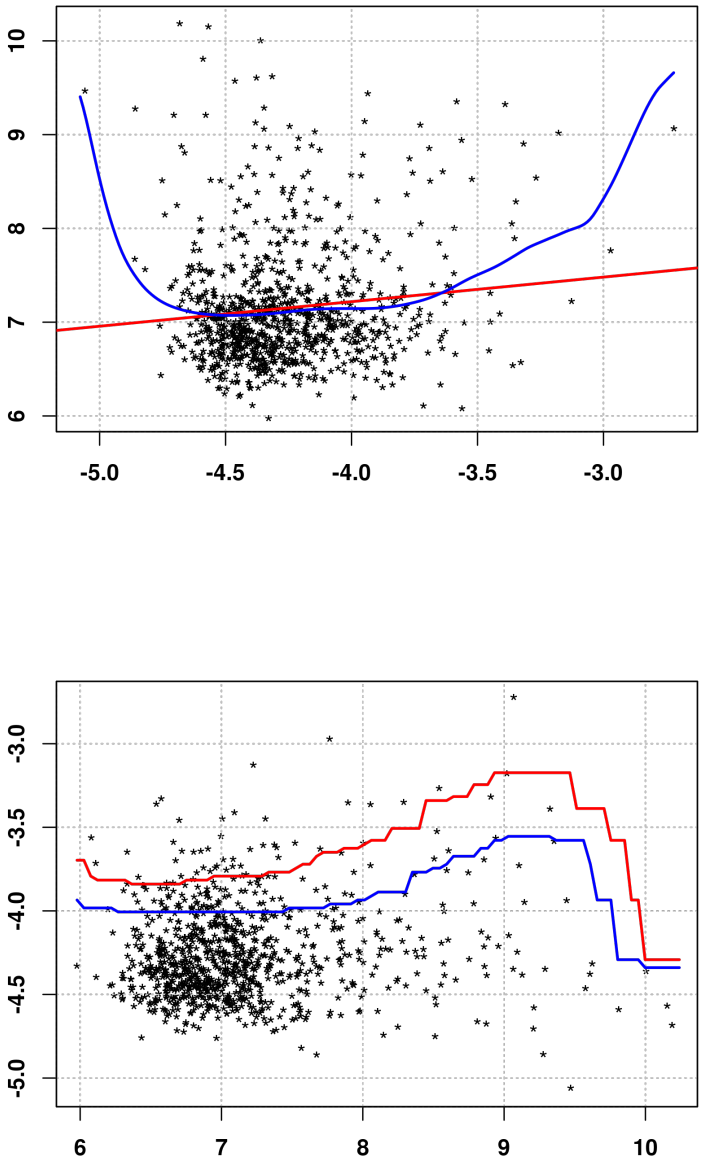


Figure 5 Farm income data. Top: scatter-plot $(Y_i, X_i^t\,\hat{v}^{(1)}(y^{(1)}))$, $i = 1, \ldots, n$ in log scale (horizontally: Y_i, vertically: $X_i^t\,\hat{v}^{(1)}(y^{(1)})$). The regression line (red) and a kernel estimate of the link function (blue) are superimposed. Bottom: scatter-plot $(X_i^t\,\hat{v}^{(1)}(y^{(1)}), Y_i)$, $i = 1, \ldots, n$ in log scale (horizontally: $X_i^t\,\hat{v}^{(1)}(y^{(1)})$, vertically: Y_i). The estimated conditional quantiles $x \mapsto \hat{q}(\alpha|x^t\,\hat{v}(y^{(1)}))$ are superimposed ($\alpha = 0.15$: blue line, $\alpha = 0.05$: red line).

The proposed method can then be used to facilitate the estimation of extreme conditional quantiles or expectiles, thanks to the dimension reduction which circumvents the curse of dimensionality. Quantifying the gain in terms of convergence rates would be of great interest and is the subject of our current work, leveraging the theoretical tools introduced in (Girard et al., 2021).

Acknowledgments The authors would like to thank two anonymous reviewers for their remarks which led to an improved presentation of their results. This work is supported by the French National Research Agency (ANR) in the framework of the Investissements d'Avenir Program (ANR-15-IDEX-02). S. Girard also acknowledges the support of the Chair Stress Test, Risk Management and Financial Steering, led by the French Ecole Polytechnique and its Foundation and sponsored by BNP Paribas.

2.A Appendix: Proofs

Some preliminary lemmas are first provided in 2.A.1. They will reveal useful in the proofs of the main results collected in 2.A.2.

2.A.1 Preliminary results

We begin with a technical tool to compute limits of integrals involving a regularly varying function.

Lemma 2.A.1. *Let $q_2 > 0$ and suppose Y is a random variable satisfying $(\mathbf{A}_1)$ with $\gamma q_2 < 1$. Let $\phi(\cdot)$ be a continuous function on $[1, \infty)$ such that $\phi(t) \to \kappa > 1$ as $t \to \infty$. Then,*

$$\lim_{t\to\infty} \int_1^{+\infty} |x - \phi(t)|^{q_2} \frac{f(tx)}{f(t)} dx = \int_1^{+\infty} |x - \kappa|^{q_2} x^{-\frac{1}{\gamma}-1} dx < \infty.$$

Proof. Potter bounds entail that for all $\epsilon > 0$, $x \geq 1$ and t large enough:

$$0 \leq \frac{f(tx)}{f(t)} \leq (1+\epsilon) x^{-\frac{1}{\gamma}-1+\epsilon}, \tag{2.8}$$

see for example (de Haan and Ferreira, 2007, Proposition B.1.9). Besides, for t large enough, $\kappa/2 \leq \phi(t) \leq 2\kappa$, and therefore:

$$(1-2\kappa)x \leq x - 2\kappa \leq x - \phi(t) \leq x - \kappa/2 \leq x \leq (2\kappa - 1)x.$$

It follows that $|x - \phi(t)| \leq (2\kappa - 1)x$ and, since $q_2 > 0$, we have: $|x - \phi(t)|^{q_2} \leq (2\kappa-1)^{q_2} x^{q_2}$.

Collecting the latter inequality with (2.8) yields:

$$0 \leq |x - \phi(t)|^{q_2} \frac{f(tx)}{f(t)} \leq (2\kappa - 1)^{q_2} x^{q_2 - \frac{1}{\gamma} - 1 + \epsilon}.$$

Recalling that $1/\gamma - q_2 > 0$, one can choose $0 < \epsilon < 1/\gamma - q_2$ such that $x \mapsto x^{q_2 - \frac{1}{\gamma} - 1 + \epsilon}$ is integrable on $[1, \infty)$. Then, Lebesgue's dominated convergence theorem together with the regular variation property of f conclude the proof. □

The next lemma is an adaptation of (Bingham et al., 1987, Proposition 1.5.10) to our setting. It provides an asymptotic equivalent of conditional expectations above a large threshold.

Lemma 2.A.2. *Suppose $\rho \in RV_\mu$ with $\mu \geq 0$ and Y is a random variable satisfying $(\mathbf{A}_1)$ with $\gamma\mu < 1$. Then, as $y \to \infty$,*

$$\mathbb{E}\left[\rho(Y) | Y \geq y\right] \sim \frac{1}{1 - \gamma\mu} \rho(y).$$

Proof. Let us consider

$$\mathbb{E}\left[\rho(Y) | Y \geq y\right] = \frac{1}{\bar{F}(y)} \int_y^{+\infty} \rho(t) f(t) dt.$$

Since $\rho(\cdot)f(\cdot) \in RV_{\mu - 1/\gamma - 1}$, there exists a slowly-varying function L such that $\rho(t)f(t) = t^{\mu - \frac{1}{\gamma} - 1} L(t)$. Then, (Bingham et al., 1987, Proposition 1.5.10) shows that, as $y \to \infty$,

$$\mathbb{E}\left[\rho(Y) | Y \geq y\right] \sim \frac{1}{\bar{F}(y)} \frac{y^{\mu - \frac{1}{\gamma}} L(y)}{1/\gamma - \mu} = \frac{\gamma}{1 - \gamma\mu} \frac{y\rho(y) f(y)}{\bar{F}(y)}.$$

Recalling that $f \in RV_{-1/\gamma - 1}$ and using again (Bingham et al., 1987, Proposition 1.5.10) prove that $\gamma y f(y) \sim \bar{F}(y)$ as $y \to \infty$. Finally, $\mathbb{E}\left[\rho(Y) | Y \geq y\right] \sim \rho(y)/(1 - \gamma\mu)$, as $y \to \infty$ and the conclusion follows. □

The following lemma establishes sufficient conditions such that the moment conditions of Proposition 2.2.1 hold in the context of the inverse regression model $(\mathbf{M}_1)$.

Lemma 2.A.3. *Assume $(\mathbf{M}_1)$, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ and $(\mathbf{A}_3)$ hold with $\gamma(c+1) < 1$. Then, $\mathbb{E}(|Y| \mathbb{1}_{\{Y \geq y\}}) < \infty$, $\mathbb{E}(\|XY\| \mathbb{1}_{\{Y \geq y\}}) < \infty$ and $\mathbb{E}(\|X\| \mathbb{1}_{\{Y \geq y\}}) < \infty$ for all $y \in \mathbb{R}$.*

Proof. Let $y \in \mathbb{R}$. First, let us recall that the existence of $\mathbb{E}(|Y| \mathbb{1}_{\{Y \geq y\}})$ in the Fréchet maximum domain of attraction is a consequence of $\gamma < 1$. Second, the triangle inequality yields:

$$\mathbb{E}(\|X\| \mathbb{1}_{\{Y \geq y\}}) < \mathbb{E}(|g(Y)| \mathbb{1}_{\{Y \geq y\}}) + \mathbb{E}\|\varepsilon\|.$$

Let us note that $\mathbb{E}(|g(Y)| \mathbb{1}_{\{Y \geq y\}}) < \infty$ since $c\gamma < 1$. Besides, for all $q \geq 1/(\gamma c) \geq 1$, Jensen's inequality entails $(\mathbb{E}\|\varepsilon\|)^q \leq \mathbb{E}(\|\varepsilon\|^q) < \infty$ under $(\mathbf{A}_3)$. Hence, $\mathbb{E}\|\varepsilon\| < \infty$ and

$\mathbb{E}(\|X\|\mathbb{1}_{\{Y\geq y\}}) < \infty$. Third,

$$\mathbb{E}(\|XY\|\mathbb{1}_{\{Y\geq y\}}) < \mathbb{E}(|Yg(Y)|\mathbb{1}_{\{Y\geq y\}}) + \mathbb{E}(\|Y\varepsilon\|\mathbb{1}_{\{Y\geq y\}}),$$

and $\mathbb{E}(|Yg(Y)|\mathbb{1}_{\{Y\geq y\}}) < \infty$ in view of $\gamma(c+1) < 1$. Furthermore, Hölder inequality shows that

$$\mathbb{E}(\|Y\varepsilon\|\mathbb{1}_{\{Y\geq y\}}) < [\mathbb{E}(\|\varepsilon\|^q)]^{1/q}[\mathbb{E}(|Y|^{q_2}\mathbb{1}_{\{Y\geq y\}})]^{1/q_2},$$

for all $q_2 \geq 1$ such that $1/q + 1/q_2 = 1$. As already remarked, $\mathbb{E}(\|\varepsilon\|^q) < \infty$ from $(\mathbf{A}_3)$ with $q > 1/(\gamma c)$. Moreover, taking account of condition $\gamma(c+1) < 1$ yields $q > 1/(1-\gamma)$ which is equivalent to $q_2 < 1/\gamma$, and therefore $\mathbb{E}(|Y|^{q_2}\mathbb{1}_{\{Y\geq y\}}) < \infty$ as well. As a conclusion, $\mathbb{E}(\|XY\|\mathbb{1}_{\{Y\geq y\}}) < \infty$ and the result is proved. □

Lemma 2.A.4 provides, in the framework of model $(\mathbf{M}_1)$, an alternative expression of $v(y)$ defined in Proposition 2.2.1.

Lemma 2.A.4. *Assume* $(\mathbf{M}_1)$ *and the assumptions of Proposition 2.1 hold. Then, for all* $y \in \mathbb{R}$, $v(y)$ *can be rewritten as*

$$v(y) = \bar{F}(y)\mathbb{E}[g(Y)\Psi_y(Y)]\,(\beta + \eta(y))\,, \tag{2.9}$$

where

$$\eta(y) := \frac{\mathbb{E}[\varepsilon\Psi_y(Y)]}{\mathbb{E}[g(Y)\Psi_y(Y)]} \text{ and } \Psi_y(Y) := \left(Y - \frac{m_Y(y)}{\bar{F}(y)}\right)\mathbb{1}_{\{Y\geq y\}}. \tag{2.10}$$

Proof. Proposition 2.1 states that $v(y) = \bar{F}(y)\mathbb{E}[XY\mathbb{1}_{\{Y\geq y\}}] - \mathbb{E}[X\mathbb{1}_{\{Y\geq y\}}]m_Y(y)$. Recalling that $X = g(Y)\beta + \varepsilon$ from model $(\mathbf{M}_1)$ yields

$$\begin{aligned}
v(y) &= \bar{F}(y)\mathbb{E}[(g(Y)\beta+\varepsilon)Y\mathbb{1}_{\{Y\geq y\}}] - \mathbb{E}[(g(Y)\beta+\varepsilon)\mathbb{1}_{\{Y\geq y\}}]m_Y(y)\\
&= \beta\mathbb{E}[g(Y)\mathbb{1}_{\{Y\geq y\}}(\bar{F}(y)Y - m_Y(y))] + \mathbb{E}[\varepsilon\mathbb{1}_{\{Y\geq y\}}(\bar{F}(y)Y - m_Y(y))]\\
&= \beta\bar{F}(y)\mathbb{E}[g(Y)\Psi_y(Y)] + \bar{F}(y)\mathbb{E}[\varepsilon\Psi_y(Y)]\\
&= \bar{F}(y)\mathbb{E}[g(Y)\Psi_y(Y)]\left(\beta + \frac{\mathbb{E}[\varepsilon\Psi_y(Y)]}{\mathbb{E}[g(Y)\Psi_y(Y)]}\right).
\end{aligned}$$

Hence the result. □

We first establish a precise control of the moments of the random variable $\Psi_y(Y)$ appearing in the numerator of the remainder term $\eta(y)$, see (2.10) in Lemma 2.A.4.

Lemma 2.A.5. *Let* $q_2 > 0$ *and suppose* Y *is a random variable satisfying* $(\mathbf{A}_1)$ *with* $\gamma q_2 < 1$.

Then,

$$\mathbb{E}\left(|\Psi_y(Y)|^{q_2}\right) \sim \lambda_1(\gamma, q_2) y^{q_2+1} f(y),$$

as $y \to \infty$ *and where* $\lambda_1(\gamma, q_2)$ *is a positive constant.*

Proof. From (2.10), one has

$$\mathbb{E}|\Psi_y(Y)|^{q_2} = \int_y^{+\infty} \left| t - \frac{m_Y(y)}{\bar{F}(y)} \right|^{q_2} f(t)dt = y^{q_2+1} f(y) \int_1^{+\infty} \left| x - \frac{1}{y}\mathbb{E}[Y|Y \geq y] \right|^{q_2} \frac{f(yx)}{f(y)} dx,$$

thanks to the change of variable $x = t/y$ and recalling that $m_Y(y)/\bar{F}(y) = \mathbb{E}[Y|Y \geq y]$. Since $f \in RV_{-1/\gamma-1}$, $\gamma \in (0,1)$, Lemma 2.A.2 applied with $\rho(t) = t$ and $\mu = 1$ entails that

$$\phi(y) := \frac{1}{y}\mathbb{E}[Y|Y \geq y] \to \frac{1}{1-\gamma} =: \kappa \geq 1,$$

as $y \to \infty$. Lemma 2.A.1 then yields, as $y \to \infty$,

$$\int_1^{+\infty} \left| x - \frac{1}{y}\mathbb{E}[Y|Y \geq y] \right|^{q_2} \frac{f(yx)}{f(y)} dx \to \int_1^{+\infty} \left| x - \frac{1}{1-\gamma} \right|^{q_2} x^{-\frac{1}{\gamma}-1} dx =: \lambda_1(\gamma, q_2),$$

As a conclusion, $\mathbb{E}\left(|\Psi_y(Y)|^{q_2}\right) \sim \lambda_1(\gamma, q_2) y^{q_2+1} f(y)$, as $y \to \infty$ and the result is proved. □

Similarly, we provide an asymptotic equivalent of the moments of the random variable $g(Y)\Psi_y(Y)$ appearing in the denominator of the remainder term $\eta(y)$, see (2.10) in Lemma 2.A.4.

Lemma 2.A.6. *Let* Y *be a random variable satisfying* $(\mathbf{A}_1)$ *and* $(\mathbf{A}_2)$ *with* $\gamma(c+1) < 1$. *Then,*

$$\mathbb{E}[g(Y)\Psi_y(Y)] \sim \lambda_2(\gamma, c)\ y g(y) \bar{F}(y),$$

as $y \to \infty$ *and where* $\lambda_2(\gamma, c) := \frac{\gamma^2 c}{(1-\gamma(c+1))(1-\gamma c)(1-\gamma)}$.

Proof. From (2.10), we have:

$$\begin{aligned}\frac{\mathbb{E}[g(Y)\Psi_y(Y)]}{\bar{F}(y)} &= \frac{\mathbb{E}[Y g(Y) \mathbb{1}_{\{Y \geq y\}}]}{\bar{F}(y)} - \frac{\mathbb{E}[Y \mathbb{1}_{\{Y \geq y\}}]}{\bar{F}(y)} \frac{\mathbb{E}[g(Y)\mathbb{1}_{\{Y \geq y\}}]}{\bar{F}(y)} \\ &= \mathbb{E}[Y g(Y)|Y \geq y] - \mathbb{E}[Y|Y \geq y]\mathbb{E}[g(Y)|Y \geq y].\end{aligned}$$

Let us consider $\rho_1(y) = y g(y) \in RV_{c+1}$, $\rho_2(y) = y \in RV_1$ and $\rho_3(y) = g(y) \in RV_c$.

Lemma 2.A.2 entails as $y \to \infty$:

$$\begin{aligned}
\mathbb{E}(\rho_1(Y)|Y \geq y) &\sim \frac{1}{1-\gamma(c+1)}\, yg(y), \\
\mathbb{E}(\rho_2(Y)|Y \geq y) &\sim \frac{1}{1-\gamma}\, y, \\
\mathbb{E}(\rho_3(Y)|Y \geq y) &\sim \frac{1}{1-\gamma c}\, g(y),
\end{aligned}$$

which concludes the proof. □

The next lemma applied successively with $\zeta = 0$ and $\zeta = 1$ yields asymptotic equivalents for these two quantities in the two situations where $\beta_j = 0$ and $\beta_j \neq 0$.

Lemma 2.A.7. *Assume* $(\mathbf{M}_1)$, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ *and* $(\mathbf{A}_3)$ *hold with* $\gamma(c+1) < 1$. *Let* $\zeta \in \{0, 1\}$. *Then,*

(i) For all $j \in \{1, \ldots, p\}$ *such that* $\beta_j \neq 0$, *we have*

$$m_{X_{\cdot,j}Y^\zeta}(y) := \mathbb{E}(X_{\cdot,j} Y^\zeta \mathbb{1}_{\{Y \geq y\}}) \sim \frac{\beta_j}{1-\gamma(c+\zeta)}\, y^\zeta g(y) \bar{F}(y),$$

as $y \to \infty$.

(ii) For all $j \in \{1, \ldots, p\}$ *such that* $\beta_j = 0$, *we have*

$$m_{X_{\cdot,j}Y^\zeta}(y) := \mathbb{E}(\varepsilon_{\cdot,j} Y^\zeta \mathbb{1}_{\{Y \geq y\}}) = O\left(y^\zeta \bar{F}(y)^{1-1/q}\right),$$

as $y \to \infty$.

Remark that, in view of the above lemma, condition $(\mathbf{A}_3)$ implies that, for all $j \in \{1, \ldots, p\}$ such that $\beta_j = 0$, $m_{X_{\cdot,j}Y^\zeta}(y)$ is negligible compared to each $m_{X_{\cdot,\ell}Y^\zeta}(y)$ associated with $\beta_\ell \neq 0$.

Proof. From $(\mathbf{M}_1)$, we have:

$$\begin{aligned}
\mathbb{E}(X_{\cdot,j} Y^\zeta \mathbb{1}_{\{Y \geq y\}}) &= \mathbb{E}(Y^\zeta g(Y) \mathbb{1}_{\{Y \geq y\}})\beta_j + \mathbb{E}(Y^\zeta \varepsilon_{\cdot,j} \mathbb{1}_{\{Y \geq y\}}) \\
&= \mathbb{E}(Y^\zeta g(Y)|Y \geq y)\bar{F}(y)\beta_j + \mathbb{E}(Y^\zeta \varepsilon_{\cdot,j} \mathbb{1}_{\{Y \geq y\}}) \\
&= \frac{\beta_j}{1-\gamma(c+\zeta)}\, y^\zeta g(y)\bar{F}(y)(1+o(1)) + \mathbb{E}(Y^\zeta \varepsilon_{\cdot,j} \mathbb{1}_{\{Y \geq y\}}), \quad (2.11)
\end{aligned}$$

in view of Lemma 2.A.2. Let $q_2 \geq 1$ such that $1/q + 1/q_2 = 1$. Combining Hölder inequality

with $(\mathbf{A}_3)$ yields:

$$\begin{aligned}
\mathbb{E}|\varepsilon_{.,j}Y^{\zeta}\mathbb{1}_{\{Y\geq y\}}| &\leq [\mathbb{E}|\varepsilon_{.,j}|^q]^{1/q}\,[\mathbb{E}(|Y|^{\zeta q_2}\mathbb{1}_{\{Y\geq y\}})]^{1/q_2} \\
&= [\mathbb{E}|\varepsilon_{.,j}|^q]^{1/q}\,[\mathbb{E}(|Y|^{\zeta q_2}|Y\geq y)]^{1/q_2}\bar{F}(y)^{1/q_2} \\
&= [\mathbb{E}|\varepsilon_{.,j}|^q]^{1/q}\left(\frac{|y|^{\zeta q_2}}{1-q_2\zeta\gamma}\right)^{1/q_2}\bar{F}(y)^{1/q_2}(1+o(1)) \\
&= O\left(y^{\zeta}\bar{F}(y)^{1-1/q}\right), \qquad (2.12)
\end{aligned}$$

as $y\to\infty$, according to Lemma 2.A.2 and since $q_2\zeta\gamma<1$. This proves (ii) when $\beta_j=0$. Focusing on the situation where $\beta_j\neq 0$, we have from (2.11) and (2.12),

$$\begin{aligned}
\mathbb{E}(X_{.,j}Y^{\zeta}\mathbb{1}_{\{Y\geq y\}}) &= \frac{\beta_j}{1-\gamma(c+\zeta)}\,y^{\zeta}g(y)\bar{F}(y)(1+o(1))+O(y^{\zeta}\bar{F}(y)^{1-1/q}) \\
&= \frac{\beta_j}{1-\gamma(c+\zeta)}\,y^{\zeta}g(y)\bar{F}(y)\left(1+o(1)+O\left(\frac{1}{\bar{F}(y)^{1/q}g(y)}\right)\right).
\end{aligned}$$

As a consequence of $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$, $\bar{F}(\cdot)^{1/q}g(\cdot)$ is a regularly varying function with index $c-1/(q\gamma)>0$. Therefore, $\bar{F}(y)^{1/q}g(y)\to\infty$ when $y\to\infty$ and (i) is proved. □

The last lemma proves that the noise term ε does not contribute to the asymptotic distribution of the estimators.

Lemma 2.A.8. *Assume* $(\mathbf{M}_1)$, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ *and* $(\mathbf{A}_3)$ *hold with* $2\gamma(c+1)<1$. *For all* $\zeta\in\{0,1\}$ *let*

$$T_{.,n}^{(\zeta)}=\sqrt{\bar{F}(y_n)}\left(\sum_{\beta_j\neq 0}\frac{\alpha_j^{(\zeta)}\varepsilon_{.,j}}{m_{X_{.,j}Y^{\zeta}}(y_n)}\right)Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}},$$

where $\alpha_j^{(\zeta)}\in\mathbb{R}$ *for all* $j=1,\ldots,p$. *Then,* $\chi_n^{(\zeta)}:=\frac{1}{\sqrt{n}}\sum_{i=1}^n(T_{i,n}^{(\zeta)}-\mathbb{E}(T_{i,n}^{(\zeta)}))\xrightarrow{P}0$.

Proof. Clearly, $\chi_n^{(\zeta)}$ is centered by definition. Let us consider its variance:

$$\begin{aligned}
\mathrm{var}(\chi_n^{(\zeta)}) &= \mathrm{var}(T_{.,n}^{(\zeta)}) \\
&= \sum_{\beta_j\neq 0}\sum_{\beta_\ell\neq 0}\frac{\alpha_j^{(\zeta)}\alpha_\ell^{(\zeta)}\bar{F}(y_n)}{m_{X_{.,j}Y^{\zeta}}(y_n)m_{X_{.,\ell}Y^{\zeta}}(y_n)}\mathrm{cov}(\varepsilon_{.,j}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}},\varepsilon_{.,\ell}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}}) \\
&\sim \frac{(1-\gamma(c+\zeta))^2}{y_n^{2\zeta}g(y_n)^2\bar{F}(y_n)}\sum_{\beta_j\neq 0}\sum_{\beta_l\neq 0}\frac{\alpha_j^{(\zeta)}\alpha_\ell^{(\zeta)}}{\beta_j\beta_\ell}\mathrm{cov}(\varepsilon_{.,j}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}},\varepsilon_{.,\ell}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}}),
\end{aligned}$$

as $n\to\infty$, from Lemma 2.A.7(i). The covariance can be expanded as

$$\mathrm{cov}(\varepsilon_{.,j}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}},\varepsilon_{.,\ell}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}})=\mathbb{E}(\varepsilon_{.,j}\varepsilon_{.,\ell}Y^{2\zeta}\mathbb{1}_{\{Y\geq y_n\}})-\mathbb{E}(\varepsilon_{.,j}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}})\mathbb{E}(\varepsilon_{.,\ell}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}}).$$

The first term is bounded using Hölder inequality, applied for all $q_3\geq 1$ such that $2/q+1/q_3=$

1:

$$\begin{aligned}\mathbb{E}(\varepsilon_{.,j}\varepsilon_{.,\ell}Y^{2\zeta}\mathbb{1}_{\{Y\geq y_n\}}) &\leq [\mathbb{E}|\varepsilon_{.,j}|^q]^{1/q}[\mathbb{E}|\varepsilon_{.,\ell}|^q]^{1/q}[\mathbb{E}|Y^{2\zeta}\mathbb{1}_{\{Y\geq y_n\}}|^{q_3}]^{1/q_3}\\ &\leq [\mathbb{E}|\varepsilon_{.,j}|^q]^{1/q}[\mathbb{E}|\varepsilon_{.,\ell}|^q]^{1/q}[\mathbb{E}|Y^{2\zeta q_3}|Y\geq y_n|]^{1/q_3}\bar{F}(y_n)^{1/q_3}\\ &= [\mathbb{E}|\varepsilon_{.,j}|^q]^{1/q}[\mathbb{E}|\varepsilon_{.,\ell}|^q]^{1/q}\left(\frac{y_n^{2\zeta q_3}}{1-2\zeta\gamma p'}\right)^{1/q_3}\bar{F}(y_n)^{1/q_3}(1+o(1))\\ &= O\left(y_n^{2\zeta}\bar{F}(y_n)^{1-2/q}\right), \end{aligned} \tag{2.13}$$

in view of Lemma 2.A.2 and $(\mathbf{A}_3)$. Indeed, condition $2\gamma(c+1)<1$ is equivalent to $\gamma c < 1/2-\gamma$. Besides, from $(\mathbf{A}_3)$, $q>1/(\gamma c)$ and thus $q>2/(1-2\gamma)$ leading to $2\gamma q_3<1$. The second term is controlled with Lemma 2.A.7(ii) and is negligible compared to the first one:

$$|\mathbb{E}(\varepsilon_{.,j}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}})\mathbb{E}(\varepsilon_{.,\ell}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}})| = O(y_n^{2\zeta}\bar{F}(y_n)^{2-2/q}) = o\left(y_n^{2\zeta}\bar{F}(y_n)^{1-2/q}\right). \tag{2.14}$$

Taking account of (2.13) and (2.14) yields

$$\operatorname{cov}(\varepsilon_{.,j}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}},\varepsilon_{.,\ell}Y^{\zeta}\mathbb{1}_{\{Y\geq y_n\}}) = O\left(y_n^{2\zeta}\bar{F}(y_n)^{1-2/q}\right), \tag{2.15}$$

and therefore

$$\operatorname{var}(\chi_n^{(\zeta)}) = O\left(\frac{1}{\bar{F}(y_n)^{2/q}g(y_n)^2}\right)\to 0,$$

as $n\to\infty$ since $\bar{F}(\cdot)^{2/q}g(\cdot)^2$ is regularly-varying with index $2(c-1/(q\gamma))>0$. The conclusion follows. □

2.A.2 Proofs of main results

Proof of Proposition 2.2.1. Let us rewrite the optimization problem as

$$\begin{aligned} w(y) &= \arg\max_{\|w\|=1}\operatorname{cov}(w^tX,Y|Y\geq y)\\ &= \arg\max_{\|w\|=1}\frac{\mathbb{E}(w^tXY\mathbb{1}_{\{Y\geq y\}})}{\bar{F}(y)}-\frac{\mathbb{E}(w^tX\mathbb{1}_{\{Y\geq y\}})\mathbb{E}(Y\mathbb{1}_{\{Y\geq y\}})}{\bar{F}(y)^2}\\ &= \arg\max_{\|w\|=1}\bar{F}(y)w^tm_{XY}(y)-w^tm_X(y)m_Y(y)\\ &= \arg\max_{\|w\|=1}w^tv(y). \end{aligned}$$

This constrained optimization problem is solved using Lagrange multipliers method. Introducing

$$\mathcal{L}(w,\lambda) = w^tv(y)-\frac{\lambda}{2}(||w||^2-1),\quad \lambda\in\mathbb{R},$$

and setting the partial derivatives to zero yield $\lambda = \|v(y)\|$ and $w = v(y)/\|v(y)\|$. □

Proof of Proposition 2.2.2. From Lemma 2.A.3, $\mathbb{E}(|Y|\mathbb{1}_{\{Y\geq y\}})$, $\mathbb{E}(\|XY\|\mathbb{1}_{\{Y\geq y\}})$ and $\mathbb{E}(\|X\|\mathbb{1}_{\{Y\geq y\}})$ exist for all $y \in \mathbb{R}$. We may then apply Lemma 2.A.4 to get:

$$\cos(w(y), \beta) = w(y)^t\beta = \text{sign}(\mathbb{E}[g(Y)\Psi_y(Y)])\frac{1+\beta^t\eta(y)}{\|\beta+\eta(y)\|}, \tag{2.16}$$

with $\eta(y) = \mathbb{E}[\varepsilon\Psi_y(Y)]/\mathbb{E}[g(Y)\Psi_y(Y)]$, see (2.10), and where $\text{sign}(u) = 1$ is $u \geq 0$ and $\text{sign}(u) = -1$ otherwise. Straightforward calculations yield

$$\cos^2(w(y), \beta) - 1 = \frac{(\beta^t\eta(y))^2 - \|\eta(y)\|^2}{\|\beta+\eta(y)\|^2},$$

and therefore it is sufficient to prove that $\|\eta(y)\| \to 0$ as $y \to \infty$ to get

$$1 - \cos^2(w(y), \beta) = O(\|\eta(y)\|^2), \tag{2.17}$$

as $y \to \infty$. Under assumption $(\mathbf{A}_3)$, there exists $q > 1/(\gamma c)$ such that $\mathbb{E}\|\varepsilon\|^q < \infty$. Hölder inequality thus yields

$$\|\mathbb{E}[\varepsilon\Psi_y(Y)]\| \leq [\mathbb{E}\|\varepsilon\|^q]^{1/q}[\mathbb{E}|\Psi_y(Y)|^{q_2}]^{1/q_2},$$

for all $q_2 \geq 1$ such that $1/q + 1/q_2 = 1$. As a consequence, $\gamma(c+1) < 1$ and $\gamma c > 1/q$ imply $\gamma q_2 < 1$ and then Lemma 2.A.5 shows that $\mathbb{E}|\Psi_y(Y)|^{q_2} \sim \lambda_1(\gamma, q_2)y^{q_2+1}f(y)$ as $y \to \infty$, with $\lambda_1(\gamma, q_2) > 0$. Therefore,

$$\|\mathbb{E}[\varepsilon\Psi_y(Y)]\| \leq [\mathbb{E}\|\varepsilon\|^q]^{1/q}(\lambda_1(\gamma, q_2))^{1/q_2}y^{1+1/q_2}f(y)^{1/q_2}. \tag{2.18}$$

and Lemma 2.A.6 shows that, as $y \to \infty$,

$$\mathbb{E}[g(Y)\Psi_y(Y)] \sim \lambda_2(\gamma, c)yg(y)\bar{F}(y), \tag{2.19}$$

with $\lambda_2(\gamma, c) > 0$. Collecting (2.18) and (2.19) thus yields:

$$\|\eta(y)\| \leq \frac{[\mathbb{E}\|\varepsilon\|^q]^{1/q}(\lambda_1(\gamma, q_2))^{1/q_2}}{\lambda_2(\gamma, c)}\frac{(yf(y))^{1/q_2}}{g(y)\bar{F}(y)}(1+o(1)) = O\left(\frac{1}{g(y)\bar{F}^{1/q}(y)}\right). \tag{2.20}$$

Under assumptions $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$, $\bar{F} \in RV_{-1/\gamma}$ and $g \in RV_c$ so that $y \mapsto g(y)\bar{F}^{1/q}(y)$ is also regularly varying with index $c - 1/(q\gamma) > 0$. As a consequence, $\|\eta(y)\| \to 0$ as $y \to \infty$ and the first part of the result is proved. Second, (2.19) shows that, for y large enough, $\mathbb{E}[g(Y)\Psi_y(Y)] > 0$. Consequently, $\cos(w(y), \beta) \geq 0$ eventually in view of (2.16), and

then (2.17) entails that $\cos(w(y),\beta) \to 1$ as $y \to \infty$ leading to

$$\|w(y)-\beta\| = \sqrt{2(1-\cos(w(y),\beta))} \sim \sqrt{1-\cos^2(w(y),\beta)} = O(\|\eta(y)\|) = O\left(\frac{1}{g(y)\bar{F}^{1/q}(y)}\right),$$

as $y \to \infty$, from (2.20). The result is proved. □

Proof of Proposition 2.3.1. To establish the joint asymptotic normality of the $2(d+1)-$random vector Λ_n, we shall prove that any non-zero linear combination of its components is asymptotically Gaussian.

Set $\alpha = (\alpha_1, \alpha_2, \alpha_{3,1}, \ldots, \alpha_{3,d}, \alpha_{4,1}, \ldots, \alpha_{4,d})^t \in \mathbb{R}^{2(d+1)}$ and let us investigate the asymptotic distribution of χ_n defined as follows:

$$\begin{aligned}\chi_n &= \sqrt{n\bar{F}(y_n)}\left\{\alpha_1\left(\frac{\hat{\bar{F}}(y_n)}{\bar{F}(y_n)}-1\right)+\alpha_2\left(\frac{\hat{m}_Y(y_n)}{m_Y(y_n)}-1\right)\right\}\\ &+ \sqrt{n\bar{F}(y_n)}\left\{\sum_{j=1}^{d}\alpha_{3,j}\left(\frac{\hat{m}_{X_{\cdot,j}}(y_n)}{m_{X_{\cdot,j}}(y_n)}-1\right)+\alpha_{4,j}\left(\frac{\hat{m}_{X_{\cdot,j}Y}(y_n)}{m_{X_{\cdot,j}Y}(y_n)}-1\right)\right\},\end{aligned}$$

and which can be rewritten as $\chi_n = \sum_{i=1}^n \chi_{i,n} := \sum_{i=1}^n (Z_{i,n} - \mathbb{E}(Z_{i,n}))$, where

$$Z_{i,n} = \sqrt{\frac{\bar{F}(y_n)}{n}}\left(\frac{\alpha_1}{\bar{F}(y_n)}+\frac{\alpha_2 Y_i}{m_Y(y_n)}+\sum_{j=1}^{d}\frac{\alpha_{3,j}X_{i,j}}{m_{X_{\cdot,j}}(y_n)}+\sum_{j=1}^{d}\frac{\alpha_{4,j}X_{i,j}Y_i}{m_{X_{\cdot,j}Y}(y_n)}\right)\mathbb{1}_{\{Y_i\geq y_n\}}.$$

Under model $(\mathbf{M}_1)$, $X_{i,j} = g(Y_i)\beta_j + \varepsilon_{i,j}$, for $j \in \{1,\ldots,d\}$ and $i \in \{1,\ldots,n\}$, we get the following decomposition:

$$Z_{i,n} \overset{d}{=} \frac{1}{\sqrt{n}}(T_{1,i,n}+T_{2,i,n}+T_{3,i,n}+T'_{3,i,n}+T_{4,i,n}+T'_{4,i,n}),$$

where

$$T_{1,i,n} = \frac{\alpha_1}{\sqrt{\bar{F}(y_n)}}\mathbb{1}_{\{Y_i\geq y_n\}}, \qquad T_{2,i,n} = \sqrt{\bar{F}(y_n)}\frac{\alpha_2}{m_Y(y_n)}Y_i\mathbb{1}_{\{Y_i\geq y_n\}},$$

$$T_{3,i,n} = \sqrt{\bar{F}(y_n)}\left(\sum_{j=1}^{d}\frac{\alpha_{3,j}\beta_j}{m_{X_{\cdot,j}}(y_n)}\right)g(Y_i)\mathbb{1}_{\{Y_i\geq y_n\}}, \qquad T'_{3,i,n} = \sqrt{\bar{F}(y_n)}\left(\sum_{j=1}^{d}\frac{\alpha_{3,j}\varepsilon_{i,j}}{m_{X_{\cdot,j}}(y_n)}\right)\mathbb{1}_{\{Y_i\geq y_n\}},$$

$$T_{4,i,n} = \sqrt{\bar{F}(y_n)}\left(\sum_{j=1}^{d}\frac{\alpha_{4,j}\beta_j}{m_{X_{\cdot,j}Y}(y_n)}\right)Y_i g(Y_i)\mathbb{1}_{\{Y_i\geq y_n\}}, \quad T'_{4,i,n} = \sqrt{\bar{F}(y_n)}\left(\sum_{j=1}^{d}\frac{\alpha_{4,j}\varepsilon_{i,j}}{m_{X_{\cdot,j}Y}(y_n)}\right)Y_i\mathbb{1}_{\{Y_i\geq y_n\}}.$$

Substituting yields $\chi_n = \chi'_{0,n} + \chi'_{3,n} + \chi'_{4,n}$ where

$$\begin{aligned}
\chi'_{0,n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{\ell=1}^{4} (T_{\ell,i,n} - \mathbb{E}(T_{\ell,i,n})), \\
\chi'_{3,n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (T'_{3,i,n} - \mathbb{E}(T'_{3,i,n})), \\
\chi'_{4,n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (T'_{4,i,n} - \mathbb{E}(T'_{4,i,n})).
\end{aligned}$$

Lemma 2.A.8 shows that both $\chi'_{3,n}$ and $\chi'_{4,n}$ converge in probability to zero and we therefore focus on the limiting distribution of $\chi'_{0,n}$. Clearly, $\mathbb{E}(\chi'_{0,n}) = 0$. Turning to the variance of $\chi'_{0,n}$ and, since we deal with independent and identically distributed random variables, one has

$$\operatorname{var}(\chi'_{0,n}) = \operatorname{var}\left(\sum_{\ell=1}^{4} T_{\ell,i,n}\right) = \sum_{\ell=1}^{4} \operatorname{var}\left(T_{\ell,i,n}\right) + 2 \sum_{1 \le \ell < m \le 4} \operatorname{cov}\left(T_{\ell,i,n}, T_{m,i,n}\right),$$

for all $i \in \{1, \ldots, n\}$. As a preliminary result, the following asymptotic equivalent holds for all $(\zeta, \omega) \in \{0, 1, 2\}^2$:

$$\mathbb{E}(Y^{\zeta} g(Y)^{\omega} \mathbb{1}_{\{Y \ge y_n\}}) \sim \frac{1}{1 - \gamma(\omega c + \zeta)} \, y_n^{\zeta} g(y_n)^{\omega} \bar{F}(y_n), \tag{2.21}$$

as $n \to \infty$, in view of Lemma 2.A.2 and assumption $2\gamma(c+1) < 1$. Moreover, Lemma 2.A.2 and Lemma 2.A.7(i) yield the following asymptotic equivalents:

$$m_Y(y_n) \sim \frac{1}{1-\gamma} y_n \bar{F}(y_n), \tag{2.22}$$

$$m_{X_{\cdot,j}}(y_n) \sim \frac{\beta_j}{1-\gamma c} g(y_n) \bar{F}(y_n), \tag{2.23}$$

$$m_{X_{\cdot,j}Y}(y_n) \sim \frac{\beta_j}{1-\gamma(c+1)} y_n g(y_n) \bar{F}(y_n), \tag{2.24}$$

which will reveal useful in evaluation of variances and covariances below. We shall also use the notation $\langle \alpha_\ell \rangle = \sum_{j=1}^{d} \alpha_{\ell,j}$ for $\ell \in \{3, 4\}$. A straightforward calculation yields:

$$\operatorname{var}(T_{1,i,n}) = \alpha_1^2 (1 - \bar{F}(y_n)) \to \alpha_1^2 \text{ as } n \to \infty.$$

Combining (2.21) and (2.22) leads to:

$$\operatorname{var}(T_{2,i,n}) = \alpha_2^2 \bar{F}(y_n) \left(\frac{\mathbb{E}(Y^2 \mathbb{1}_{\{Y \ge y_n\}})}{m_Y^2(y_n)} - 1 \right) \to b_{22} \, \alpha_2^2 \text{ as } n \to \infty.$$

Similarly, and taking account of Lemma 2.A.2 and Lemma 2.A.7(i), one has:

$$\begin{aligned}
\mathrm{var}(T_{3,i,n}) &= \left(\sqrt{\bar{F}(y_n)}\sum_{j=1}^{d}\frac{\alpha_{3,j}\beta_j}{m_{X_{\cdot,j}}(y_n)}\right)^2\left[\mathbb{E}(g(Y)^2\mathbb{1}_{\{Y\geq y_n\}})-\mathbb{E}(g(Y)\mathbb{1}_{\{Y\geq y_n\}})^2\right]\\
&\sim \frac{(1-\gamma c)^2\langle\alpha_3\rangle^2}{g(y_n)^2\bar{F}(y_n)}\left(\frac{g(y_n)^2\bar{F}(y_n)}{1-2\gamma c}-\frac{g(y_n)^2\bar{F}(y_n)^2}{(1-\gamma c)^2}(1+o(1))\right)\\
&\to b_{33}\langle\alpha_3\rangle^2 \text{ as } n\to\infty,
\end{aligned}$$

$$\begin{aligned}
\mathrm{var}(T_{4,i,n}) &= \left(\sqrt{\bar{F}(y_n)}\sum_{j=1}^{d}\frac{\alpha_{4,j}\beta_j}{m_{X_{\cdot,j}Y}(y_n)}\right)^2\left[\mathbb{E}(Y^2g(Y)^2\mathbb{1}_{\{Y\geq y_n\}})-\mathbb{E}(Yg(Y)\mathbb{1}_{\{Y\geq y_n\}})^2\right]\\
&\sim \frac{(1-\gamma(c+1))^2\langle\alpha_4\rangle^2}{y_n^2g(y_n)^2\bar{F}(y_n)}\left(\frac{(y_ng(y_n))^2\bar{F}(y_n)}{1-2\gamma(c+1)}-\frac{(y_ng(y_n)\bar{F}(y_n))^2}{(1-\gamma(c+1))^2}(1+o(1))\right)\\
&\to b_{44}\langle\alpha_4\rangle^2 \text{ as } n\to\infty.
\end{aligned}$$

The covariances are evaluated in a similar way. First, terms involving $T_{1,i,n}$ can be readily calculated:

$$\begin{aligned}
\mathrm{cov}(T_{1,i,n},T_{2,i,n}) &= \mathrm{cov}\left(\frac{\alpha_1}{\sqrt{\bar{F}(y_n)}}\mathbb{1}_{\{Y\geq y_n\}},\sqrt{\bar{F}(y_n)}\frac{\alpha_2}{m_Y(y_n)}Y\mathbb{1}_{\{Y\geq y_n\}}\right)\\
&= \alpha_1\alpha_2(1-\bar{F}(y_n))\\
&\to \alpha_1\alpha_2 \text{ as } n\to\infty,
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}(T_{1,i,n},T_{3,i,n}) &= \mathrm{cov}\left(\frac{\alpha_1}{\sqrt{\bar{F}(y_n)}}\mathbb{1}_{\{Y\geq y_n\}},\sqrt{\bar{F}(y_n)}\left(\sum_{j=1}^{d}\frac{\alpha_{3,j}\beta_j}{m_{X_{\cdot,j}}(y_n)}\right)g(Y)\mathbb{1}_{\{Y\geq y_n\}}\right)\\
&\sim \frac{(1-\gamma c)\alpha_1\langle\alpha_3\rangle}{g(y_n)\bar{F}(y_n)}\mathbb{E}(g(Y)\mathbb{1}_{\{Y\geq y_n\}})(1-\bar{F}(y_n))\\
&\to \alpha_1\langle\alpha_3\rangle \text{ as } n\to\infty,
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}(T_{1,i,n},T_{4,i,n}) &= \mathrm{cov}\left(\frac{\alpha_1}{\sqrt{\bar{F}(y_n)}}\mathbb{1}_{\{Y\geq y_n\}},\left(\sqrt{\bar{F}(y_n)}\sum_{j=1}^{d}\frac{\alpha_{4,j}\beta_j}{m_{X_{\cdot,j}Y}(y_n)}\right)Yg(Y)\mathbb{1}_{\{Y\geq y_n\}}\right)\\
&\sim \frac{(1-\gamma(c+1))\alpha_1\langle\alpha_4\rangle}{y_ng(y_n)\bar{F}(y_n)}\mathbb{E}(Yg(Y)\mathbb{1}_{\{Y\geq y_n\}})(1-\bar{F}(y_n))\\
&\to \alpha_1\langle\alpha_4\rangle \text{ as } n\to\infty.
\end{aligned}$$

Second, the remaining terms require repeated uses of Lemma 2.A.7(i):

$$\begin{aligned}
\operatorname{cov}(T_{2,i,n}, T_{3,i,n}) &= \operatorname{cov}\left(\sqrt{\bar{F}(y_n)}\frac{\alpha_2}{m_Y(y_n)}Y\mathbb{1}_{\{Y\geq y_n\}}, \left(\sqrt{\bar{F}(y_n)}\sum_{j=1}^{d}\frac{\alpha_{3,j}\beta_j}{m_{X_{\cdot,j}}(y_n)}\right)g(Y)\mathbb{1}_{\{Y\geq y_n\}}\right)\\
&\sim \frac{(1-\gamma)(1-\gamma c)\alpha_2\langle\alpha_3\rangle}{y_n g(y_n)\bar{F}(y_n)}\left[\mathbb{E}(Yg(Y)\mathbb{1}_{\{Y\geq y_n\}}) - \mathbb{E}(Y\mathbb{1}_{\{Y\geq y_n\}})\mathbb{E}(g(Y)\mathbb{1}_{\{Y\geq y_n\}})\right]\\
&\sim \frac{(1-\gamma)(1-\gamma c)\alpha_2\langle\alpha_3\rangle}{y_n g(y_n)\bar{F}(y_n)}\left(\frac{y_n g(y_n)\bar{F}(y_n)}{1-\gamma(c+1)} - \frac{y_n g(y_n)\bar{F}(y_n)^2}{(1-\gamma)(1-\gamma c)}(1+o(1))\right)\\
&\to b_{23}\,\alpha_2\langle\alpha_3\rangle \text{ as } n\to\infty,
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(T_{2,i,n}, T_{4,i,n}) &= \operatorname{cov}\left(\sqrt{\bar{F}(y_n)}\frac{\alpha_2}{m_Y(y_n)}Y\mathbb{1}_{\{Y\geq y_n\}}, \left(\sqrt{\bar{F}(y_n)}\sum_{j=1}^{d}\frac{\alpha_{4,j}\beta_j}{m_{X_{\cdot,j}Y}(y_n)}\right)Yg(Y)\mathbb{1}_{\{Y\geq y_n\}}\right)\\
&\sim \frac{(1-\gamma(c+1))(1-\gamma)\alpha_2\langle\alpha_4\rangle}{y_n^2 g(y_n)\bar{F}(y_n)}\\
&\times \left[\mathbb{E}(Y^2 g(Y)\mathbb{1}_{\{Y\geq y_n\}}) - \mathbb{E}(Y\mathbb{1}_{\{Y\geq y_n\}})\mathbb{E}(Yg(Y)\mathbb{1}_{\{Y\geq y_n\}})\right]\\
&\sim \frac{(1-\gamma(c+1))(1-\gamma)\alpha_2\langle\alpha_4\rangle}{y_n^2 g(y_n)\bar{F}(y_n)}\\
&\times \left(\frac{y_n^2 g(y_n)\bar{F}(y_n)}{1-\gamma(c+2)} - \frac{y_n^2 g(y_n)\bar{F}(y_n)^2}{(1-\gamma)(1-\gamma(c+1))}(1+o(1))\right)\\
&\to b_{24}\,\alpha_2\langle\alpha_4\rangle \text{ as } n\to\infty,
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(T_{3,i,n}, T_{4,i,n}) &= \operatorname{cov}\left(\sqrt{\bar{F}(y_n)}\sum_{j=1}^{d}\frac{\alpha_{3,j}\beta_j}{m_{X_{\cdot,j}}(y_n)}g(Y)\mathbb{1}_{\{Y\geq y_n\}}, \sqrt{\bar{F}(y_n)}\sum_{j=1}^{d}\frac{\alpha_{4,j}\beta_j}{m_{X_{\cdot,j}Y}(y_n)}Yg(Y)\mathbb{1}_{\{Y\geq y_n\}}\right)\\
&\sim \frac{(1-\gamma c)(1-\gamma(c+1))\langle\alpha_3\rangle\langle\alpha_4\rangle}{y_n g(y_n)^2\bar{F}(y_n)}\\
&\times \left[\mathbb{E}(Yg(Y)^2\mathbb{1}_{\{Y\geq y_n\}}) - \mathbb{E}(g(Y)\mathbb{1}_{\{Y\geq y_n\}})\mathbb{E}(Yg(Y)\mathbb{1}_{\{Y\geq y_n\}})\right]\\
&\sim \frac{(1-\gamma c)(1-\gamma(c+1))\langle\alpha_3\rangle\langle\alpha_4\rangle}{y_n g(y_n)^2\bar{F}(y_n)}\\
&\times \left(\frac{y_n g(y_n)^2\bar{F}(y_n)}{1-\gamma(2c+1)} - \frac{y_n g(y_n)^2\bar{F}(y_n)^2}{(1-\gamma c)(1-\gamma(c+1))}(1+o(1))\right)\\
&\to b_{34}\,\langle\alpha_3\rangle\langle\alpha_4\rangle \text{ as } n\to\infty.
\end{aligned}$$

Finally, it follows that, as $n\to\infty$,

$$\begin{aligned}
\operatorname{var}(\chi'_{0,n}) &\to \alpha_1^2 + b_{22}\alpha_2^2 + b_{33}\langle\alpha_3\rangle^2 + b_{44}\langle\alpha_4\rangle^2 + 2\alpha_1\left(\alpha_2 + \langle\alpha_3\rangle + \langle\alpha_4\rangle\right) + 2\alpha_2\left(b_{23}\langle\alpha_3\rangle + b_{24}\langle\alpha_4\rangle\right)\\
&+ 2b_{34}\langle\alpha_3\rangle\langle\alpha_4\rangle\\
&= \alpha^t B\alpha,
\end{aligned}$$

where B is given in the statement of the Proposition. Remarking that $\chi'_{0,n}$ is the sum of a triangular array of independent, identically distributed and centered random variables, one may use Lyapunov criterion (Billingsley, 1995, Theorem 27.3), to prove its asymptotic normality. To this end, consider $\delta \in (0, \frac{1}{2(c+1)} - \gamma)$ and let us show that

$$n\mathbb{E}\left|\sum_{\ell=1}^{4} \frac{T_{\ell,1,n} - \mathbb{E}(T_{\ell,1,n})}{\sqrt{n}}\right|^{2+\delta} \to 0 \text{ as } n \to \infty. \tag{2.25}$$

Using both triangle and Jensen inequalities, we get

$$\begin{aligned}\left\{\mathbb{E}\left|\sum_{\ell=1}^{4}[T_{\ell,1,n} - \mathbb{E}(T_{\ell,1,n})]\right|^{2+\delta}\right\}^{1/(2+\delta)} &\leq \sum_{\ell=1}^{4}\left(\{\mathbb{E}|T_{\ell,1,n}|^{2+\delta}\}^{1/(2+\delta)} + \mathbb{E}|T_{\ell,1,n}|\right)\\ &\leq 8\max_{1\leq\ell\leq 4}\{\mathbb{E}|T_{\ell,1,n}|^{2+\delta}\}^{1/(2+\delta)}.\end{aligned}$$

Lemma 2.A.2 and Lemma 2.A.7(i) yield, as $n \to \infty$,

$$\begin{aligned}\mathbb{E}|T_{1,1,n}|^{2+\delta} &= \bar{F}(y_n)^{-\delta/2}\,|\alpha_1|^{2+\delta},\\ \mathbb{E}|T_{2,1,n}|^{2+\delta} &\sim \bar{F}(y_n)^{-\delta/2}\,|\alpha_2|^{2+\delta}\,\frac{(1-\gamma)^{2+\delta}}{1-\gamma(2+\delta)},\\ \mathbb{E}|T_{3,1,n}|^{2+\delta} &\sim \bar{F}(y_n)^{-\delta/2}\,|\langle\alpha_3\rangle|^{2+\delta}\,\frac{(1-\gamma c)^{2+\delta}}{1-\gamma c(2+\delta)},\\ \mathbb{E}|T_{4,1,n}|^{2+\delta} &\sim \bar{F}(y_n)^{-\delta/2}\,|\langle\alpha_4\rangle|^{2+\delta}\,\frac{(1-\gamma(c+1))^{2+\delta}}{1-\gamma(2+\delta)(c+1)},\end{aligned}$$

leading to

$$n\mathbb{E}\left|\sum_{\ell=1}^{4} \frac{T_{\ell,1,n} - \mathbb{E}(T_{\ell,1,n})}{\sqrt{n}}\right|^{2+\delta} = O\left([n\bar{F}(y_n)]^{-\delta/2}\right) \to 0 \text{ as } n \to \infty,$$

which proves (2.25). As a conclusion, $\chi'_{0,n} \xrightarrow{d} \mathcal{N}(0, \alpha^t B\alpha)$ and $\sqrt{n\bar{F}(y_n)}\Lambda_n \xrightarrow{d} \mathcal{N}(0, B)$. □

Proof of Proposition 2.3.2. Let us denote by $\sigma_n^{-1} := \sqrt{n\bar{F}(y_n)}$ and prove in a first step that

$$\sigma_n^{-1}\left(\frac{\hat{v}_j(y_n) - v_j(y_n)}{\|v(y_n)\|}\right)_{1\leq j\leq d} \xrightarrow{d} \xi\,(\beta_j)_{1\leq j\leq d} \text{ with } \xi \sim \mathcal{N}\left(0, \lambda(c,\gamma)\right), \tag{2.26}$$

using the notations introduced in Proposition 2.3.1. Denoting by $\vartheta_n := \sigma_n^{-1}\Lambda_n$, Proposition 2.3.1 shows that:

$$\begin{aligned}\hat{\bar{F}}(y_n) &= \bar{F}(y_n)(1+\sigma_n\vartheta_{1,n}), & \hat{m}_Y(y_n) &= m_Y(y_n)(1+\sigma_n\vartheta_{2,n}),\\ \hat{m}_{X_{\cdot,j}}(y_n) &= m_{X_{\cdot,j}}(y_n)(1+\sigma_n\vartheta_{j+2,n}), & \hat{m}_{X_{\cdot,j}Y}(y_n) &= m_{X_{\cdot,j}Y}(y_n)(1+\sigma_n\vartheta_{j+2+d,n}),\end{aligned}$$

for all $j \in \{1, \dots, d\}$ and with $\vartheta_n \xrightarrow{d} \mathcal{N}(0, B)$. Substituting in (2.4), we get

$$\begin{aligned}
\hat{v}_j(y_n) &= \hat{\bar{F}}(y_n)\hat{m}_{X_{\cdot,j}Y}(y_n) - \hat{m}_{X_{\cdot,j}}(y_n)\hat{m}_Y(y_n) \\
&= \bar{F}(y_n)m_{X_{\cdot,j}Y}(y_n) - m_{X_{\cdot,j}}(y_n)m_Y(y_n) \\
&+ \bar{F}(y_n)m_{X_{\cdot,j}Y}(y_n)\sigma_n\left[\vartheta_{1,n} + \vartheta_{j+2+d,n} + \sigma_n\vartheta_{1,n}\vartheta_{j+2+d,n}\right] \\
&- m_{X_{\cdot,j}}(y_n)m_Y(y_n)\sigma_n\left[\vartheta_{2,n} + \vartheta_{j+2,n} + \sigma_n\vartheta_{2,n}\vartheta_{j+2,n}\right],
\end{aligned}$$

and taking account of $v_j(y_n) = F(y_n)m_{X_{\cdot,j}Y}(y_n) - m_{X_{\cdot,j}}(y_n)m_Y(y_n)$ and $\sigma_n \to 0$ yields

$$\begin{aligned}
\sigma_n^{-1}(\hat{v}_j(y_n) - v_j(y_n)) &= \bar{F}(y_n)m_{X_{\cdot,j}Y}(y_n)\left[\vartheta_{1,n} + \vartheta_{j+2+d,n} + o_P(1)\right] \\
&- m_{X_{\cdot,j}}(y_n)m_Y(y_n)\left[\vartheta_{2,n} + \vartheta_{j+2,n} + o_P(1)\right].
\end{aligned}$$

From (2.22)–(2.24) in the proof of Proposition 2.3.1, it follows that, for all $j \in \{1, \dots, D\}$:

$$\begin{aligned}
\frac{\sigma_n^{-1}}{y_n g(y_n)\bar{F}(y_n)^2}(\hat{v}_j(y_n) - v_j(y_n)) &= \frac{\beta_j}{1-\gamma(c+1)}\left[\vartheta_{1,n} + \vartheta_{j+2+d,n} + o_P(1)\right] \\
&- \frac{\beta_j}{(1-\gamma)(1-\gamma c)}\left[\vartheta_{2,n} + \vartheta_{j+2,n} + o_P(1)\right]. \quad (2.27)
\end{aligned}$$

Besides, Lemma 2.A.4 yields

$$\|v(y_n)\| = \bar{F}(y_n)\mathbb{E}[g(Y)\Psi_{y_n}(Y)]\|\beta + \eta(y_n)\|,$$

with $\mathbb{E}[g(Y)\Psi_{y_n}(Y)] \sim \lambda_2(\gamma, c)\ y_n g(y_n)\bar{F}(y_n)$ and $\|\beta + \eta(y_n)\| \to \|\beta\| = 1$ as $y_n \to \infty$, from (2.19) and (2.20) in the proof of Proposition 2.2.2. It follows that

$$\|v(y_n)\| = \lambda_2(\gamma, c)y_n g(y_n)\bar{F}^2(y_n)(1 + o(1)). \quad (2.28)$$

Collecting (2.27) and (2.28) entails, for all $j = 1, \dots, d$:

$$\frac{\sigma_n^{-1}}{\|v(y_n)\|}(\hat{v}_j(y_n) - v_j(y_n)) = \beta_j\left[a_1(\vartheta_{1,n} + \vartheta_{j+2+d,n}) - a_2(\vartheta_{2,n} + \vartheta_{j+2,n})\right] + o_P(1),$$

where we have defined $a_1 = (1-\gamma)(1-\gamma c)/(\gamma^2 c)$ and $a_2 = (1-\gamma(c+1))/(\gamma^2 c)$. Consequently, we have proved that

$$\frac{\sigma_n^{-1}}{\|v(y_n)\|}(\hat{v}_j(y_n) - v_j(y_n))_{1\le j\le d} \xrightarrow{d} \mathcal{N}\left(0, \mathrm{diag}(\beta_1, \dots, \beta_d)ABA^t\mathrm{diag}(\beta_1, \dots, \beta_d)\right), \quad (2.29)$$

where A is the $d \times 2(d+1)$ matrix defined as follows

$$A = \left(\begin{array}{c|c|cccccc|ccccc} a_1 & -a_2 & -a_2 & 0 & \dots & \dots & 0 & a_1 & 0 & \dots & \dots & 0 \\ a_1 & -a_2 & 0 & -a_2 & \ddots & & 0 & 0 & a_1 & \ddots & & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & \ddots & -a_2 & 0 & \vdots & & \ddots & a_1 & 0 \\ a_1 & -a_2 & 0 & \dots & \dots & 0 & -a_2 & 0 & \dots & \dots & 0 & a_1 \end{array}\right).$$

Straightforward algebra shows that

$$\mathrm{diag}(\beta_1, \dots, \beta_d) A B A^t \mathrm{diag}(\beta_1, \dots, \beta_d) = \lambda(c, \gamma)(\beta_1, \dots, \beta_d)^t (\beta_1, \dots, \beta_d),$$

so that the limiting Gaussian distribution is non-degenerated in the only direction $(\beta_1, \dots, \beta_d)^t$. Therefore, the convergence in distribution (2.29) can be rewritten as in (2.26). The second step consists in proving that

$$\sigma_n^{-1} \left(\frac{\hat{v}_j(y_n) - v_j(y_n)}{\|v(y_n)\|} \right)_{d+1 \leq j \leq p} \xrightarrow{\mathbb{P}} 0. \tag{2.30}$$

For all $j = d+1, \dots, p$ and $\zeta \in \{0, 1\}$, the inverse model $(\mathbf{M}_1)$ shows that

$$\hat{m}_{X_{\cdot,j} Y^\zeta}(y_n) = \frac{1}{n} \sum_{i=1}^n X_{ij} Y_i^\zeta \mathbb{1}_{\{Y_i^\zeta \geq y_n\}} = \frac{1}{n} \sum_{i=1}^n \varepsilon_{ij} Y_i^\zeta \mathbb{1}_{\{Y_i^\zeta \geq y_n\}}.$$

Moreover, from (2.15) in the proof of Lemma 2.A.8, we have

$$\mathrm{var}(\hat{m}_{X_{\cdot,j} Y^\zeta}(y_n)) = O\left(\frac{y_n^{2\zeta} \bar{F}(y_n)^{1-2/q}}{n} \right),$$

and recalling that $\mathbb{E}\left(\hat{m}_{X_{\cdot,j} Y^\zeta}(y_n)\right) = m_{X_{\cdot,j} Y^\zeta}(y_n)$, a straightforward application of Markov inequality yields

$$\hat{m}_{X_{\cdot,j} Y^\zeta}(y_n) = m_{X_{\cdot,j} Y^\zeta}(y_n) + O_P\left(\frac{y_n^\zeta \bar{F}(y_n)^{\frac{1}{2} - \frac{1}{q}}}{\sqrt{n}} \right).$$

Substituting in (2.4) and taking into account Lemma 2.A.7(ii) yield

$$\begin{aligned}\hat{v}_j(y_n) &= \hat{\bar{F}}(y_n)\hat{m}_{X_{\cdot,j}Y}(y_n) - \hat{m}_{X_{\cdot,j}}(y_n)\hat{m}_Y(y_n) \\ &= \bar{F}(y_n)(1+\sigma_n\vartheta_{1,n})\left(m_{X_{\cdot,j}Y}(y_n) + O_P\left(\frac{y_n\bar{F}(y_n)^{\frac{1}{2}-\frac{1}{q}}}{\sqrt{n}}\right)\right) \\ &- m_Y(y_n)(1+\sigma_n\vartheta_{2,n})\left(m_{X_{\cdot,j}}(y_n) + O_P\left(\frac{\bar{F}(y_n)^{\frac{1}{2}-\frac{1}{q}}}{\sqrt{n}}\right)\right) \\ &= v_j(y_n) + O_P\left(\frac{y_n\bar{F}(y_n)^{\frac{3}{2}-\frac{1}{q}}}{\sqrt{n}}\right).\end{aligned}$$

Therefore, in view of (2.28), we have

$$\frac{\sigma_n^{-1}}{\|v(y_n)\|}(\hat{v}_j(y_n) - v_j(y_n)) = O_P\left(\frac{1}{g(y_n)\bar{F}(y_n)^{1/q}}\right),$$

for all $j \in \{d+1,\ldots,p\}$, as $n \to \infty$ and since $\bar{F}(\cdot)^{1/q}g(\cdot)$ is regularly-varying with index $c - 1/(q\gamma) > 0$. Finally, we can then infer that (2.30) holds, hence the result. □

Proof of Theorem 2.3.1. Let us recall that $\sigma_n^{-1} = \sqrt{n\bar{F}(y_n)}$ and consider the expansion

$$\sigma_n^{-1}\left(\frac{\hat{v}(y_n)}{\|v(y_n)\|} - \beta\right) = \sigma_n^{-1}\left(\frac{\hat{v}(y_n)}{\|v(y_n)\|} - w(y_n)\right) + \sigma_n^{-1}\left(w(y_n) - \beta\right).$$

First, Proposition 2.3.2 shows that

$$\sigma_n^{-1}\left(\frac{\hat{v}(y_n)}{\|v(y_n)\|} - w(y_n)\right) \xrightarrow{d} \xi\beta,$$

where $\xi \sim \mathcal{N}(0, \lambda(c,\gamma))$. Second, Proposition 2.2.2 entails that

$$\sigma_n^{-2}\|w(y_n) - \beta\|^2 = O\left(\frac{n\bar{F}^{1-2/q}(y_n)}{g^2(y_n)}\right) \to 0,$$

as $y_n \to \infty$, and the result is proved. □

2.A.3 Supplementary material for simulations

This Supplementary material includes additional results on simulated data (Section 2.5 of the main paper). The finite sample behavior of the EPLS estimator $\hat{v}(y_n)$ is illustrated and compared to SIMEXQ (single-index model extreme quantile) estimator. Seven cases are investigated:

- dimension $p = 3$, Frank copula and Student distribution (Figure 6),

- dimension $p = 30$, Frank copula and Pareto distribution (Figure 7),
- dimension $p = 30$, Frank copula and Student distribution (Figure 8),
- dimension $p = 3$, Gaussian copula and Pareto distribution (Figure 9),
- dimension $p = 3$, Gaussian copula and Student distribution (Figure 10),
- dimension $p = 30$, Gaussian copula and Pareto distribution (Figure 11),
- dimension $p = 30$, Gaussian copula and Student distribution (Figure 12),

while the remaining case ($p = 3$, Frank copula and Pareto distribution) is presented in Figure 2 of the main paper.

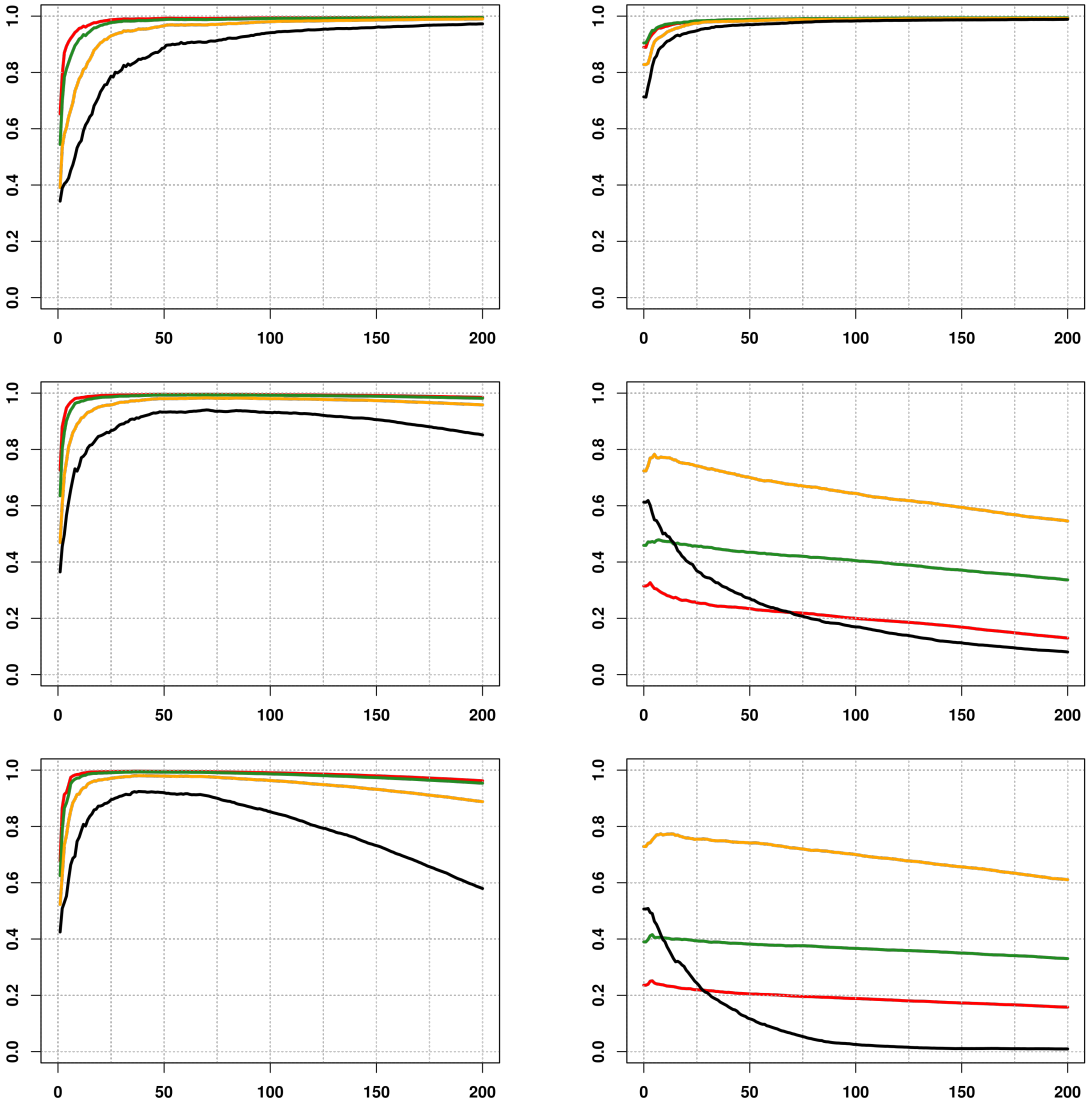


Figure 6 Finite sample behaviour of EPLS $\hat{v}(Y_{n-k+1,n})$ (left) and SIMEXQ (right) estimators, on simulated data from a Student distribution, Frank copula, dimension $p = 3$. Horizontally: number $k \in \{1, \ldots, 200\}$ of exceedances, vertically: $\mathrm{PC}(Y_{n-k+1,n})$ quality measure. From top to bottom, Frank copula parameter $\theta \in \{0, 10, 20\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

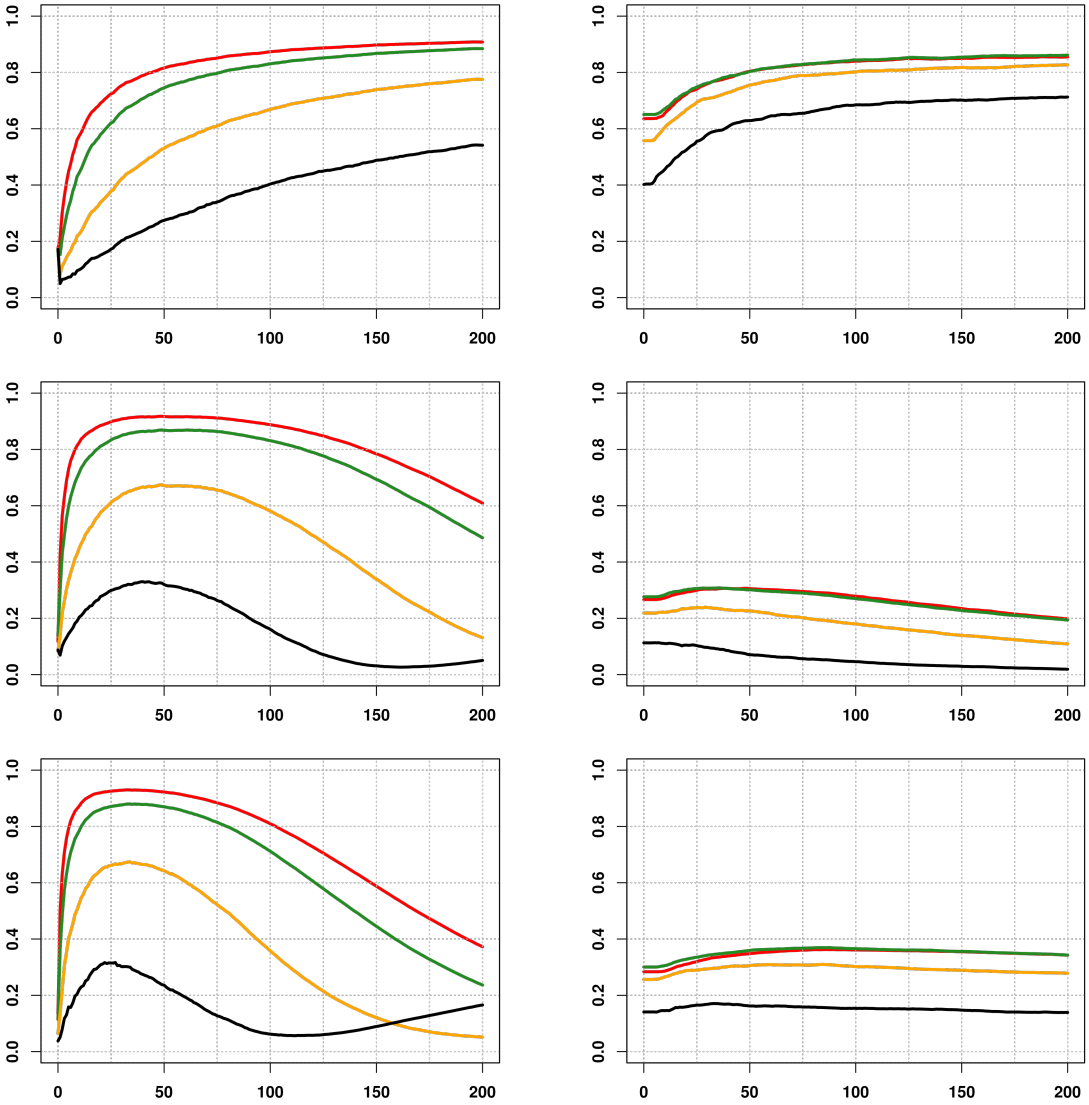


Figure 7 Finite sample behaviour of EPLS $\hat{v}(Y_{n-k+1,n})$ (left) and SIMEXQ (right) estimators, on simulated data from a Pareto distribution, Frank copula, dimension $p = 30$. Horizontally: number $k \in \{1, \ldots, 200\}$ of exceedances, vertically: $\mathrm{PC}(Y_{n-k+1,n})$ quality measure. From top to bottom, Frank copula parameter $\theta \in \{0, 10, 20\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

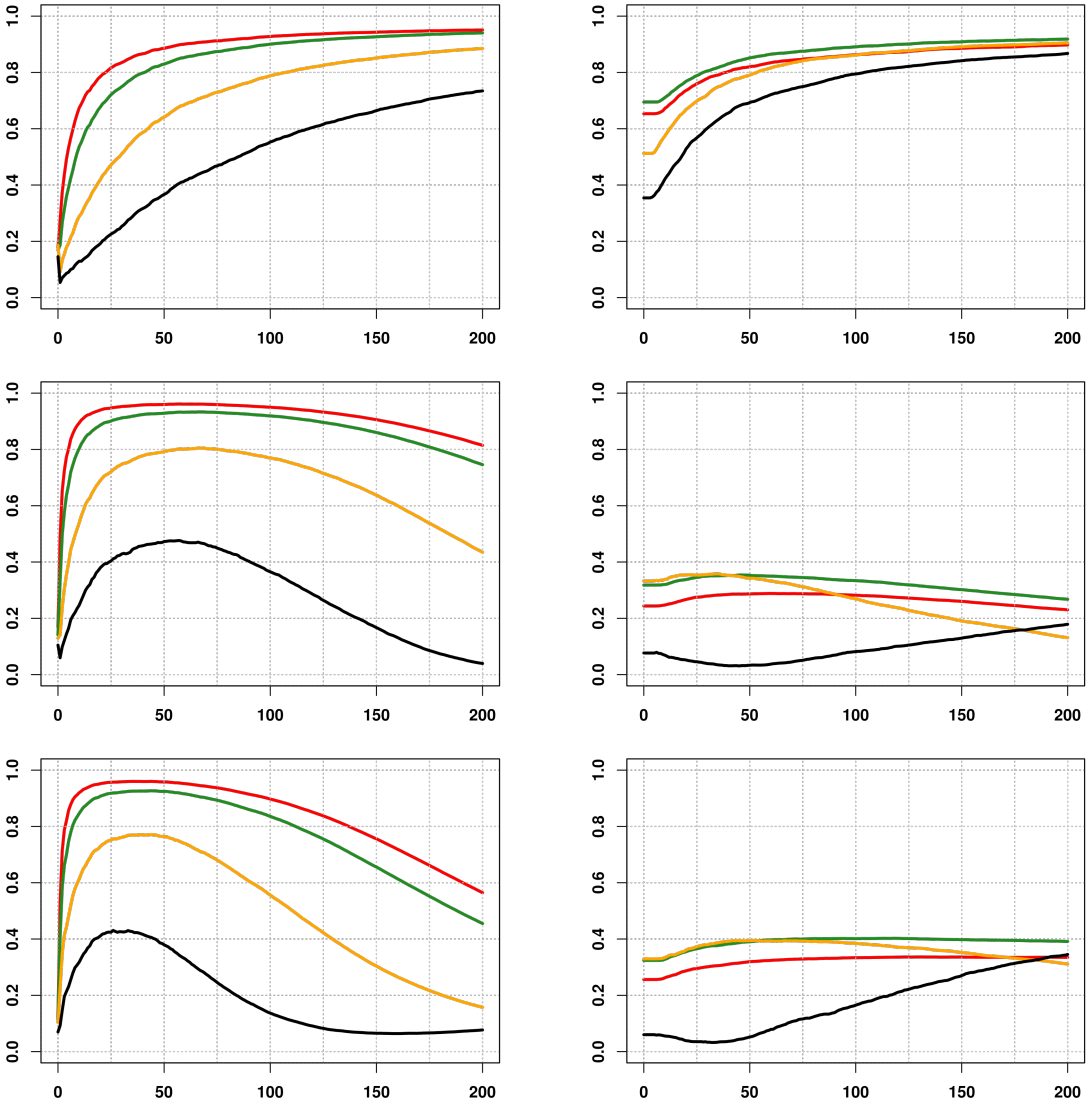


Figure 8 Finite sample behaviour of EPLS $\hat{v}(Y_{n-k+1,n})$ (left) and SIMEXQ (right) estimators, on simulated data from a Student distribution, Frank copula, dimension $p = 30$. Horizontally: number $k \in \{1, \ldots, 200\}$ of exceedances, vertically: $\mathrm{PC}(Y_{n-k+1,n})$ quality measure. From top to bottom, Frank copula parameter $\theta \in \{0, 10, 20\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

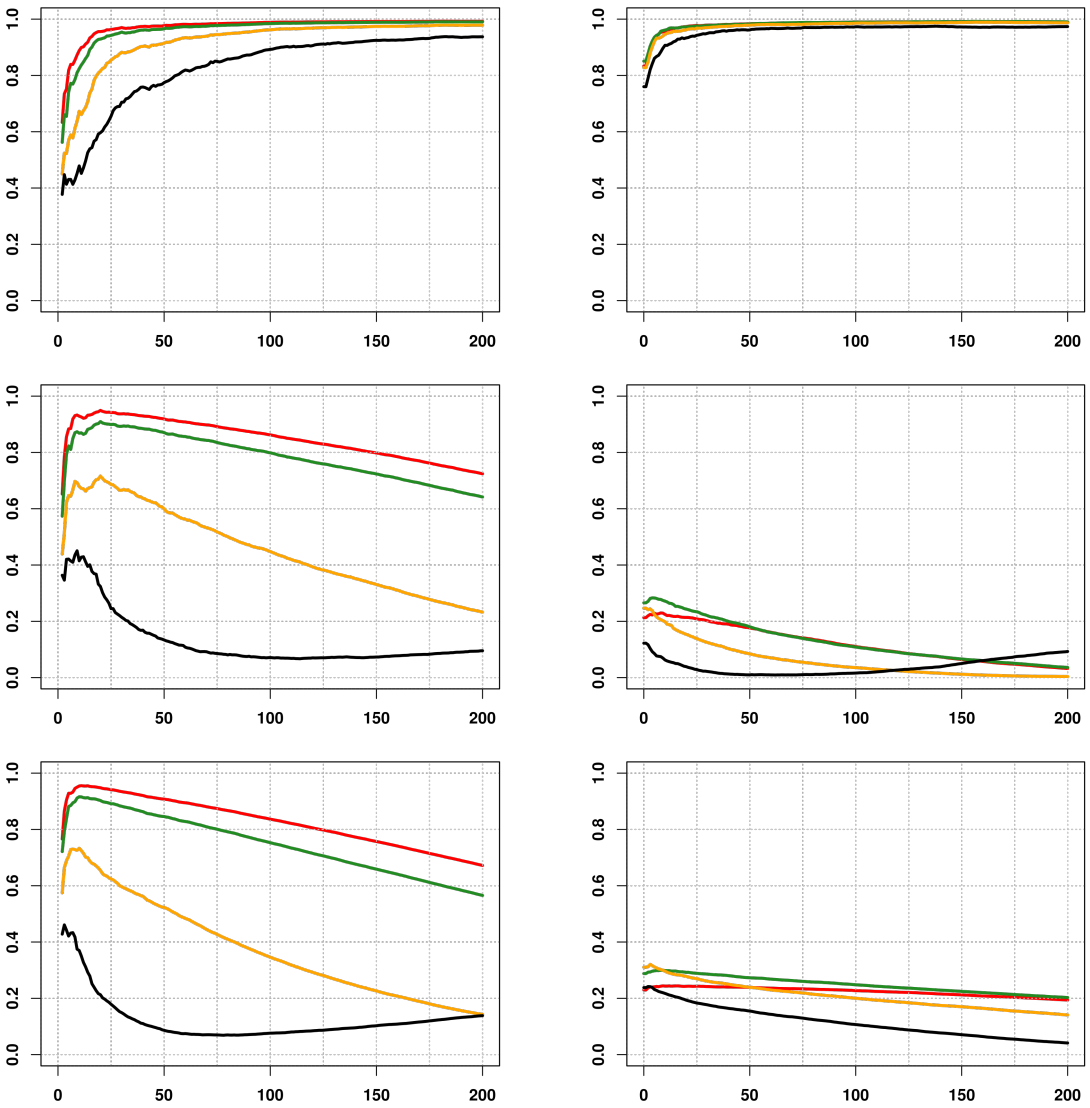


Figure 9 Finite sample behaviour of EPLS $\hat{v}(Y_{n-k+1,n})$ (left) and SIMEXQ (right) estimators, on simulated data from a Pareto distribution, Gaussian copula, dimension $p = 3$. Horizontally: number $k \in \{1, \ldots, 200\}$ of exceedances, vertically: $\mathrm{PC}(Y_{n-k+1,n})$ quality measure. From top to bottom, Gaussian copula parameter $\theta \in \{0, 0.87, 0.96\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

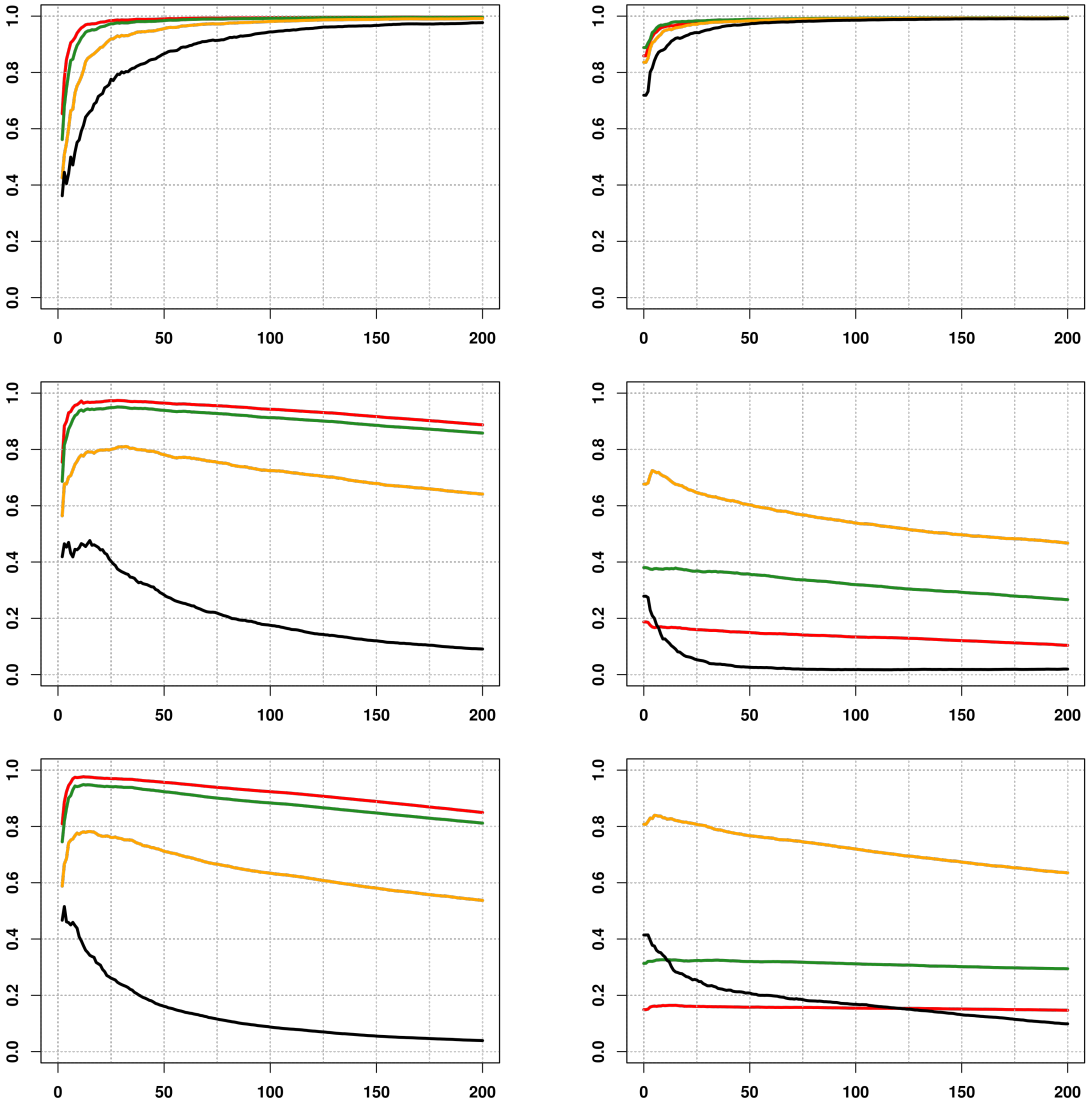


Figure 10 Finite sample behaviour of EPLS $\hat{v}(Y_{n-k+1,n})$ (left) and SIMEXQ (right) estimators, on simulated data from a Student distribution, Gaussian copula, dimension $p = 3$. Horizontally: number $k \in \{1, \ldots, 200\}$ of exceedances, vertically: $\mathrm{PC}(Y_{n-k+1,n})$ quality measure. From top to bottom, Gaussian copula parameter $\theta \in \{0, 0.87, 0.96\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

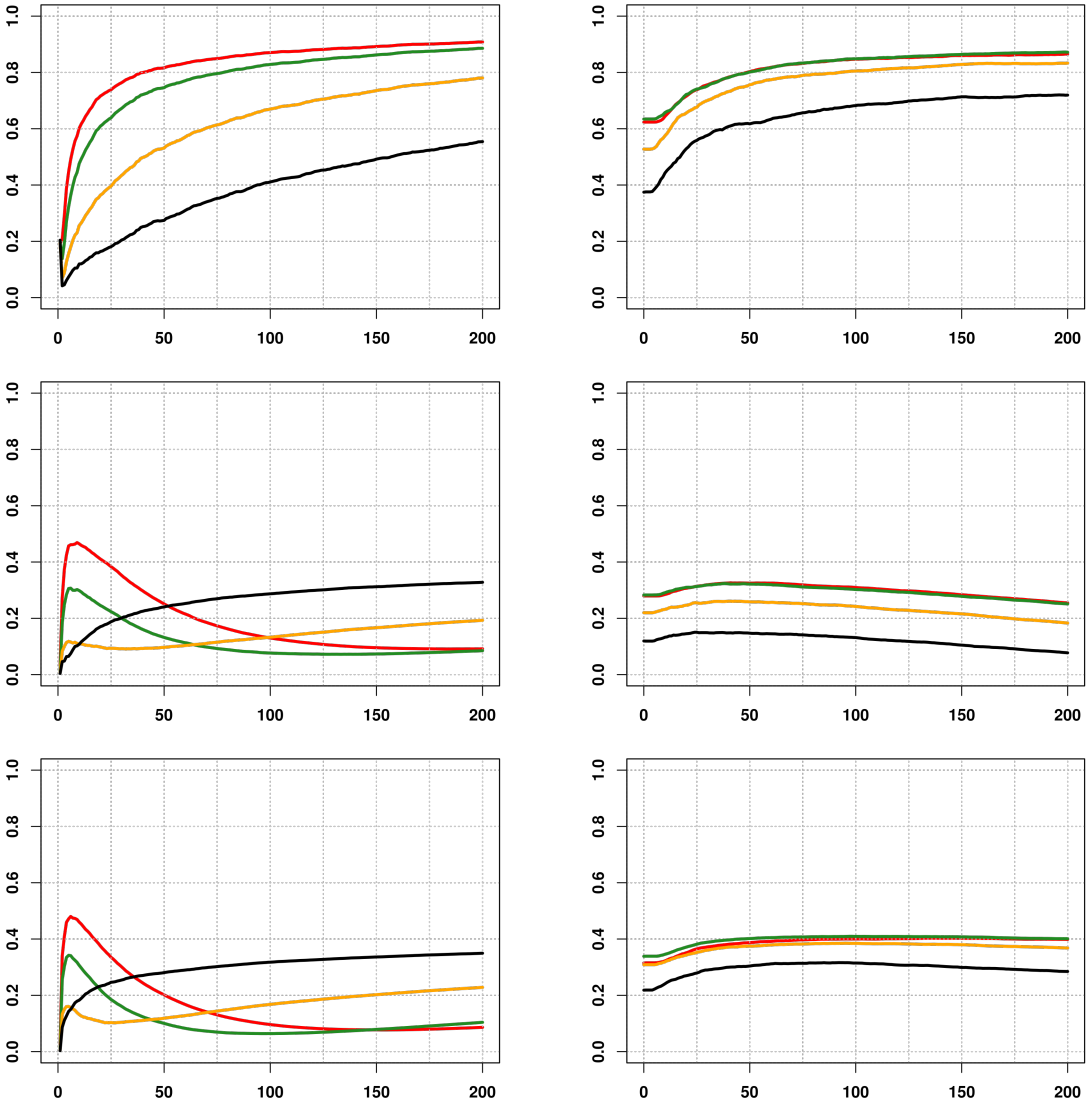


Figure 11 Finite sample behaviour of EPLS $\hat{v}(Y_{n-k+1,n})$ (left) and SIMEXQ (right) estimators, on simulated data from a Pareto distribution, Gaussian copula, dimension $p = 30$. Horizontally: number $k \in \{1, \ldots, 200\}$ of exceedances, vertically: $\mathrm{PC}(Y_{n-k+1,n})$ quality measure. From top to bottom, Gaussian copula parameter $\theta \in \{0, 0.87, 0.96\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

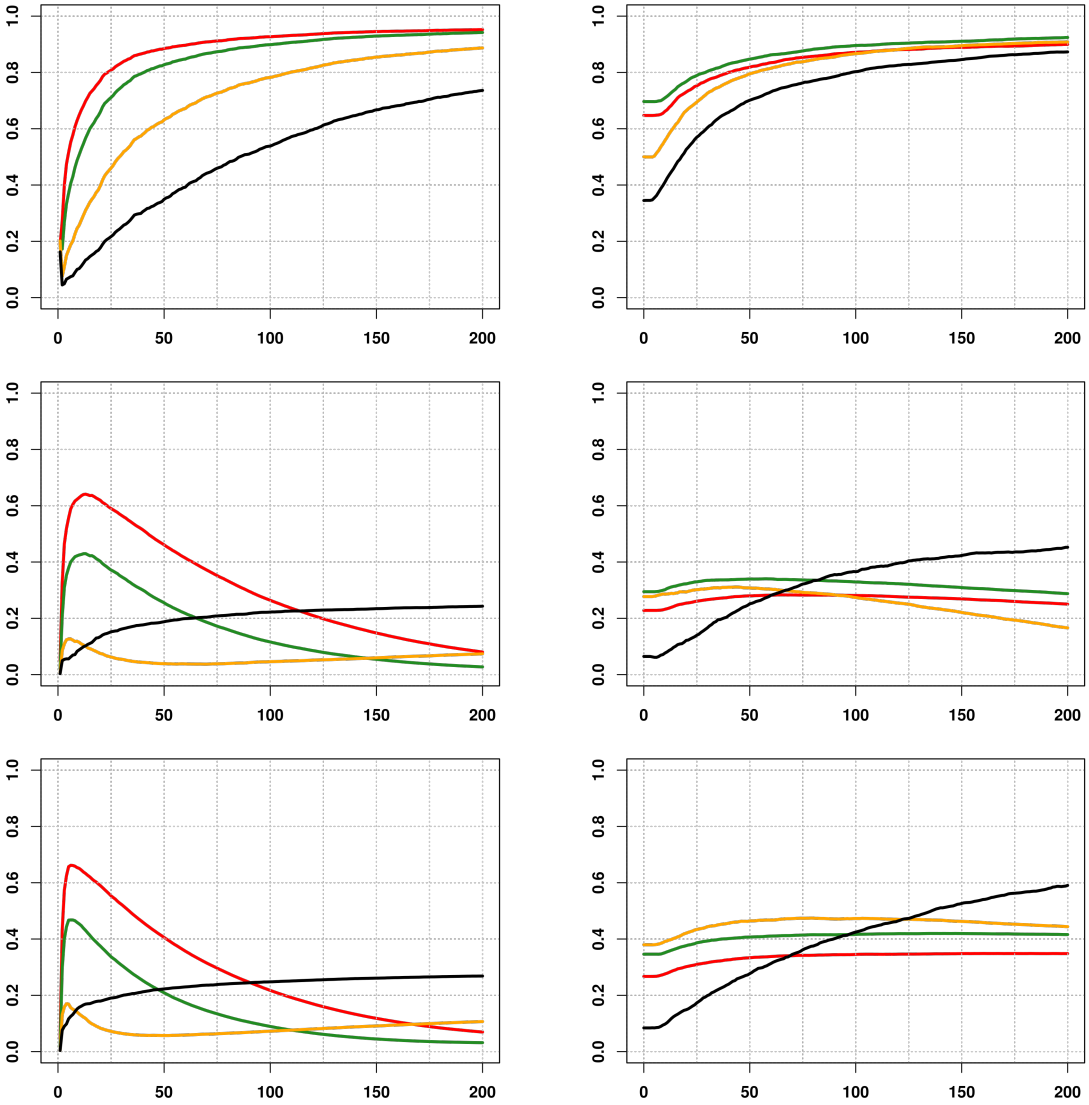


Figure 12 Finite sample behaviour of EPLS $\hat{v}(Y_{n-k+1,n})$ (left) and SIMEXQ (right) estimators, on simulated data from a Student distribution, Gaussian copula, dimension $p = 30$. Horizontally: number $k \in \{1, \ldots, 200\}$ of exceedances, vertically: $\mathrm{PC}(Y_{n-k+1,n})$ quality measure. From top to bottom, Gaussian copula parameter $\theta \in \{0, 0.87, 0.96\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

Chapter 3

Bayesian Extreme Partial Least-Squares

Contents

Abstract

The application of Bayesian inference in the context of extreme values is of great interest as it allows to deal with data scarcity problems. In this chapter, we adopt a Bayesian approach to the Extreme-PLS model, presented earlier in Chapter 2, to identify the direction of dimension reduction and introduce prior information on it. This chapter is the subject of a paper to be submitted for publication in the near future. Section 3.1 begins with a short introduction to the background of the Bayesian Extreme-PLS model. Section 3.2 proposes an adaptation of the von-Mises Fisher distribution, considered as a natural distribution for directional data, from the unit hypersphere to the hyperball. We propose in Section 3.3 a Bayesian formulation of the Extreme-PLS model and compute the posterior distribution of the dimension reduction direction. The associated likelihood function is derived from the data and is characterised by the von Mises-Fisher distribution adapted to a hyperball. Then we present three possible choices of the prior distributions, namely conjugate, hierarchical and sparse priors. The maximum a posteriori estimator of the direction is explicit for the conjugate and sparse priors case, while it has to be computed by MCMC sampling, for example, in the hierarchical one. The performance of the approach is studied through a simulation study in Section 3.4. We show that the proposed estimator is more efficient than Extreme-PLS in some situations, especially on small datasets. Section 3.5 illustrates the use of this approach on a French farm income dataset, to explain the lowest cereal yields given other factors. We conclude with a brief discussion in Section 3.6. The proofs are given in the Appendix.

Resumé

L'application de l'inférence bayésienne dans le contexte des valeurs extrêmes est d'un grand intérêt car cela permet de pallier les problèmes de rareté des données. Dans ce chapitre, nous adoptons une approche bayésienne du modèle Extreme-PLS, présenté précédemment dans le Chapitre 2, pour identifier la direction de la réduction de dimension et introduire des informations a priori à ce sujet. Ce chapitre fait l'objet d'un article qui sera soumis pour publication dans un avenir proche. La Partie 3.1 commence par une brève introduction au contexte du modèle Bayesian Extreme-PLS. La Partie 3.2 propose une adaptation de la distribution de von Mises-Fisher, considérée comme une distribution naturelle pour les données directionnelles, de l'hypersphère unité à la boule. Dans la Partie 3.3, nous proposons une formulation bayésienne du modèle Extreme-PLS et calculons la distribution postérieure de la direction de la réduction de dimension. La fonction de vraisemblance associée est dérivée des données, et est caractérisée par la distribution de von Mises-Fisher sur la boule. Nous présentons ensuite trois choix possibles de lois a priori, à savoir loi conjuguée, hiérarchique et sparse. L'estimateur maximum a posteriori de la direction est explicite pour les lois a priori conjuguées et sparse, tandis qu'il doit être calculé par échantillonnage MCMC, par exemple, pour la loi a priori hiérarchique. La performance de l'approche est évaluée par une étude sur simulations dans la Partie 3.4. Nous montrons que l'estimateur proposé est plus efficace que le modèle Extreme-PLS dans certaines situations, en particulier sur les données de petite taille. La Partie 3.5 illustre l'utilisation de cette approche sur un ensemble de données de revenus agricoles français, pour expliquer les rendements céréaliers les plus faibles compte tenu d'autres facteurs. Nous concluons par une brève discussion dans la Partie 3.6. Les preuves sont données en Annexe.

3.1 Introduction

In statistical regression for extreme values, one has to deal with problems where the scarcity of extreme events limits the available data to provide a robust regression. Furthermore, the number n of available observations is usually much smaller than the dimension p and thus one also has to deal with the well-known "curse of dimensionality".
Extracting and identifying a low-dimensional subspace of the covariates X, that maintains a high relationship between X and the response variable Y, is a crucial step. Partial least squares (PLS) regression (Wold, 1975) is one of the most popular methods combining principal component analysis (PCA) for dimension reduction and multiple regression. Sliced inverse regression (SIR) (Li, 1991) is also a very popular class of methods that estimates a central dimension reduction (DR) subspace based on the conditional distribution of X given Y, i.e. inverse regression. Several extensions have been developed for SIR (Li et al., 2007; Wu, 2008; Chiancone et al., 2017; Coudret et al., 2014) and PLS (Cook et al., 2013; Chun and Keleş, 2010), among others (see Section 1.3). While all these dimension reduction methods adopt the frequentist approach, there exists some works in the literature that use the Bayesian approach to estimate the reduction dimension space. In Reich et al. (2011), the authors propose a Bayesian method of dimension reduction by placing a prior on the central subspace, Mao et al. (2010) propose a Bayesian framework for dimension reduction using a nonparametric Bayesian mixture modeling approach, Cai et al. (2021) considers a Bayesian approach to calculate the conditional distribution of X given Y and perform dimension reduction.
On the other side, there are some few works in the literature on dimension reduction dedicated to extreme conditional quantiles (see Paragraph 1.5.4.4). One can mention Gardes (2018) who proposes a dimension reduction method and a conditional extremal quantile estimator by considering the tail dimension reduction subspace. The work of Xu et al. (2020) introduces a semi-parametric approach for the estimation of extreme conditional quantiles basing on a tail single-index model. The dimension reduction direction β is estimated through fitting a misspecified linear quantile regression model. Extreme-PLS model, proposed in Chapter 2, is a dimension reduction method based on PLS to find the linear combinations of covariates X that best explain the extreme values of Y. To our best of knowledge, there is no existing work that adopts the Bayesian approach for dimension reduction in the regression context and conditional extremes. The latter is of great interest as it allows us to deal with problems with a very small amount of data, or to incorporate domain knowledge of covariates structures by using appropriate priors.

In this work, we develop a Bayesian methodology, called Bayesian Extreme-PLS, which extends the model-based approach proposed in Chapter 2 to make it more efficient for small

data problems. In particular, we remain within the Extreme-PLS framework to identify the dimension reduction direction and introduce some prior information on it. First, a Bayesian formulation is provided for computing the posterior distribution of β. To this end, the likelihood function of X given Y and β is derived from the data and is characterised by the von Mises-Fisher distribution. This distribution, which naturally arises for directional data distributed on the unit sphere (Mardia and Jupp, 2009) is here adapted to hyperballs. Then, we establish the posterior distribution of β from the likelihood function and a desired prior distribution. Several criteria are investigated for the prior distribution, including for instance a sparsity assumption, where only certain coordinates of X will be useful to explain Y. Second, once the posterior distribution is computed, one can use some usual statistics, such as the posterior mean or mode to estimate β in a Bayesian way.

This chapter is organized as follows. First, we propose in Section 3.2 an adaptation of the von Mises-Fisher distribution from the hypersphere to the hyperball. Bayesian inference on the Extreme Partial Least Squares model is described in Section 3.3, where three possible priors on the direction β are discussed. The behaviour of Bayesian estimators is illustrated on simulated data in Section 3.4, while an application on French farm income data is described in Section 3.5. A discussion is provided in Section 3.6 and proofs are postponed to the Appendix.

3.2 von Mises-Fisher distribution: From the hypersphere to the hyperball

The von Mises-Fisher distribution (Mardia, 1975) is a fundamental probability distribution used in the analysis of directional data, and it is considered as a spherical analogue of the multivariate Gaussian distribution in $\mathbb{R}^p$. This distribution was introduced by Watson and Williams (1956) and has been discussed in details by Mardia (1975). We focus on the use of the von Mises–Fisher distribution since it appears naturally for directional data, and then we propose an extension of this distribution from the hypersphere to the hyperball.

3.2.1 von Mises-Fisher distribution on the unit hypersphere

A p-dimensional unit random vector X is said to have a p-variate von Mises-Fisher distribution vMF/S(μ, κ) on the unit hypersphere $S^{p-1} = \{x \in \mathbb{R}^p / \|x\| = 1\}$, if its probability density is given for all $x \in S^{p-1}$ by:

$$f(x|\mu, \kappa) = c_p(\kappa) e^{\kappa \mu^t x}, \tag{3.1}$$

where $\mu \in S^{p-1}$, $\kappa \geq 0$, $p \geq 2$. The normalising constant $c_p(\kappa)$ is given by:

$$c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)}, \tag{3.2}$$

where $I_p(\cdot)$ is the modified Bessel function of the first kind and order p, defined as (Abramowitz and Stegun, 1964):

$$I_p(\kappa) = \sum_{l \geq 0} \frac{1}{\Gamma(p+l+1) l!} \left(\frac{\kappa}{2}\right)^{2l+p}, \tag{3.3}$$

where $\Gamma(\cdot)$ is the Gamma function which verifies the property: $\Gamma(p+1) = p\Gamma(p)$. The density $f(\cdot|\mu, \kappa)$ is parameterised by the a location parameter μ (μ is the modal location on the hypersphere, see the following Lemma 3.2.1), and the concentration parameter κ which characterises the degree of concentration of the unit vectors drawn from $f(\cdot|\mu, \kappa)$ around μ. Large values of κ imply a strong concentration around the modal location μ. In contrast, when $\kappa = 0$, $f(x|\mu, \kappa)$ reduces to the uniform density on S^{p-1}.

The following Lemma gives the mode and the first moment of a von Mises-Fisher distribution.

Lemma 3.2.1. *Let X be a p-dimensional unit vector having a von Mises-Fisher distribution $vMF/S(\mu, \kappa)$ on the unit hypersphere S^{p-1}, with $\mu \in S^{p-1}$ and $\kappa \geq 0$. Then,*

(i) The mode of X is μ.

(ii) $\mathbb{E}(X) = \mu A_p(\kappa)$, where $A_p(\kappa) = I_{p/2}(\kappa)/I_{p/2-1}(\kappa)$.

Note that the von Mises-Fisher distribution is unimodal with mode μ. The expectation of the latter is collinear with μ and since $A_p(\kappa)$ is the ratio of Bessel functions, we cannot obtain a closed form. Nevertheless, using the asymptotic expansion of the modified Bessel function (see Abramowitz and Stegun (1964), p. 377) when $\kappa \to \infty$, given by

$$I_p(\kappa) = \frac{e^\kappa}{\sqrt{2\pi\kappa}} \left(1 - \frac{4p^2 - 1}{8\kappa}\right) + O\left(\frac{1}{\kappa^2}\right), \tag{3.4}$$

the following expansion of $A_p(\kappa)$ can be established:

$$A_p(\kappa) = 1 - \frac{p-1}{2\kappa} + \frac{(p-1)(p-3)}{8\kappa^2} + O\left(\frac{1}{\kappa^3}\right), \tag{3.5}$$

as $\kappa \to \infty$, see Chatelain and Le Bihan (2013). As a result, when $\kappa \to \infty$, $\mathbb{E}(X) \to \mu$. Besides, in view of (3.3), one also has, when $\kappa \to 0$,

$$I_p(\kappa) = \frac{\kappa^p}{2^p p!} + \frac{\kappa^{2+p}}{2^{p+2}(1+p)!} + O\left(\kappa^{p+4}\right), \tag{3.6}$$

leading to

$$A_p(\kappa) = \frac{\kappa}{p} + O\left(\kappa^3\right), \tag{3.7}$$

as $\kappa \to 0$ and in this case $\mathbb{E}(X) \sim \mu\kappa/p$.

3.2.2 von Mises-Fisher distribution on the hyperball

Our goal is to define a similar distribution on the hyperball of radius $r > 0$ denoted by $B^p(r) = \{x \in \mathbb{R}^p / \|x\| \leq r\}$. To this end, it is sufficient to adapt the von Mises-Fisher distribution defined on the unit hypersphere S^{p-1} to the hyperball $B^p(r)$ by considering for all $x \in B^p(r)$,

$$f(x|\mu, \kappa, r) = \frac{c'_p(\kappa)}{r^p} e^{\kappa \mu^t x / r}, \tag{3.8}$$

with $\mu \in S^{p-1}$, $\kappa \geq 0$, $r > 0$, $p \geq 2$ and where $c'_p(\kappa)$ is the normalizing constant to be determined.

Lemma 3.2.2. *The normalizing constant of the von Mises-Fisher distribution defined on the hyperball $B^p(r)$ is given by:*

$$c'_p(\kappa) = 2\pi c_{p+2}(\kappa). \tag{3.9}$$

The von Mises-Fisher distribution on the hyperball $B^p(r)$ is denoted by $\text{vMF/B}(\mu, \kappa, r)$.

3.3 Bayesian inference for Extreme Partial Least Squares

3.3.1 Framework

In the framework of the adaptation of Partial Least Squares to the extreme-value case, the following single-index non linear inverse regression model is introduced in Section 2.2:

(M) $X = g(Y)\beta + \varepsilon$, where X and ε are p-dimensional random vectors, Y is a real random variable, $g : \mathbb{R} \to \mathbb{R}$ is an unknown link function, $\beta \in \mathbb{R}^p$ is an unknown unit vector.

Here, Y is the response variable, X is the multidimensional covariate and ε is a multidimensional error term. Model (**M**) is referred to as an inverse regression model since the covariates are written as functions of the response variable. Considering (X_i, Y_i), $i \in \{1, \ldots, n\}$ a n sample with same distribution as (X, Y), the Extreme-PLS estimate of the direction $\beta \in S^{p-1}$ is obtained by maximizing the covariance between $\hat{\beta}^t X$ and Y conditionally on large values of Y:

$$\hat{\beta}(y) = \underset{\|\beta\|=1}{\text{argmax}}\, \beta^t \left\{ \hat{\bar{F}}(y) \hat{m}_{XY}(y) - \hat{m}_X(y) \hat{m}_Y(y) \right\}, \tag{3.10}$$

with $\hat{\bar{F}}$ the empirical survival function of Y,

$$\hat{m}_{XY}(y) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \mathbb{1}_{\{Y_i \geq y\}}, \hat{m}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\{Y_i \geq y\}}, \hat{m}_X(y) = \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}_{\{Y_i \geq y\}},$$

as y is some high threshold, see Section 2.3 for details. Here, we aim at introducing **a prior** probability distribution on the direction β. The latter may reflect our prior information on this direction. Then, we update the prior distribution using the information contained in the observed data to obtain the so-called posterior distribution. To do so, one can rewrite the optimization problem (3.10) as follows:

$$\begin{aligned}\hat{\beta}(y) &= \underset{\|\beta\|=1}{\operatorname{argmax}} \exp\left(\beta^t \left\{\frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}_{\{Y_i \geq y\}} \left[\hat{\bar{F}}(y) Y_i - \hat{m}_Y(y)\right]\right\}\right) \\ &= \underset{\|\beta\|=1}{\operatorname{argmax}} \prod_{i=1}^{n} \exp\left(\beta^t X_i \Phi_y(Y_{1:n})\right), \end{aligned} \tag{3.11}$$

where $Y_{1:n}$ denotes the vector $(Y_1, \ldots, Y_n)^t \in \mathbb{R}^n$ and $\Phi_y(Y_{1:n}) = \frac{1}{n}\mathbb{1}_{\{Y_i \geq y\}}[\hat{\bar{F}}(y)Y_i - \hat{m}_Y(y)]$. Remarking that, under model (**M**) and using the triangle inequality, $\|X_i\| \leq |g(Y_i)| + \|\varepsilon_i\|$, the optimization problem (3.11) can be interpreted in terms of density associated with the vMF/B distribution

$$\hat{\beta}(y) = \underset{\|\beta\|=1}{\operatorname{argmax}} \prod_{i=1}^{n} f_{\text{vMF/B}}\left(X_i|\beta, \kappa = \Phi_y(Y_{1:n})(|g(y)| + \|\varepsilon_i\|), r = |g(y)| + \|\varepsilon_i\|\right).$$

It appears that $\hat{\beta}$ can be interpreted as the estimator maximizing the likelihood conditionally on $Y_{1:n}$ and $\varepsilon_{1:n}$. Since the distribution $p(\cdot,\cdot)$ of $(Y_{1:n}, \varepsilon_{1:n})$ does not depend on β, one also has

$$\hat{\beta}(y) = \underset{\|\beta\|=1}{\operatorname{argmax}} \left(\prod_{i=1}^{n} f_{\text{vMF/B}}\left(X_i|\beta, \kappa = \Phi_y(Y_{1:n})(|g(y)| + \|\varepsilon_i\|), r = |g(y)| + \|\varepsilon_i\|\right)\right) p(Y_{1:n}, \varepsilon_{1:n}),$$

and $\hat{\beta}(y)$ can also be viewed as the unconditional maximum likelihood estimator of β. Thus, by introducing a prior distribution $\pi(\cdot)$ on β, denoted by $\pi(\beta)$, and in view of Bayes' rule, the corresponding posterior distribution of β is given by

$$\begin{aligned} p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) &\propto \prod_{i=1}^{n} f_{\text{vMF/B}}\left(X_i|\beta, \kappa = \Phi_y(Y_{1:n})(|g(Y_i)| + \|\varepsilon_i\|), r = |g(Y_i)| + \|\varepsilon_i\|\right) \\ &\times\ p(Y_{1:n}, \varepsilon_{1:n})\pi(\beta). \end{aligned}$$

Since $p(Y_{1:n}, \varepsilon_{1:n})$ does not depend on β, the posterior distribution can be simplified as

$$p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) \propto \pi(\beta) \prod_{i=1}^{n} f_{\text{vMF/B}}\left(X_i|\beta, \kappa = \Phi_y(Y_{1:n})(|g(Y_i)| + \|\varepsilon_i\|), r = |g(Y_i)| + \|\varepsilon_i\|\right).$$

This formalism opens the door to the construction of Bayesian estimators for β. Let us remark that, the posterior distribution can be further simplified as

$$p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) \propto \pi(\beta) \prod_{i=1}^{n} \exp\left(\beta^t X_i \Phi_y(Y_{1:n})\right) = \pi(\beta) \exp\left(\beta^t S_n(y)\right), \tag{3.12}$$

where we have defined $S_n(y) = \sum_{i=1}^{n} X_i \Phi_y(Y_{1:n})$.

Let us now discuss the choice of the prior distribution $\pi(\cdot)$ which is a crucial step in Bayesian approaches. There are several possible criteria for choosing the prior distribution. Some are based on practical considerations. For example, some prior distributions lead to simple computational posterior distributions, such as the conjugate family case. Using a family of conjugate distributions makes the calculations quite simple when the posterior parameters are explicitly expressed in terms of the prior parameters and the data. Indeed, if one knows how to simulate according to the considered distribution, or compute some aspects such as the mean or the mode, the simulation according to the posterior distribution is a special case and can be less complex. This situation is illustrated in Paragraph 3.3.2. It is also possible to use several levels of prior distributions, which leads to so-called hierarchical methods, see Paragraph 3.3.3. Another criteria is based on past experience, the expertise of specialists in a certain field or the intuition of the statistician. Some ideas for selecting variables, regularising or encouraging the estimator to be sparse can be expressed through a well chosen prior distribution (Van Erp et al., 2019). Such an example is derived in Paragraph 3.3.4.

3.3.2 Conjugate prior

Using the fact that the vMF distribution belongs to the exponential family, one may take the corresponding conjugate standard prior for β, see Nunez-Antonio and Gutiérrez-Pena (2005). Indeed, it is convenient for Bayesian inference to have conjugate priors for the parameters of a distribution since posterior distributions remain in the same class as the prior distribution. We thus assume a vMF/S prior distribution for the direction β on the hypersphere S^{p-1}, with location vector $\mu \in S^{p-1}$ and concentration parameter $\kappa_1 \geq 0$. Using (3.12), the posterior is written as:

$$\begin{aligned} p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) &\propto \exp(\kappa_1 \beta^t \mu) \exp\left(\beta^t S_n(y)\right) \\ &= \exp\left(\beta^t \left\{S_n(y) + \kappa_1 \mu\right\}\right). \end{aligned} \tag{3.13}$$

Note that the posterior is a von Mises-Fisher distribution with location parameter $(S_n(y) + \kappa_1\mu)/\|S_n(y)+\kappa_1\mu\|$ and concentration $\|S_n(y)+\kappa_1\mu\|$. In this case, the maximum a posteriori (MAP) estimator coincides with the mode of the vMF/S distribution:

$$\hat{\beta}^C_{MAP} = \frac{S_n(y) + \kappa_1\mu}{\|S_n(y) + \kappa_1\mu\|},$$

from Lemma 3.2.1. The MAP estimator is a linear combination of the prior direction μ with the Extreme-PLS estimator $S_n(y)/\|S_n(y)\|$. Setting $\kappa_1 = 0$ amounts to assuming an uniform prior distribution for the direction β and we thus recover the Extreme-PLS framework. In contrast, letting $\kappa_1 \to \infty$ yields $\hat{\beta}^C_{MAP} \to \mu$.

3.3.3 Hierarchical prior

In Bayesian statistics, recall that it is assumed that the data distribution is driven by a parameter β which is itself random and following a prior distribution. In some situations, the distribution of β may in turn depend on a parameter μ which is unknown. A fully Bayesian approach consists in considering μ itself random and choosing a prior distribution on μ. The parameter μ is called a hyperparameter.

vMF/S(μ, κ_1) prior for β with κ_1 fixed and uniform prior on μ. Assume that β given μ follows a vMF/S(μ, κ_1) distribution and that μ is uniformly distributed on the unit hypersphere S^{p-1}. The prior density of μ is constant over S^{p-1} and we have:

$$p(\beta, \mu | X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) \propto \exp\left(\beta^t \{S_n(y) + \kappa_1\mu\}\right),$$

in view of (3.13). Thus, the posterior distribution is defined by

$$\begin{aligned} p(\beta | X_{1:n}, Y_{1:n}, \varepsilon_{1:n}, \mu) &= \int_{S^{p-1}} p(\beta, \mu | X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) d\mu \\ &\propto \exp\left(\beta^t S_n(y)\right) \int_{S^{p-1}} \exp\left(\kappa_1 \beta^t \mu\right) d\mu \\ &\propto \exp\left(\beta^t S_n(y)\right). \end{aligned}$$

As a consequence, the posterior distribution of β coincides with the likelihood, the prior distribution is not taken into account in the estimation.

vMF/S(μ, κ_1) prior for β with prior on (μ, κ_1). Assume that β given μ follows a vMF/S(μ, κ_1) distribution, that μ given κ_1 follows a vMF/S$(\mu_0, \kappa_1 q_0)$ distribution and that κ_1 is Gamma distributed:

$$p(\mu, \kappa_1) = p(\mu|\kappa_1)p(\kappa_1) = f_{\text{vMF/S}}(\mu|\mu_0, \kappa_1 q_0) f_{\mathcal{G}}(\kappa_1|a_0, b_0),$$

where $\mu_0 \in S^{p-1}$, $q_0 \geq 0$ and $f_{\mathcal{G}}(\cdot|a_0, b_0)$ is the Gamma density function with shape parameter $a_0 > 0$ and inverse scale parameter $b_0 > 0$. The choice of Gamma distribution is motivated by the fact κ_1 is scalar and positive. Moreover, the Gamma density is flexible enough to approximate well the posterior marginal density of κ_1 (see Taghia et al. (2014), p. 1703). Note that when $q_0 = 0$, we recover a uniform distribution on S^{p-1} for μ, as seen above. From (3.12), one has

$$\begin{aligned} p(\beta, \mu, \kappa_1 | X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) &\propto f_{\text{vMF/S}}(\beta|\mu, \kappa_1) f_{\text{vMF/S}}(\mu|\mu_0, \kappa_1 q_0) f_{\mathcal{G}}(\kappa_1|a_0, b_0) p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) \\ &\propto c_p(\kappa_1) \exp\left(\beta^t S_n(y)\right) \exp\left(\mu^t\{\kappa_1\beta + q_0\kappa_1\mu_0\}\right) \exp(-b_0\kappa_1)\kappa_1^{a-1}. \end{aligned}$$

Integrating with respect to μ and κ_1, the posterior distribution is given by

$$\begin{aligned} p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}, \mu, \kappa_1) &= \int_0^\infty \int_{S^{p-1}} p(\beta, \mu, \kappa_1|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) d\mu d\kappa_1 \\ &\propto e^{\beta^t S_n} \int_0^\infty c_p(\kappa_1) e^{-b_0\kappa_1} \kappa_1^{a-1} \int_{S^{p-1}} \exp\left(\mu^t\{\kappa_1\beta + q_0\kappa_1\mu_0\}\right) d\mu \; d\kappa_1. \end{aligned}$$

Moreover, in view of Section 3.2.1, we have:

$$\begin{aligned} \int_{S^{p-1}} \exp\left(\mu^t\{\kappa_1\beta + q_0\kappa_1\mu_0\}\right) d\mu &= \int_{S^{p-1}} \exp\left(\|\kappa_1\beta + q_0\kappa_1\mu_0\| \mu^t \frac{\kappa_1\beta + q_0\kappa_1\mu_0}{\|\kappa_1\beta + q_0\kappa_1\mu_0\|}\right) d\mu \\ &= 1/c_p(\kappa_1\|\beta + q_0\mu_0\|), \end{aligned}$$

and therefore

$$p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}, \mu, \kappa_1) \propto e^{\beta^t S_n} \int_0^\infty \frac{c_p(\kappa_1)}{c_p(\kappa_1\|\beta + q_0\mu_0\|)} e^{-b_0\kappa_1} \kappa_1^{a-1} d\kappa_1.$$

It appears that the posterior distribution is not explicit. The computation of the MAP should then rely on simulation methods such as Importance Sampling or Monte-Carlo Markov Chains (Besag, 2001; Gamerman and Lopes, 2006; Hastings, 1970; Metropolis et al., 1953) presented in Section 1.4.1. This will be the subject of our future work.

3.3.4 Sparse prior

The Extreme-PLS method can be improved by combining the assumption that only a few covariates X are useful to explain the response variable Y. Thus, our goal is to impose sparsity on the direction of dimension reduction β. Introducing a sparsity prior on β corresponds to the assumption that many coordinates β_j can be set to zero without affecting the fit of the model. Let us consider a Laplace distribution $\pi(\beta|\lambda) = b_p(\lambda)\exp(-\lambda\|\beta\|_1)$ as a prior for $\beta \in S^{p-1}$, with $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ is the L_1 norm, $\lambda \geq 0$ is a concentration parameter and $b_p(\lambda)$ is an appropriate normalizing constant. Such a prior on β is likely to induce sparsity

since the shape of Laplace distribution has a peak at zero. Indeed, robust Lasso regression (L_1 regularization) in a Bayesian setting is equivalent to using a Laplace prior (see Tibshirani (1996), p. 277). From (3.12), the posterior is written as:

$$p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) \propto \exp\left(\beta^t S_n - \lambda\|\beta\|_1\right). \tag{3.14}$$

In this case, taking the logarithm, the MAP can be interpreted as penalizing the covariance with a L_1 term to enforce sparse solutions. The MAP estimator is given by:

$$\hat{\beta}^S_{MAP}(y) = \underset{\|\beta\|^2=1}{\operatorname{argmax}}\, \beta^t S_n(y) - \lambda\|\beta\|_1, \tag{3.15}$$

This linear optimization problem under a quadratic constraint benefits from a closed-form solution obtained with Lagrange multipliers method and given in the next Proposition.

Proposition 3.3.1. *The unique solution of the optimization problem (3.15) is given, for all $y \in \mathbb{R}$ and $\lambda \geq 0$, by:*

$$\hat{\beta}^S_{MAP}(y) = \tilde{\beta}(y)/\|\tilde{\beta}(y)\|, \text{ where } \tilde{\beta}_j(y) = G_\lambda(S_{n,j}(y)), \tag{3.16}$$

for $j \in \{1, \ldots, p\}$, and G_λ is the shrinkage operator defined as

$$G_\lambda(x) = sign(x)\,(|x| - \lambda)\,\mathbb{1}_{\{\lambda<|x|\}}, \tag{3.17}$$

see Figure 1 for an illustration.

Note that solution (3.16) is given by shrinking the coordinates of $S_n(y)$ (which are associated with the Extreme-PLS estimator) towards zero. Besides, when the concentration parameter $\lambda = 0$, we recover the Extreme-PLS method. In the non-extreme case, a sparse extension of PLS was proposed in Chun and Keleş (2010), which uses the lasso to promote sparsity in the dimension reduction step.

3.4 Illustration on simulated data

Let us consider the simulation framework of Extreme-PLS (see Section 2.5) based on a sample of size $n = 1000$ and dimension $p = 30$ from model (**M**) with a power link function $g(t) = t^c$, $t > 0$, $c \in \{1/4, 1/2, 1, 3/2\}$. The behaviour of the Bayesian Extreme-PLS estimator is illustrated on this regression model where Y is heavy-tailed, with tail-index $\gamma = 1/5$, distributed from a Pareto distribution with survival function $\bar{F}(y) = (y/2)^{-5}$, $y \geq 2$. Each component $\varepsilon^{(j)}$, $j = 1, \ldots, p$ of the error ε is simulated from the $\mathcal{N}(0, \sigma^2)$ distribution and

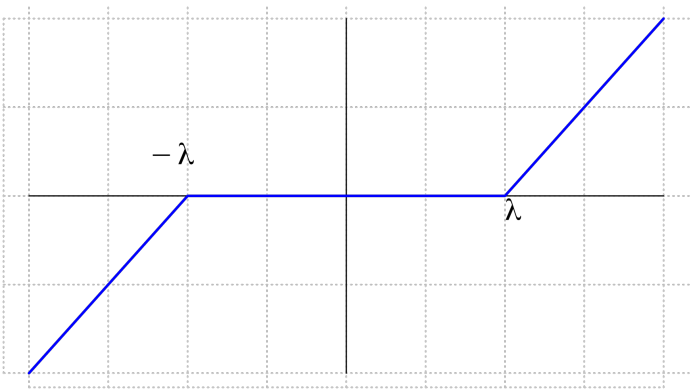


Figure 1 Plot of the shrinkage operator for a fixed value of λ.

depending on Y using Frank copula defined for all $(u, v) \in [0, 1]^2$ by

$$C_\theta(u, v) = -\frac{1}{\theta} \log \left(1 + \frac{(\mathrm{e}^{-\theta u} - 1)(\mathrm{e}^{-\theta v} - 1)}{\mathrm{e}^{-\theta} - 1} \right),$$

where $\theta \in \mathbb{R}$ is a parameter tuning the dependence between the margins. Frank copula is an Archimedean copula, see for instance (Nelsen, 2007, Section 4.2), able to model the full range of dependence: $\theta \to -\infty$ yields the counter-monotonicity copula, $\theta \to +\infty$ yields the co-monotonicity copula while $\theta = 0$ corresponds to independence. Here, we choose $\theta \in \{0, 10, 20\}$ corresponding to Kendall's $\tau \in \{0, 0.67, 0.82\}$. The standard deviation σ is selected such that the Signal to Noise Ratio (SNR) defined as SNR:= $g(\bar{F}^{-1}(1/n))/\sigma$ is equal to 10. Note that $g(\bar{F}^{-1}(1/n))$ represents the approximate maximum value of g on a n-sample from the distribution with associated survival function $\bar{F}$. We denote by $Y_{1,n} \leq Y_{2,n} \leq \cdots \leq Y_{n,n}$ the order statistics of the sample $(Y_1, \ldots, Y_n)$. The performance of the Bayesian Extreme-PLS estimator is investigated using the conjugate prior with $\beta = \beta_1 = (1, \ldots, 1, 0, \ldots, 0)^t/\sqrt{15}$ and the sparse prior with $\beta = \beta_2 = (1, 1, 0, \ldots, 0)^t/\sqrt{2}$.

Finally, the proximity between $\hat{\beta}_{MAP}(y)$ and another vector β computed on $N = 100$ replications is defined as follows:

$$\mathrm{PC}(y) = \frac{1}{N} \sum_{r=1}^{N} \cos^2 \left(\hat{\beta}^{(r)}_{MAP}(y), \beta \right). \tag{3.18}$$

The closer PC is to 1, the larger the proximity is.

Conjugate prior. The location parameter of the prior vMF/S distribution is set to $\mu = (1,\dots,1)^t/\sqrt{30}$, see Paragraph 3.3.2 for details.

In Figures 2–4, the proximity criterion $\mathrm{PC}(Y_{n-k+1,n})$ is displayed as a function of the number of exceedances $k \in \{1,\dots,200\}$, along various scenarios including: $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and β_1 (first column), $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and μ (second column), small to large concentrations $\kappa_1 \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$ (represented in each row), low to high dependence $\theta \in \{0, 10, 20\}$ (depicted in each figure).

Let us first recall that, in the top panels, corresponding to $\kappa_1 = 0$, we find back the Extreme-PLS estimator. In comparison, it appears that the Bayesian Extreme-PLS estimator $\hat{\beta}^C_{MAP}$ improves the estimation of β_1 for small numbers of exceedances (first column). Indeed, whatever the dependence coefficient θ and when the concentration increases $\kappa_1 \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, the Bayesian method yields accurate results ($\mathrm{PC}\geq 0.8$ for $c \geq 1$ and $\mathrm{PC}\geq 0.5$ for $c \leq 1/2$) for a wide range of choices of k, particularly for small values. In contrast, when the concentration parameter becomes large $\kappa = 10^{-2}$, the curves become smoother and PC tends to decrease. At the same time, we can see that the PC between $\hat{\beta}^C_{MAP}$ and μ (second column) gets closer to 1 as κ increases. This is due to the fact that larger values of κ_1 imply a stronger concentration around the prior direction μ (see Subsection 3.2.1).

Figures 5–7 provide another representation of the proximity criterion $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and β_1 as a function of both the number of exceedances $k \in \{1,\dots,200\}$ and the concentration parameter $\kappa_1 \in \{0, 10^{-4}, 2.10^{-4}, \dots, 10^{-2}\}$. The purple and blue bands represent respectively a $\mathrm{PC}\geq 0.8$ and $\mathrm{PC}\geq 0.6$. One can see that this band is quite large for a wide range of choices of κ_1 and k. When $\kappa_1 \to 0$, this implies that no prior is considered and that the EPLS estimator is recovered, while $\kappa_1 \to \infty$ implies that $\hat{\beta}^C_{MAP} \to \mu$. The Bayesian EPLS improves clearly the estimation of β, compared to EPLS, especially with exponent $c \leq 1/2$, for a well chosen κ_1 and k, i.e. not very large κ_1 and not very small k. Therefore, there is a trade-off between the choice of the concentration parameter and the number of exceedances.

Sparse prior. In Figures 8–10, the proximity criterion $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^S_{MAP}$ (see Paragraph 3.3.4) and β_2, and the norm $\|\beta_2\|_1$ are plotted respectively in the first and second column, as a function of the number of exceedances $k \in \{1,\dots,200\}$. The graphs are similar to the ones drawn before, except we use the concentration parameter $\lambda \in \{0, 10^{-5}, 10^{-4}, 10^{-3}\}$. In particular, top panels represent the Extreme-PLS estimator obtained when $\lambda = 0$. When the concentration parameter increases, $\lambda \in \{10^{-5}, 10^{-4}\}$, the Bayesian EPLS estimator $\hat{\beta}^S_{MAP}$ provides very good results, with a clear improvement in terms of proximity criterion ($\mathrm{PC}\geq 0.8$) for small values of k when $c \geq 1$ for all dependence situations. Good results ($\mathrm{PC}\geq 0.6$) can also be seen when $c = 1/2$ for well-chosen values of k. On the other hand, we can see that

when the concentration parameter is large ($\lambda = 10^{-3}$), the model still provides good results for exponents $c \geq 1$ for a limited choices of k, but seems to be very sensitive to small exponents such as $c = 1/4$. Indeed, the L_1 norm penalty in the prior enforces sparsity in the estimated direction. Therefore, when λ becomes very large, it is more likely that β coefficients are all equal to zero and then the proximity criterion gets closer to 1.
Figures 11–13 represent the proximity criterion $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^S_{MAP}$ and β_2 as a function of both the number of exceedances $k \in \{1, \ldots, 200\}$ and the concentration parameter $\lambda \in \{0, 10^{-5}, 2.10^{-5}, \ldots, 10^{-3}\}$. The introduction of the sparsity assumption in the prior improves the estimation of β compared to EPLS (large purple band for well-chosen λ and k).

3.5 Application to farm income modelling

In the context of farm income modelling, we aim to investigate the impact of various factors on low yields. Our approach is applied to data extracted from the FADN used in the numerical example of the Chapter 2. This dataset contains 949 farms described by 13 continuous attributes in 2014. Recall that the response variable Y is the inverse of the wheat yield (in quintals/hectare) and the covariate X includes 12 variables: selling prices (euro/quintal), pesticides, fertilizers, crop insurance purchase, insurance claims, farm subsidies, seeds and seedlings costs, works and services purchase for crops, other insurance premiums, farm income taxes, farmer's personal social security costs (euro/hectare) and temperature average (degree Celsius).
A number of checks have been carried out, in Section 2.6 of Chapter 2, to verify whether the heavy tail hypothesis of Y makes sense using the Hill plots and the quantile-quantile plots. It appears that the heavy-tailed assumption on Y is appropriate. Furthermore, the number of exceedances has been set at $k = 97$, since it corresponds to the highest correlation between $X^t\hat{v}(y)$ and Y, with $\hat{v}(y)$ the Extreme-PLS estimator (see Section 2.6 for more details). As the number of exceeds has been set, the choice of the concentration parameters κ_1 and λ, for both conjugate and sparse priors, remains to be discussed. We firstly compute the two MAP estimators $\beta^C_{MAP}(y)$ and $\beta^S_{MAP}(y)$ corresponding to conjugate and sparse priors, with $y = Y_{n-97+1,n}$, $\kappa_1 \in \{0, 10, \ldots, 3 \times 10^3\}$ and $\lambda \in \{0, 0.01, \ldots, 0.3\}$. Secondly, we define the following two conditional correlations:

$$\rho(X^t\hat{\beta}_{MAP}(y), Y|Y \geq y) = \frac{\mathrm{cov}(X^t\hat{\beta}_{MAP}(y), Y|Y \geq y)}{\sigma(X^t\hat{\beta}_{MAP}(y)|Y \geq y)\sigma(Y|Y \geq y)}, \tag{3.19}$$

$$\rho(X^t\hat{\beta}_{MAP}(y), X^{(j)}|Y \geq y) = \frac{\mathrm{cov}(X^t\hat{\beta}_{MAP}(y), X^{(j)}|Y \geq y)}{\sigma(X^t\hat{\beta}_{MAP}(y)|Y \geq y)\sigma(X^{(j)}|Y \geq y)}. \tag{3.20}$$

The results are depicted on Figures 14 and 15. For the conjugate prior, the location parameter of the vMF/S distribution is set to $\mu = (0, 0, 1, 0, \ldots, 0, 1)^t/\sqrt{2}$. Indeed, according to the agricultural expert, it has been observed that wheat yields are affected by financial market prices and weather conditions such as temperature. Therefore, we have assigned a weight of 1 to these two variables. The top panel of Figure 14 displays the conditional correlation (3.19) between the projected covariate $X^t\hat{\beta}^C_{MAP}(y)$ and the response variable Y. It appears that introducing the prior information on the data, the correlation increases from 0.59 to 0.64 for a concentration parameter ranging from 0 to 220. While for the sparse prior, the correlation between $X^t\hat{\beta}^S_{MAP}(y)$ and Y, represented in the top of Figure 15, remains stable with respect to the choice of the concentration until $\lambda = 0.16$ and then it drops to 0 since $\hat{\beta}^S_{MAP}(y) = 0$ in this area. The coordinates of $|\hat{\beta}^S_{MAP}(y)|$ are represented in the bottom of Figure 15 as functions of the concentration λ. One can see on the regularization path that when the concentration parameter λ increases, the coordinates of $\hat{\beta}^S_{MAP}(y)$ gradually decrease towards 0. Low crop yields are mainly correlated with farm subsidies. In the following, we only consider the MAP estimator $\hat{\beta}^C_{MAP}(y)$ corresponding to the conjugate prior as it gives the highest correlation between the projected covariate and Y.

The conditional correlation (3.20) between the projected covariate $X^t\hat{\beta}^C_{MAP}(y)$ and each coordinate $X^{(j)}$ of the covariate is presented in the bottom panel of Figure 14. When $\kappa_1 = 220$ fixed, small yields are mainly related to agricultural inputs (fertilisers, pesticides, seeds and seedlings, purchases of works and services, personal social charges) and risk management (claims, purchase of crop insurance, agricultural subsidies). Indeed, yields were strongly impacted by agricultural inputs in 2014, despite mild winter temperatures. When κ_1 increases, becomes very large, one can see that the effect of all the variables mentioned previously decreases and the effect of the price and temperature variables increase. This is reasonable, as larges values of the concentration parameter yield $\hat{\beta}^C_{MAP} \to \mu$.

In the following, we select $\kappa_1 = 220$ which corresponds to the highest correlation between $X^t\hat{\beta}^C_{MAP}(y)$ and Y. The projected scatter plot $\left(Y_i, \hat{\beta}^C_{MAP}(y)^t X_i\right)$, $i = 1, \ldots, n$ is displayed in a logarithmic scale for the visualization sake on the top panel of Figure 16 together with two estimations (linear and non-linear) of the conditional mean $\mathbb{E}(\hat{\beta}^C_{MAP}(y)^t X|Y)$. The conditional quantiles $\hat{q}(\alpha|\hat{\beta}^C_{MAP}(y)^t X)$, computed through a kernel estimator of the conditional survival function, are reported in the bottom panel of Figure 16 together with the scatter plot $\left(\hat{\beta}^C_{MAP}(y)^t X_i, Y_i\right)$, $i = 1, \ldots, n$. The vertical and horizontal axes are represented in a logarithmic scale. One can see that both curves of the conditional quantiles corresponding to levels $\alpha = 0.15$ (blue line) and $\alpha = 0.05$ (red line) behave in a same way. The estimated conditional quantiles of inverse yields show an increasing trend for $\log(\hat{\beta}^C_{MAP}(y)^t X) \leq 3.6$: low yields are linked to agricultural inputs and risk management. For $\log(\hat{\beta}^C_{MAP}(y)^t X) > 3.6$,

the interpretation of the results becomes difficult, as the estimation is not reliable for large values of the covariate, due to the scarcity of data in this area and the boundary effects of kernel estimators.

3.6 Discussion

In this work, we propose a Bayesian approach for the Extreme-PLS model to identify the direction of dimension reduction and introduce prior information about it. We present some criteria for choosing a prior distribution, such as incorporating the sparsity of the directions with the Bayesian Lasso prior. Numerical examples show that the proposed method is particularly effective for problems with very small data sets. In future work, the posterior distribution with the hierarchical prior (see Section 3.3.3) can be computed directly by Importance Sampling or Monte-Carlo Markov Chains sampling. Furthermore, it would be interesting to test other priors such as an uninformative distribution (Jeffreys, 1946) or other shrinkage priors that aim to reduce small effects towards zero, e.g. ridge or elastic-net priors (Van Erp et al., 2019). It would also be interesting to see how the introduction of prior information on the dimension reduction direction could improve the estimation of extreme conditional quantiles on small samples.

3.A Appendix: proofs

Proof of Lemma 3.2.1. (i) The mode of von Mises-Fisher distribution is given by the argmax of the probability density function. Thus, the constrained optimization problem is:

$$\hat{x} = \underset{\|x\|=1}{\operatorname{argmax}}\, e^{\kappa \mu^t x},$$

which is equivalent to

$$\hat{x} = \underset{\|x\|=1}{\operatorname{argmax}}\, \kappa \mu^t x,$$

and can be solved using Lagrange multipliers method by introducing:

$$\mathcal{L}(\hat{x}, \lambda) = \kappa \mu^t \hat{x} - \frac{\lambda}{2}(||\hat{x}||^2 - 1), \quad \lambda \in \mathbb{R},$$

and setting the partial derivatives to zero yield:

$$\begin{cases} \nabla_\lambda \mathcal{L}(\hat{x}, \lambda) & = -\frac{1}{2}(\|\hat{x}\|^2 - 1) = 0, \\ \nabla_{\hat{x}} \mathcal{L}(\hat{x}, \lambda) & = \kappa\mu - \lambda\hat{x} = 0, \end{cases}$$

or equivalently,

$$\begin{cases} \|\hat{x}\|^2 &= 1, \\ \hat{x} &= \kappa\mu/\lambda. \end{cases}$$

It straightforwardly follows that $\lambda = \kappa$ and $\hat{x} = \mu$.

(ii) For the expectation, we have:

$$\mathbb{E}(X) = c_p(\kappa) \int_{S^{p-1}} x e^{\kappa\mu^t x} dx.$$

Let $b \in \mathbb{R}^p$ an arbitrary test vector and consider

$$b^t\mathbb{E}(X) = c_p(\kappa) b^t \int_{S^{p-1}} x e^{\kappa u^t x} dx = c_p(\kappa) \int_{S^{p-1}} \sum_{j=1}^{p} x_j (b_j e^{\kappa u^t x}) dx.$$

Now, the divergence theorem allows to transform the surface integral over S^{p-1} into a volume integral over $B^p(1)$:

$$\begin{aligned} b^t\mathbb{E}(X) &= c_p(\kappa) \int_{B^p(1)} \sum_{j=1}^{p} \frac{\partial}{\partial x_j} \left(b_j e^{\kappa u^t x} \right) dx \\ &= c_p(\kappa) \int_{B^p(1)} \sum_{j=1}^{p} b_j \kappa \mu_j e^{\kappa u^t x} dx \\ &= c_p(\kappa) \int_{B^p(1)} \kappa b^t \mu e^{\kappa u^t x} dx \\ &= \kappa b^t \mu c_p(\kappa) \int_{B^p(1)} e^{\kappa u^t x} dx. \end{aligned}$$

Using the normalization condition for a vMF distribution over the unit hyperball $B^p(1)$, defined in Section 3.2.2, and from (3.9) we have,

$$\int_{B^p(1)} e^{\kappa\mu^t x} dx = \frac{1}{c'_p(\kappa)} = \frac{1}{2\pi c_{p+2}(\kappa)},$$

and taking account of (3.2) yields

$$b^t\mathbb{E}(X) = b^t\mu \frac{\kappa c_p(\kappa)}{2\pi c_{p+2}(\kappa)} = b^t\mu \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)}.$$

Since the test vector b is arbitrary, it follows

$$\mathbb{E}(X) = \mu \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)} = \mu A_p(\kappa),$$

and the result is proved. □

Proof of Lemma 3.2.2. The normalizing constant $c'_p(\kappa)$ is derived using the fact that:

$$1/c'_p(\kappa) = \int_{B^p(r)} \frac{1}{r^p} e^{\kappa \mu^t x / r} dx = \int_{B^p(1)} e^{\kappa \mu^t x} dx.$$

Two successive changes of variable yield

$$\begin{aligned}
1/c'_p(\kappa) &= \int_{B^p(1)} e^{\kappa \mu^t x} dx \\
&= \int_0^1 \rho^{p-1} \int_{S^{p-1}} e^{(\rho\kappa)\mu^t u} du d\rho, \\
&= \int_0^1 \frac{\rho^{p-1}}{c_p(\rho\kappa)} d\rho \\
&= \frac{(2\pi)^{p/2}}{\kappa^{p/2-1}} \int_0^1 \rho^{p/2} I_{p/2-1}(\rho\kappa) d\rho \\
&= \frac{(2\pi)^{p/2}}{\kappa^p} \int_0^\kappa t^{p/2} I_{p/2-1}(t) dt.
\end{aligned}$$

From definition (3.3) of the modified Bessel function and using the fact that $I_p(t)$ is a power series of infinite radius of convergence, we have:

$$\begin{aligned}
\int_0^\kappa t^{p+1} I_p(t) dt &= \sum_{l \geq 0} \frac{1}{\Gamma(p+l+1) l!} \times \frac{1}{2^{2l+p}} \int_0^\kappa t^{2l+2p+1} dt \\
&= \sum_{l \geq 0} \frac{1}{\Gamma(p+l+1) l!} \times \frac{1}{2^{2l+p}} \times \frac{\kappa^{2l+2p+2}}{2(l+p+1)},
\end{aligned}$$

and using the fact that $\Gamma(p+l+2) = (p+l+1)\Gamma(p+l+1)$, we get

$$\int_0^\kappa t^{p+1} I_p(t) dt = \kappa^{p+1} \sum_{l \geq 0} \frac{1}{\Gamma(p+l+2) l!} \left(\frac{\kappa}{2}\right)^{2l+p+1} = \kappa^{p+1} I_{p+1}(\kappa).$$

It follows that, $\forall p \geq 2$:

$$\int_0^\kappa t^{p/2} I_{p/2-1}(t) dt = \kappa^{p/2} I_{p/2}(\kappa),$$

leading to

$$1/c'_p(\kappa) = \frac{(2\pi)^{p/2}}{\kappa^{p/2}} I_{p/2}(\kappa),$$

or, equivalently,

$$c'_p(\kappa) = 2\pi c_{p+2}(\kappa),$$

which concludes the proof. □

Proof of Proposition 3.3.1. The optimization problem can be rewritten as

$$\begin{aligned}\hat{\beta}^S_{MAP}(y) &= \underset{\|\beta\|^2=1}{\operatorname{argmin}}\, \lambda\|\beta\|_1 - \beta^t S_n(y) \\ &= \underset{\|\beta\|^2=1}{\operatorname{argmin}} \sum_{j=1}^{p} \lambda|\beta_j| - \beta_j S_{n,j}(y) \\ &= \underset{\|\beta\|^2=1}{\operatorname{argmin}} \sum_{j=1}^{p} |\beta_j|\left(\lambda - \operatorname{sign}(\beta_j) S_{n,j}(y)\right).\end{aligned}$$

Introducing $b_j = |\beta_j|$ and $s_j = \operatorname{sign}(\beta_j)$, the above optimization problem can be rewritten as

$$\hat{\beta}^S_{MAP}(y) = \underset{b,s}{\operatorname{argmin}} \left\{ \sum_{j=1}^{p} b_j(\lambda - s_j S_{n,j}(y)) \;\text{ s.t. }\; \|b\|^2 = 1,\; b_j \geq 0,\; |s_j| = 1,\; j = 1,\dots,p \right\}.$$

Clearly, the solution w.r.t. s is given by $s_j = \operatorname{sign}(S_{n,j})$ for all $j = 1,\dots,p$ and therefore

$$\hat{\beta}^S_{MAP}(y) = \underset{b}{\operatorname{argmin}} \left\{ \sum_{j=1}^{p} b_j(\lambda - |S_{n,j}(y)|) \;\text{ s.t. }\; \|b\|^2 = 1,\; b_j \geq 0,\; j = 1,\dots,p \right\}.$$

The Lagrangian is given by

$$\mathcal{L}(b, \alpha_1, \dots, \alpha_{p+1}) = -\sum_{j=1}^{p} b_j\left(\lambda - |S_{n,j}(y)|\right) + \sum_{j=1}^{p} \alpha_j b_j + \alpha_{p+1}(\|b\|^2 - 1),$$

with associated Karush-Kuhn-Tucker conditions

$$\begin{cases} |S_{n,j}(y)| - \lambda + \alpha_j + 2\alpha_{p+1}\hat{b}_j = 0, \\ \alpha_j \hat{b}_j = 0,\; \alpha_j \geq 0, \quad j = 1,\dots,p, \\ \|\hat{b}\|^2 - 1 = 0. \end{cases} \tag{3.21}$$

Multiplying the first equation in (3.21) by $\hat{b}_j$ and summing yield

$$\alpha_{p+1} = \frac{1}{2}\sum_{j=1}^{p} \hat{b}_j(\lambda - |S_{n,j}(y)|). \tag{3.22}$$

Then, two cases arise. (i) If $\lambda < |S_{n,j}(y)|$, then one can fix $\alpha_j = 0$ and (3.21) implies $\hat{b}_j = (\lambda - |S_{n,j}(y)|)/(2\alpha_{p+1})$. (ii) Conversely, if $\lambda \geq |S_{n,j}(y)|$, then one can fix $\hat{b}_j = 0$ and (3.21) implies $\alpha_j = \lambda - |S_{n,j}(y)| \geq 0$. Replacing in (3.22) yields $2\alpha^2_{p+1} = \|\hat{b}\|^2$. Taking account of $\hat{b}_j \geq 0$, this implies $\alpha_{p+1} = -\sqrt{2}\|\hat{b}\|$. Summarizing, one has $\hat{\beta}^S_{MAP}(y) = \tilde{\beta}(y)/\|\tilde{\beta}(y)\|$ with

$$\tilde{\beta}_j(y) = \operatorname{sign}(S_{n,j}(y))\left(|S_{n,j}(y)| - \lambda\right)\mathbb{1}_{\{\lambda < |S_{n,j(y)}|\}}$$

for all $j = 1, \ldots, p$ and the result is proved. □

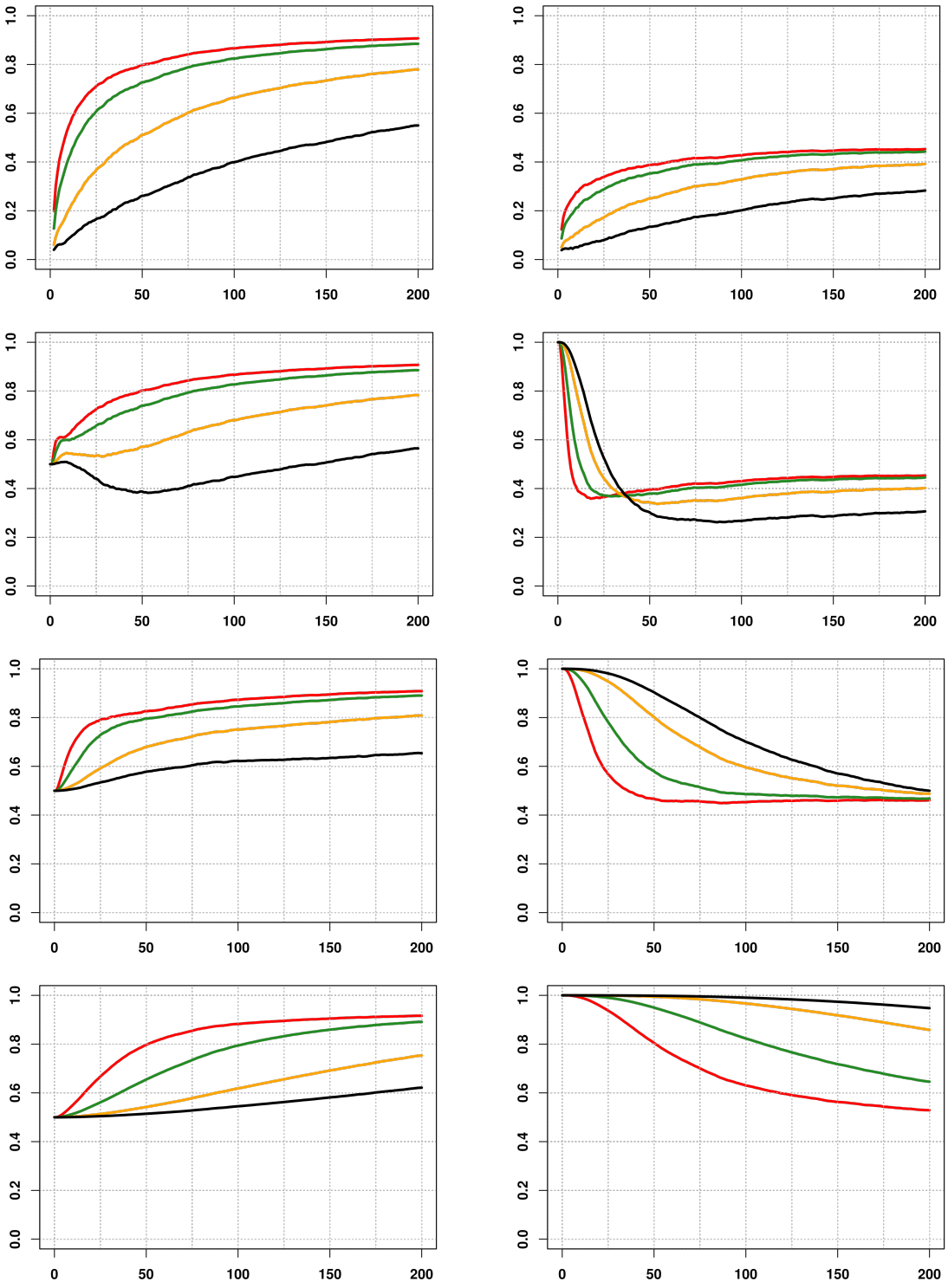


Figure 2 Finite sample behaviour of Bayesian EPLS, with conjugate prior, on simulated data from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a Frank copula parameter $\theta = 0$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{C}_{MAP}$ and β_1 (left) and $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{C}_{MAP}$ and μ (right) as a function of the number $k \in \{1, ..., 200\}$. From top to bottom, concentration parameter $\kappa_1 \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

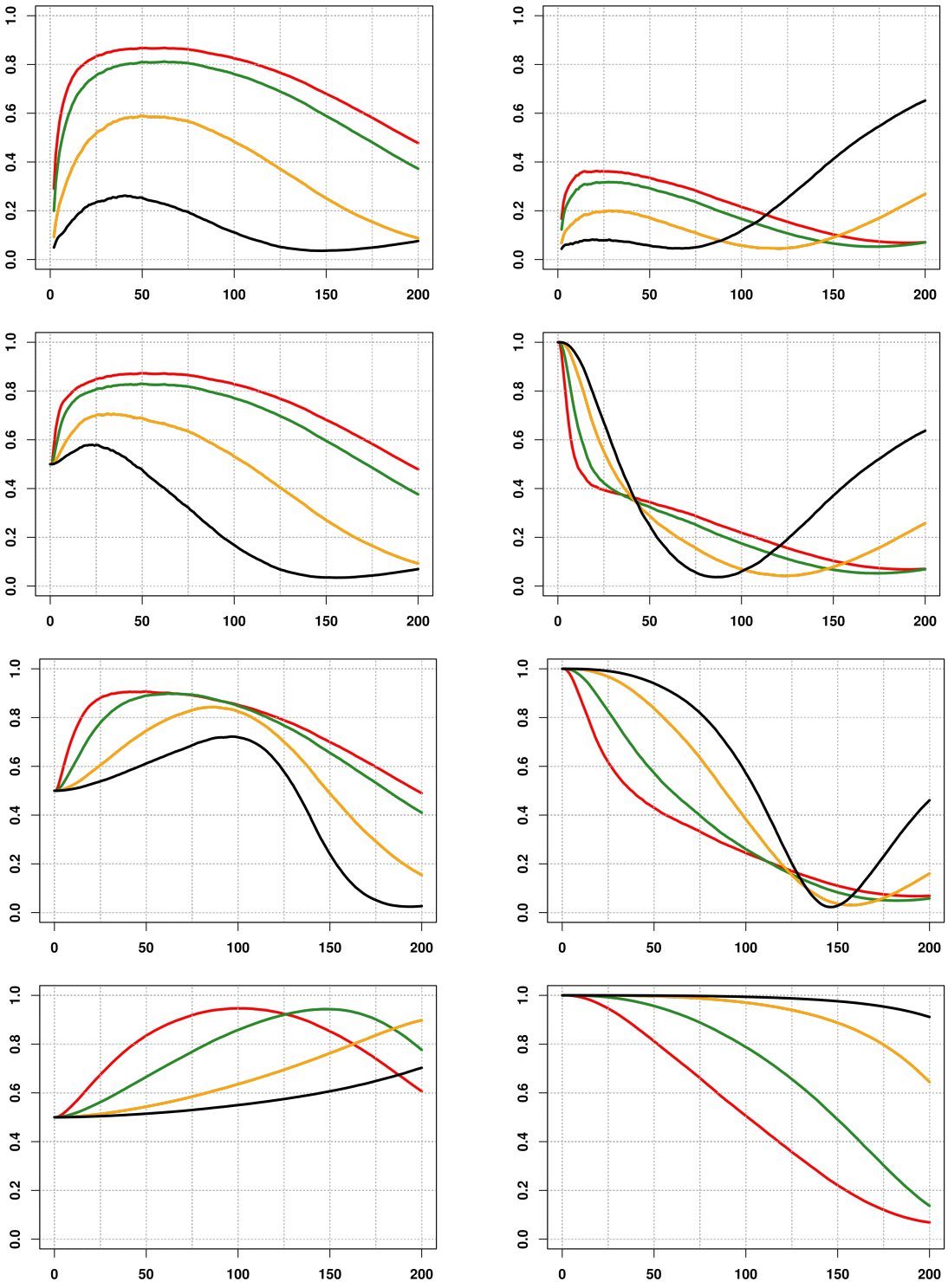


Figure 3 Finite sample behaviour of Bayesian EPLS on simulated data, with conjugate prior, from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a Frank copula parameter $\theta = 10$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and β_1 (left) and $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and μ (right) as a function of the number $k \in \{1, ..., 200\}$. From top to bottom, concentration parameter $\kappa_1 \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

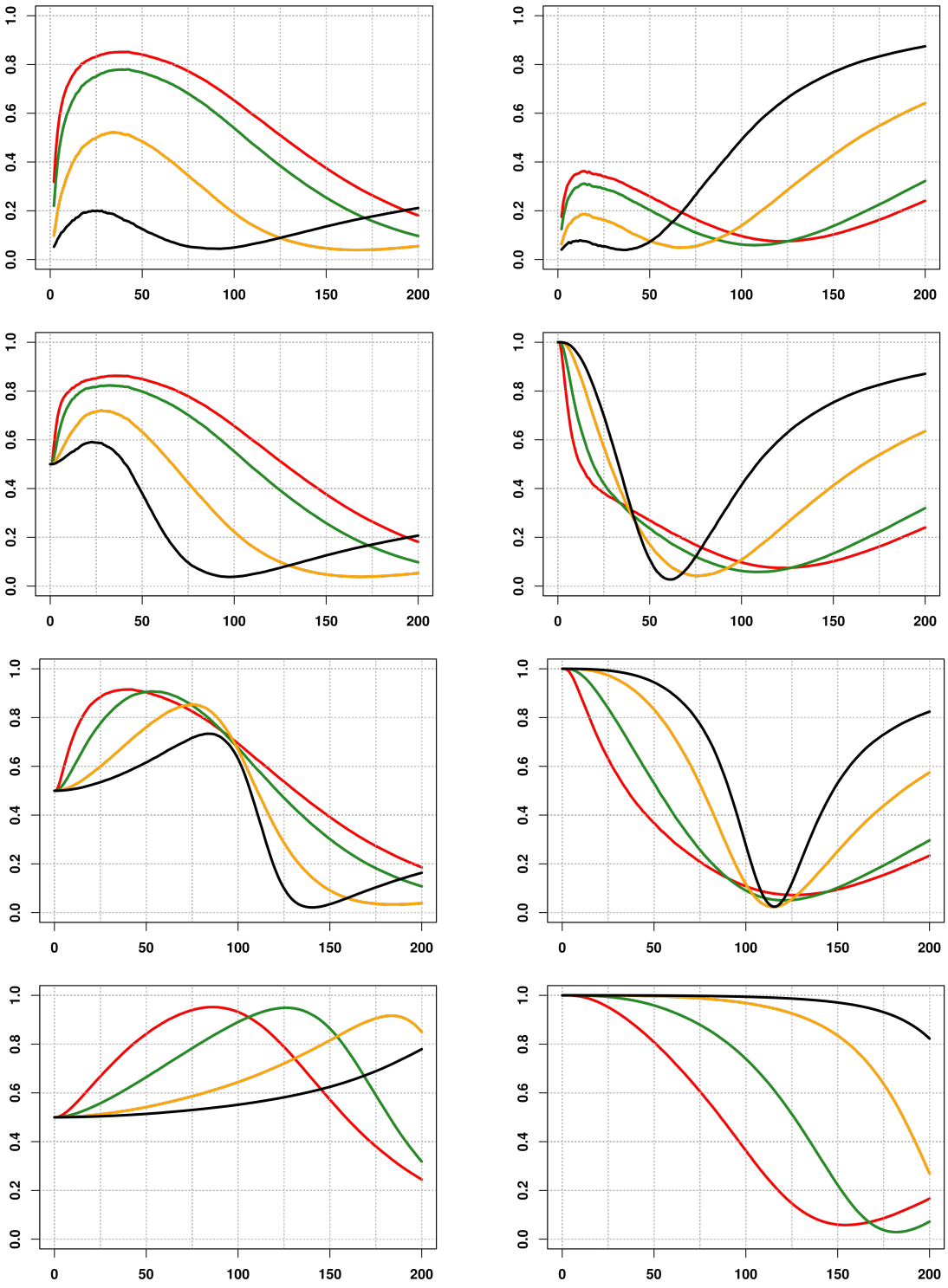


Figure 4 Finite sample behaviour of Bayesian EPLS, with conjugate prior, on simulated data from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a Frank copula parameter $\theta = 20$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and β_1 (left) and $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and μ (right) as a function of the number $k \in \{1, ..., 200\}$. From top to bottom, concentration parameter $\kappa_1 \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

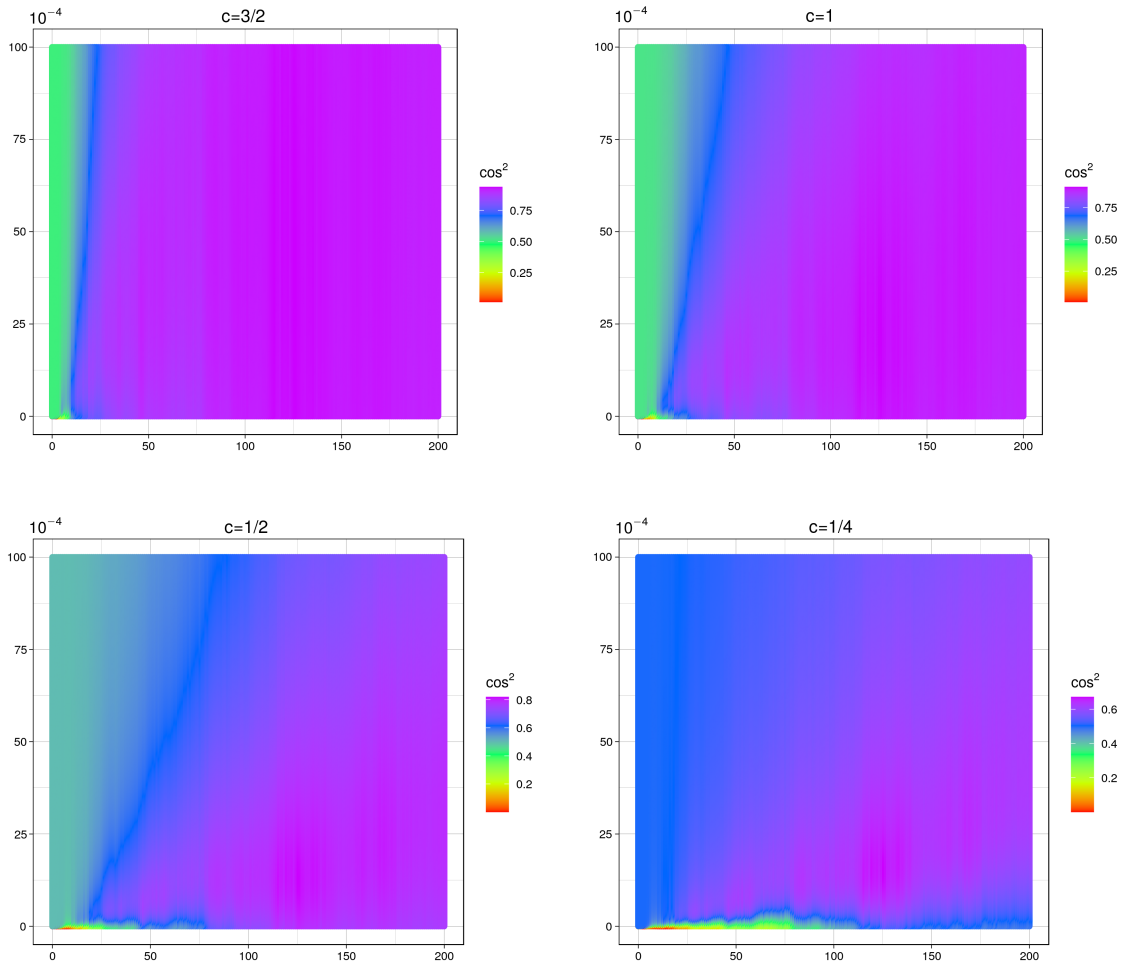


Figure 5 Finite sample behaviour of Bayesian EPLS, with conjugate prior, on simulated data from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a Frank copula parameter $\theta = 0$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and β_1 as a function of the number $k \in \{1, ..., 200\}$ and the concentration parameter $\kappa_1 \in \{0, 10^{-4}, 2.10^{-4}, \ldots, 10^{-2}\}$. From top left to bottom right, the powers $c \in \{3/2, 1, 1/2, 1/4\}$ of the link function $g(t) = t^c$.

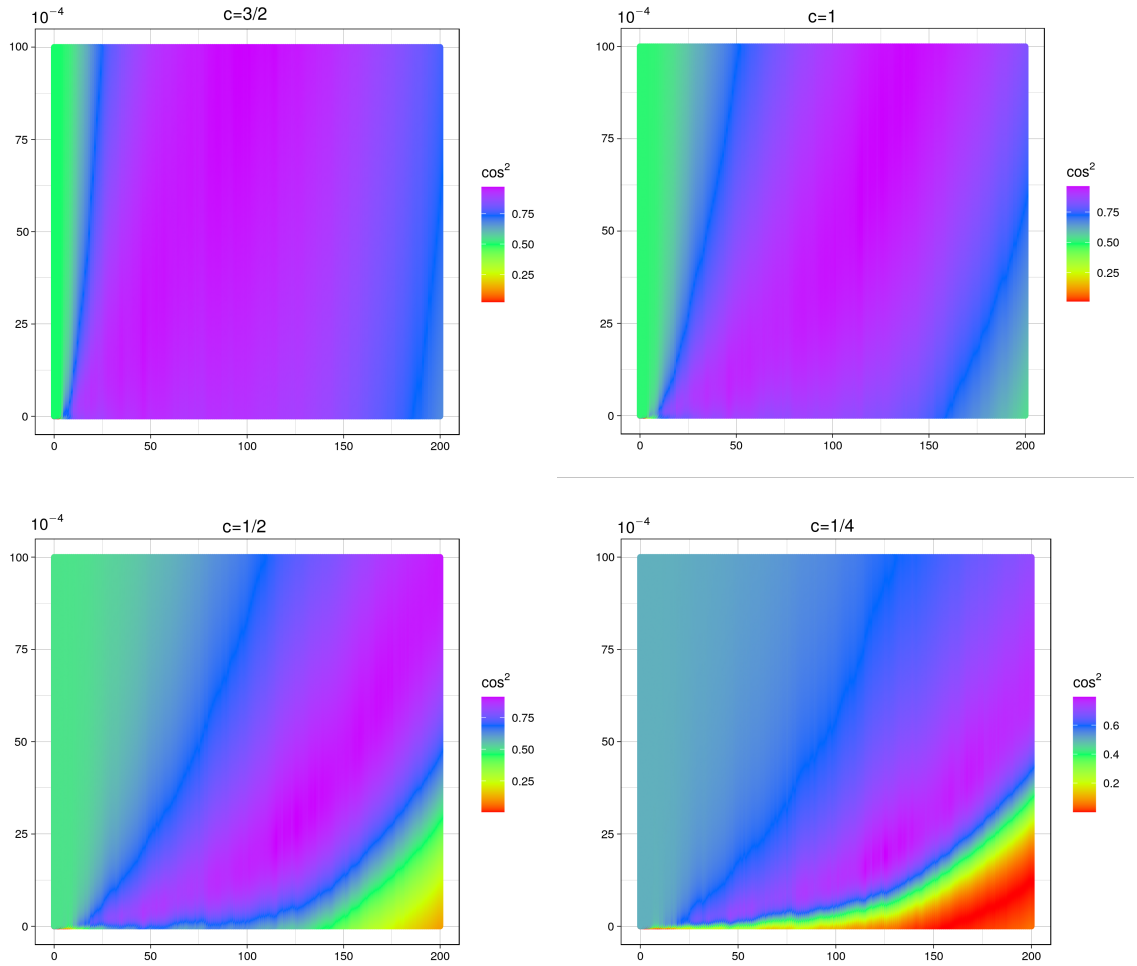


Figure 6 Finite sample behaviour of Bayesian EPLS, with conjugate prior, on simulated data from a Pareto distribution ($\gamma = 1/5, a = 2$) and a Frank copula parameter $\theta = 10$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and β_1 as a function of the number $k \in \{1, ..., 200\}$ and the concentration parameter $\kappa_1 \in \{0, 10^{-4}, 2.10^{-4}, \ldots, 10^{-2}\}$. From top left to bottom right, the powers $c \in \{3/2, 1, 1/2, 1/4\}$ of the link function $g(t) = t^c$.

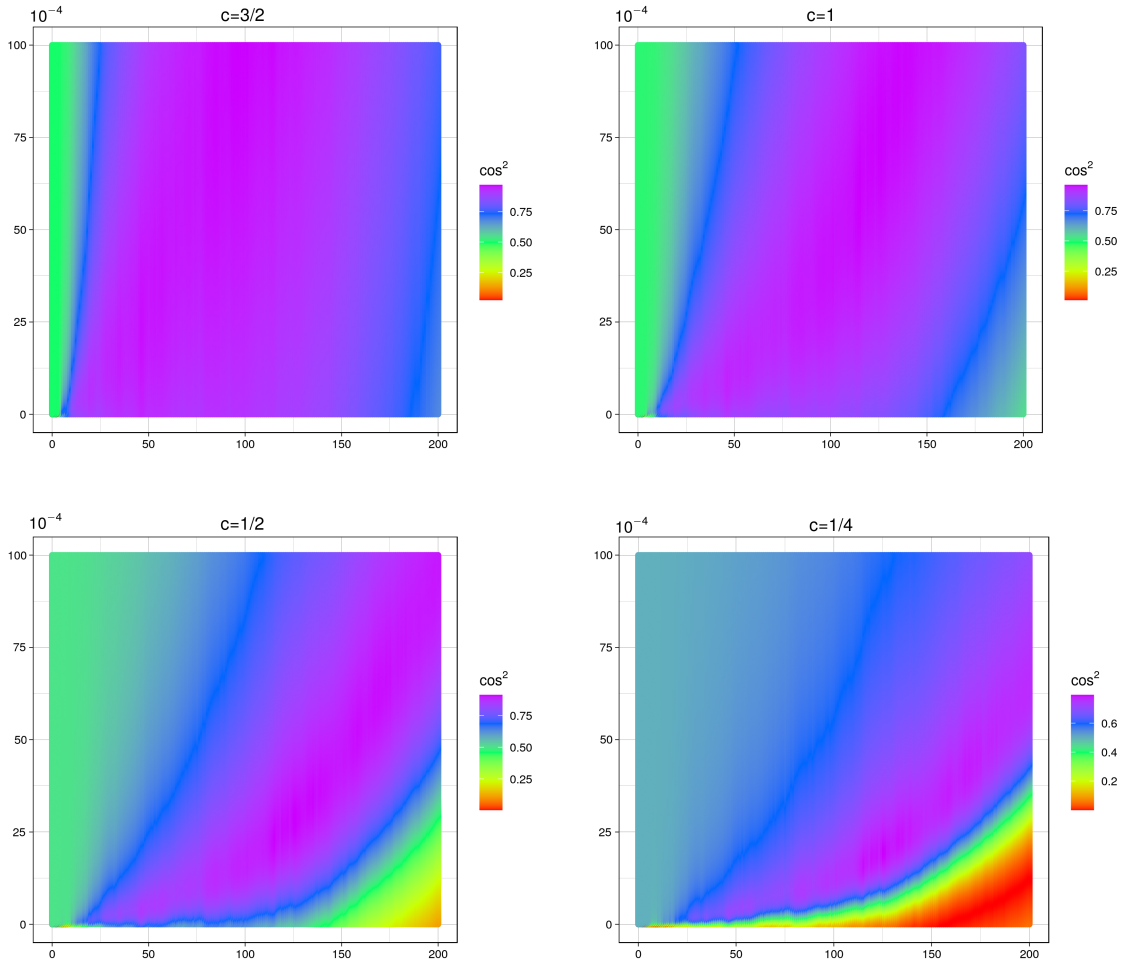


Figure 7 Finite sample behaviour of Bayesian EPLS, with conjugate prior, on simulated data from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a Frank copula parameter $\theta = 20$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^C_{MAP}$ and β_1 as a function of the number $k \in \{1, ..., 200\}$ and the concentration parameter $\kappa_1 \in \{0, 10^{-4}, 2.10^{-4}, \ldots, 10^{-2}\}$. From top left to bottom right, the powers $c \in \{3/2, 1, 1/2, 1/4\}$ of the link function $g(t) = t^c$.

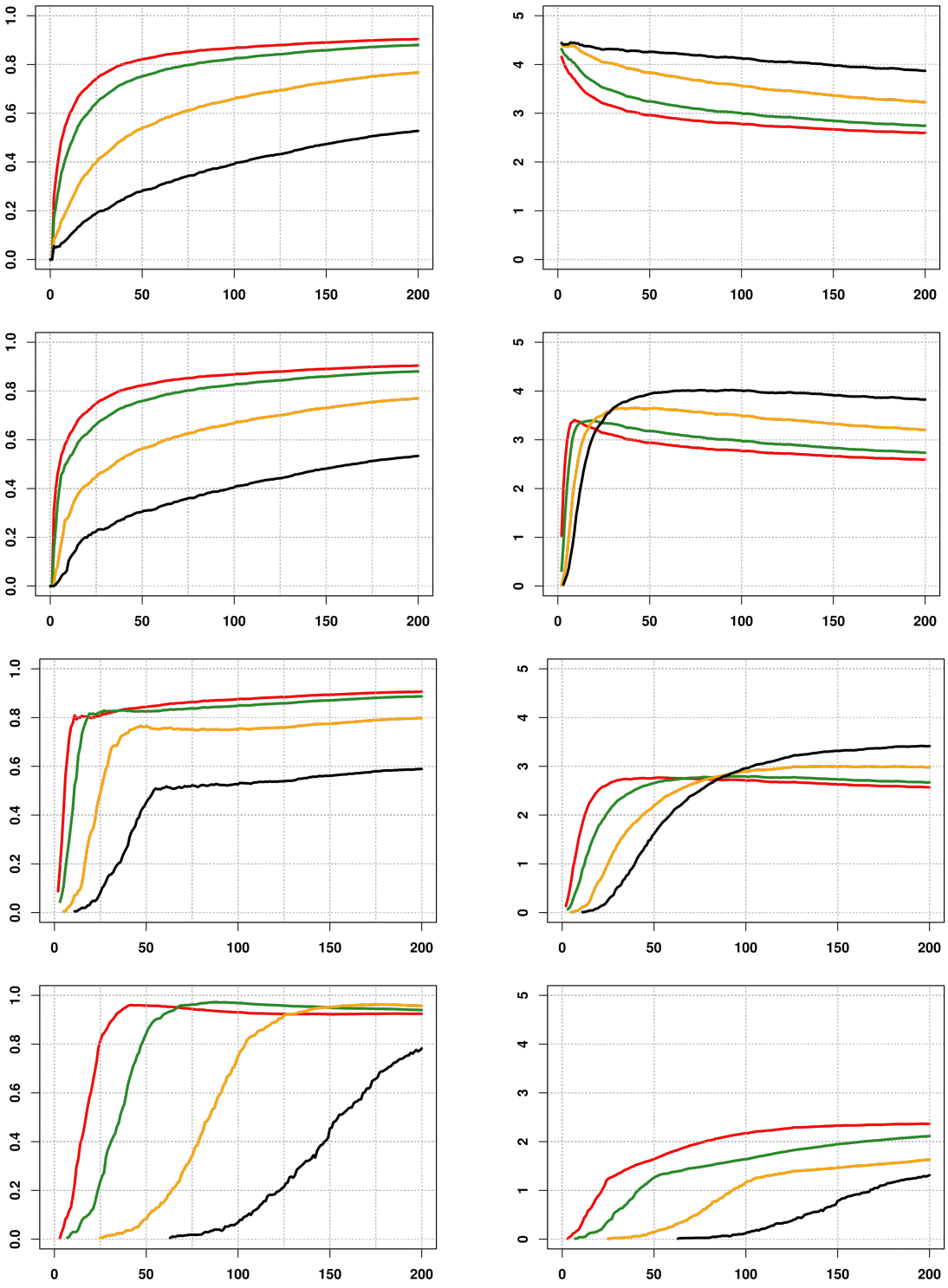


Figure 8 Finite sample behaviour of Bayesian EPLS, with sparse prior, on simulated data from a Pareto distribution ($\gamma = 1/5$, $a = 2$) and a Frank copula parameter $\theta = 0$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^S_{MAP}$ and β_2 (left) and $\|\beta_2\|_1$ (right) as a function of the number $k \in \{1, ..., 200\}$. From top to bottom, concentration parameter $\lambda \in \{0, 10^{-5}, 10^{-4}, 10^{-3}\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

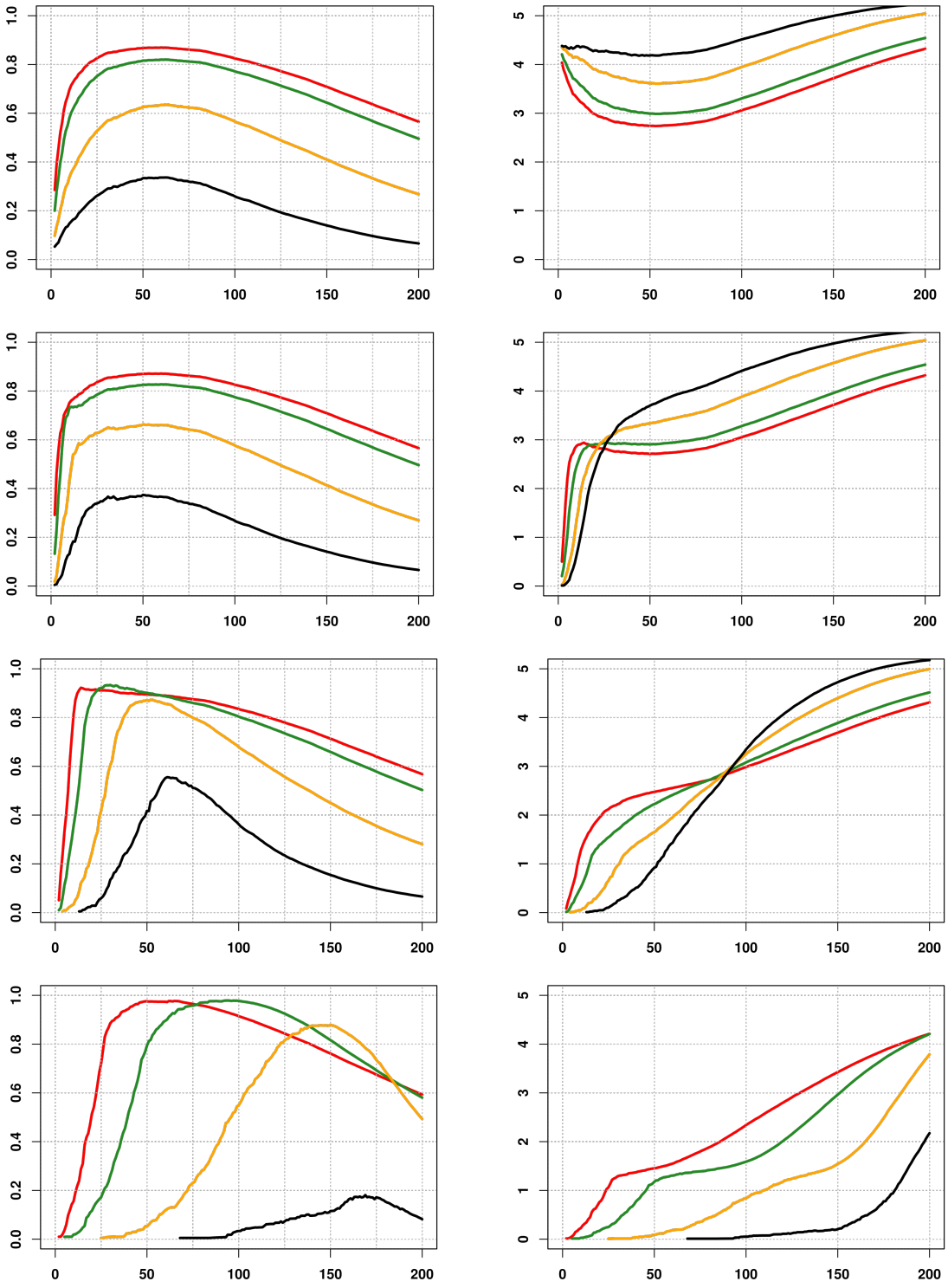


Figure 9 Finite sample behaviour of Bayesian EPLS, with sparse prior, on simulated data from a Pareto distribution ($\gamma = 1/5$, $a = 2$) and a Frank copula parameter $\theta = 10$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^S_{MAP}$ and β_2 (left) and $\|\beta_2\|_1$ (right) as a function of the number $k \in \{1, ..., 200\}$. From top to bottom, concentration parameter $\lambda \in \{0, 10^{-5}, 10^{-4}, 10^{-3}\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

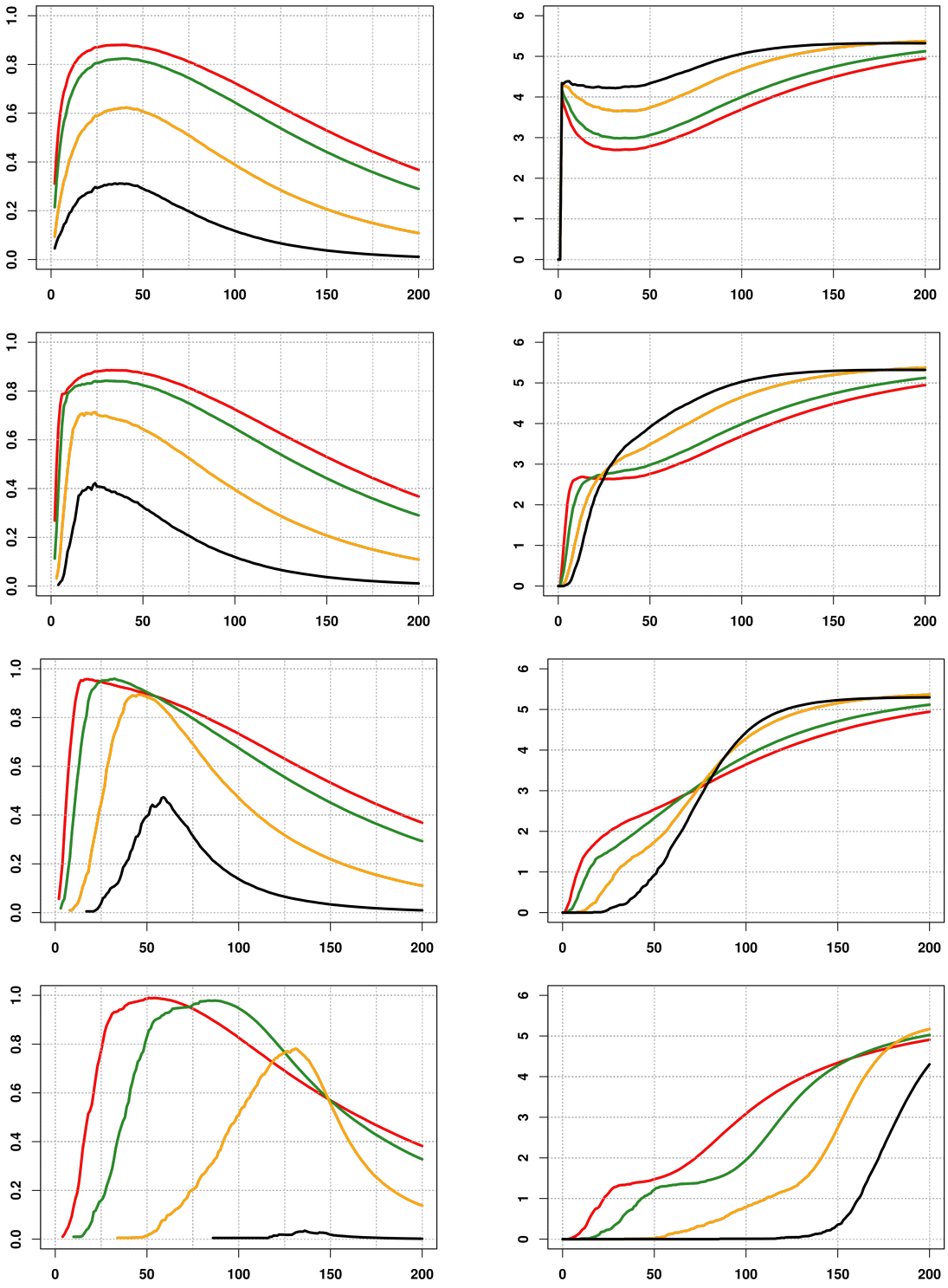


Figure 10 Finite sample behaviour of Bayesian EPLS, with sparse prior, on simulated data from a Pareto distribution ($\gamma = 1/5$, $a = 2$) and a Frank copula parameter $\theta = 20$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^S_{MAP}$ and β_2 (left) and $\|\beta_2\|_1$ (right) as a function of the number $k \in \{1, ..., 200\}$. From top to bottom, concentration parameter $\lambda \in \{0, 10^{-5}, 10^{-4}, 10^{-3}\}$. The powers $c \in \{1/4, 1/2, 1, 3/2\}$ of the link function $g(t) = t^c$ are displayed in {black, yellow, green, red}.

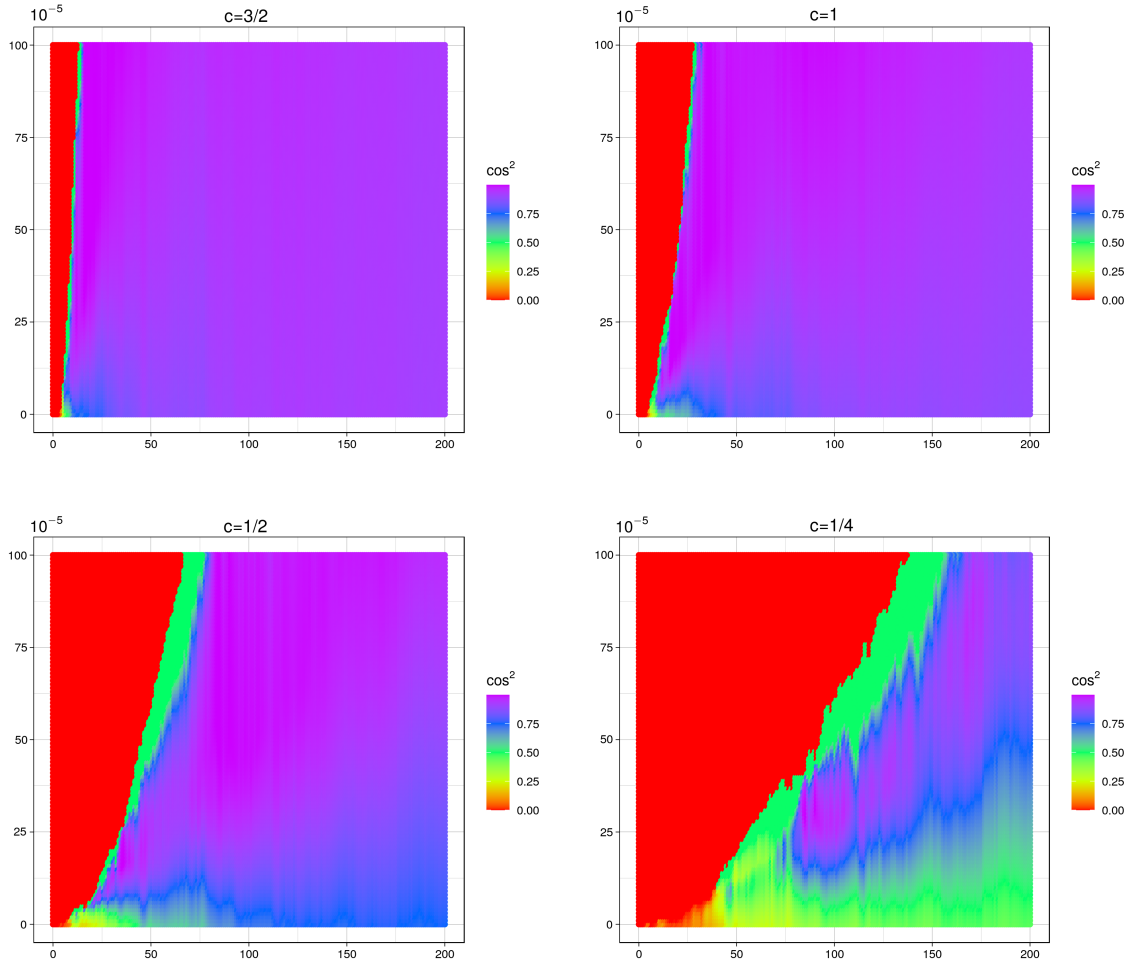


Figure 11 Finite sample behaviour of Bayesian EPLS, with sparse prior, on simulated data from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a Frank copula parameter $\theta = 0$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^S_{MAP}$ and β_2 as a function of the number $k \in \{1, ..., 200\}$ and concentration parameter $\lambda \in \{0, 10^{-5}, 2.10^{-5}, \ldots, 10^{-3}\}$. From top left to bottom right, the powers $c \in \{3/2, 1, 1/2, 1/4\}$ of the link function $g(t) = t^c$.

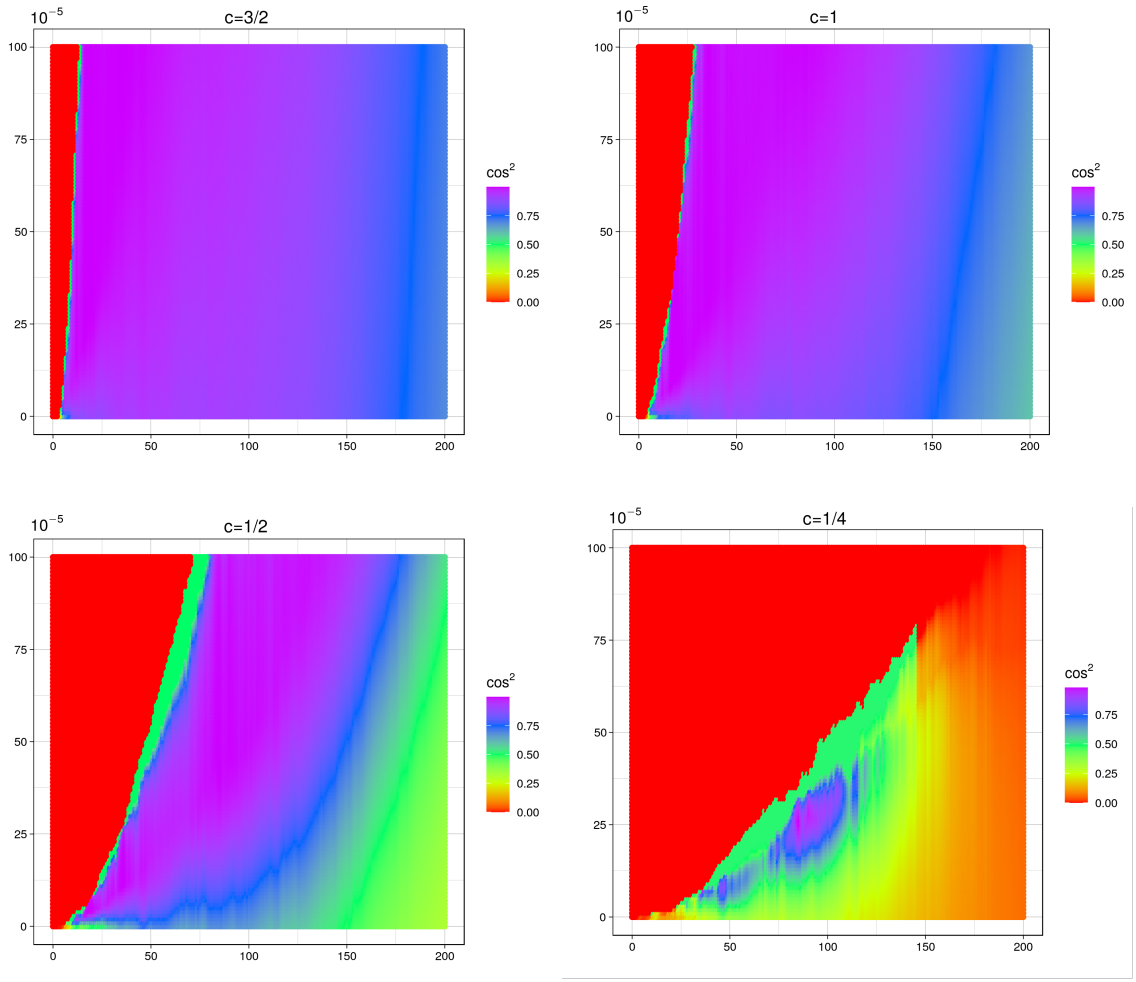


Figure 12 Finite sample behaviour of Bayesian EPLS, with sparse prior, on simulated data from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a Frank copula parameter $\theta = 10$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^S_{MAP}$ and β_2 as a function of the number $k \in \{1, ..., 200\}$ and concentration parameter $\lambda \in \{0, 10^{-5}, 2.10^{-5}, \ldots, 10^{-3}\}$. From top left to bottom right, the powers $c \in \{3/2, 1, 1/2, 1/4\}$ of the link function $g(t) = t^c$.

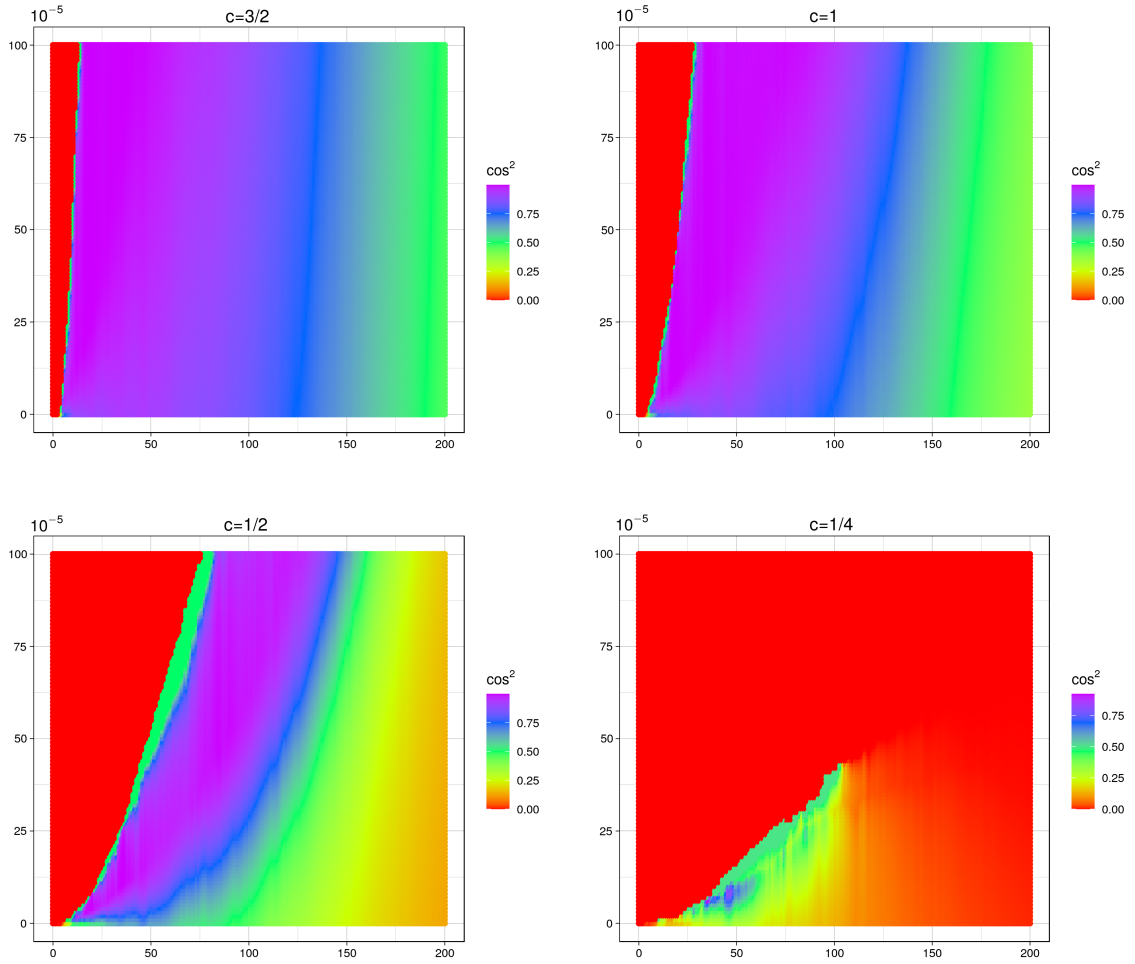


Figure 13 Finite sample behaviour of Bayesian EPLS, with sparse prior, on simulated data from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a Frank copula parameter $\theta = 20$. $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^S_{MAP}$ and β_2 as a function of the number $k \in \{1, ..., 200\}$ and concentration parameter $\lambda \in \{0, 10^{-5}, 2.10^{-5}, \ldots, 10^{-3}\}$. From top left to bottom right, the powers $c \in \{3/2, 1, 1/2, 1/4\}$ of the link function $g(t) = t^c$.

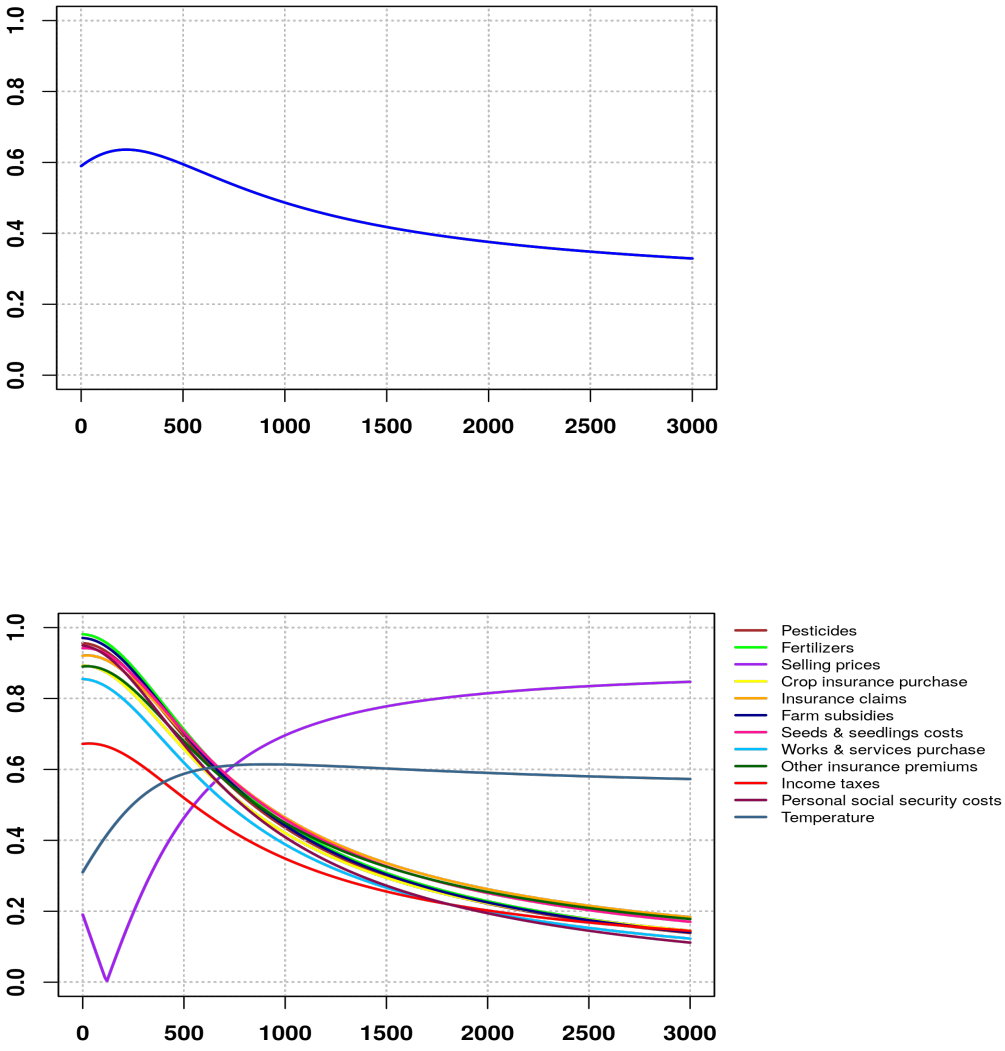


Figure 14 Farm income data with conjugate prior. Top: Graph of the estimated conditional correlation $\hat{\rho}(X^t\hat{\beta}^C_{MAP}(y), Y|Y \geq y)$ with $y = Y_{n-97+1,n}$. Bottom: Graph of the estimated conditional correlation $\hat{\rho}(X^t\hat{\beta}^C_{MAP}(y), X^{(j)}|Y \geq y)$ with $y = Y_{n-97+1,n}$ for $j = 1, \ldots, 12$ (horizontally: concentration κ_1, vertically: conditional correlation estimated by its empirical counterpart).

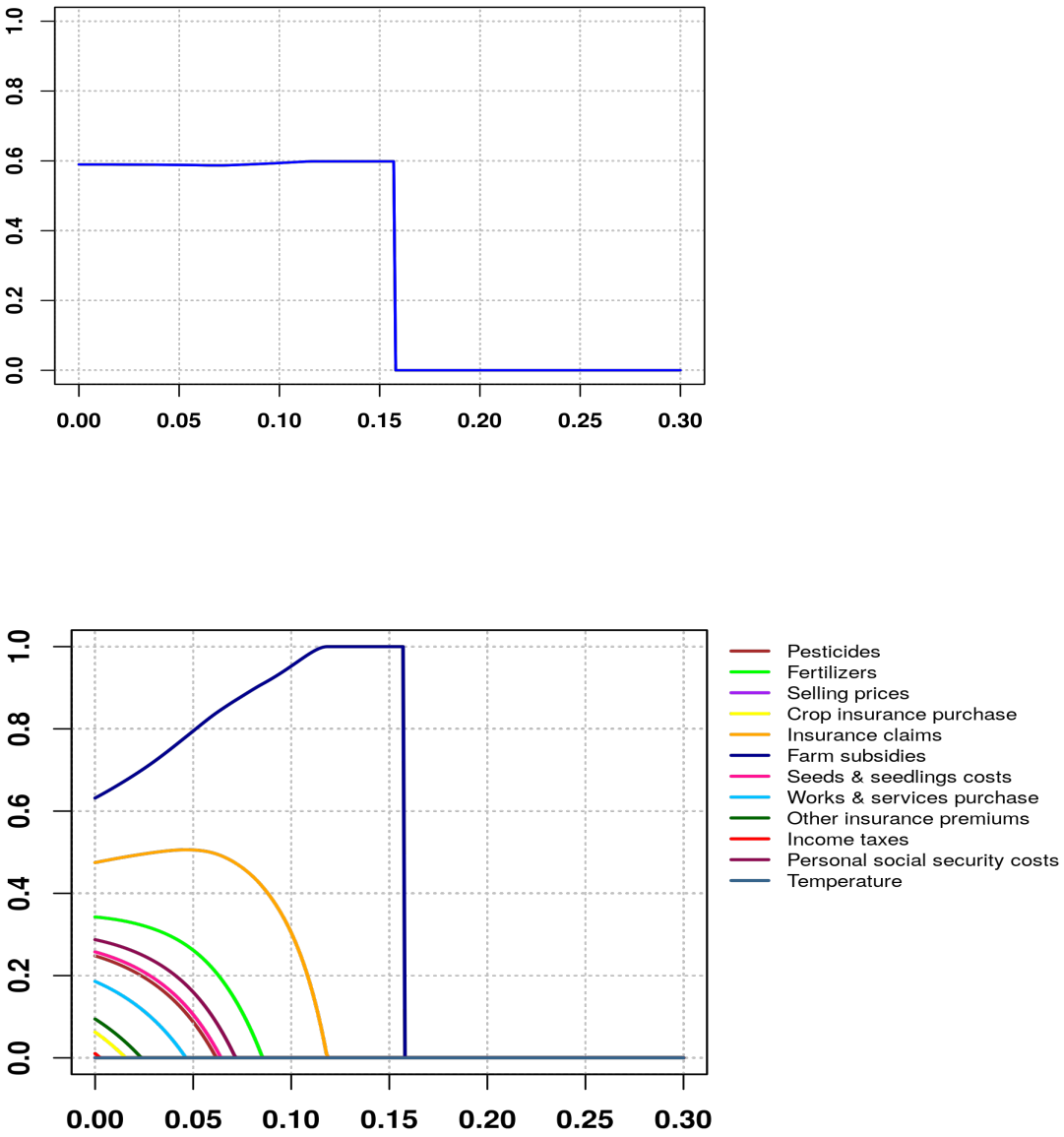


Figure 15 Farm income data with sparse prior. Top: Graph of the estimated conditional correlation $\hat{\rho}(X^t\hat{\beta}^S_{MAP}(y), Y|Y \geq y)$ with $y = Y_{n-97+1,n}$ as a function of the concentration λ. Bottom: Graph of the coordinates $|\hat{\beta}^S_{MAP,j}(y)|$ with $y = Y_{n-97+1,n}$ for $j = 1, \ldots, 12$, as functions of the concentration λ.

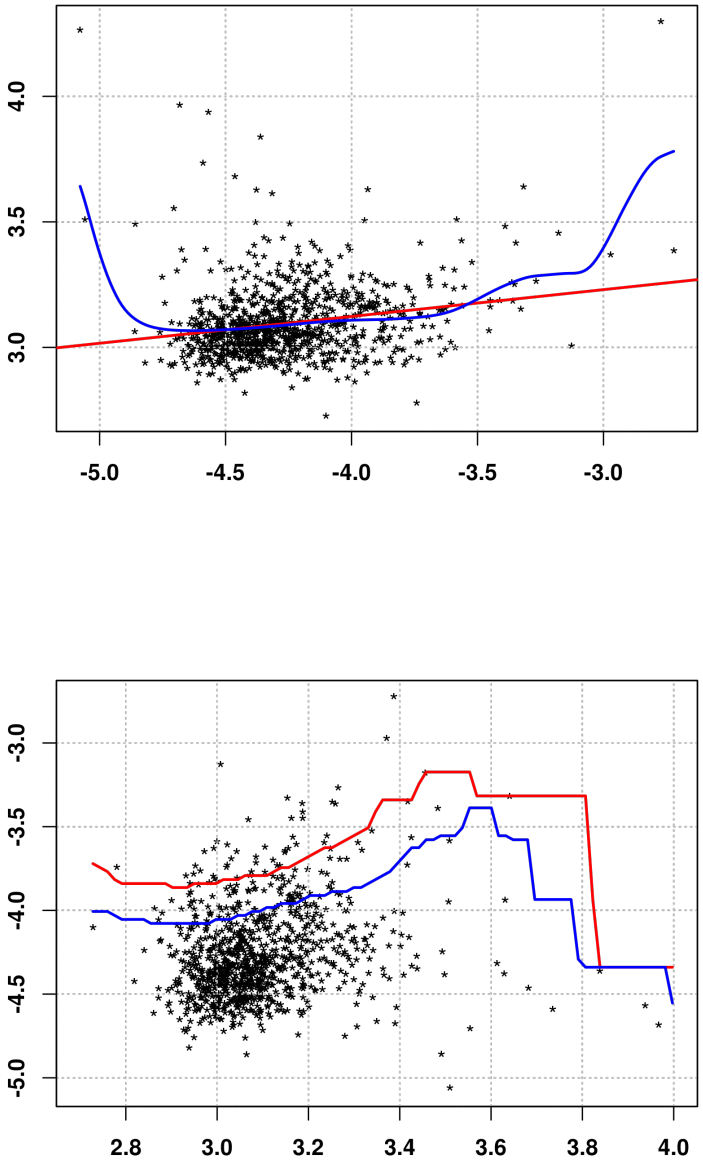


Figure 16 Farm income data. Top: scatter-plot $\left(Y_i, \hat{\beta}^C_{MAP}(y)^t X_i\right)$ in log scale obtained for $y = Y_{n-k+1,n}$ with $k = 97$ and $\kappa_1 = 2 \times 10^8$. The regression line (red) and a kernel estimate of the link function (blue) are superimposed. Bottom: scatter-plot $\left(\beta^C_{MAP}(y)^t X_i, Y_i\right)$ in log scale obtained for $y = Y_{n-k+1,n}$ with $k = 97$ and $\kappa_1 = 2 \times 10^8$. The estimated conditional quantiles $\hat{q}(\alpha|\hat{\beta}^C_{MAP}(y)^t X)$ are superimposed ($\alpha = 0.15$: blue line, $\alpha = 0.05$: red line).

Chapter 4

Yield and price dependence structures: A copula-based model of French farm income

Abstract

Revenue insurance is an agricultural risk management tool that provides farmers with joint coverage for yield and price risks. When designing this type of insurance product, it is important to model the variability of revenue risks by evaluating the interaction between crop yields and prices. This chapter presents a study on the evaluation of the dependence structure between prices and yields in the cereal and wine sectors in France, using copulas tool. This study also evaluates the conditional dependence given other factors, and provides implications for the evaluation of the possibility of establishing a revenue insurance contract. This chapter is presented as an article published in "2020-Annual Meeting of the Agricultural and Applied Economics Association" (Bousebata et al., 2020). We show that the dependence between prices and yields is relatively high and can be described by the Frank copula. We find that this dependence structure is unstable for cereals (wheat and maize) because they are standard crops whose prices follow world market trends. Wine always shows a negative correlation, as this sector is structured in terms of territory and quality and prices are controlled locally. This study also shows that French cereal and wine production is strongly influenced by extreme weather conditions, such as drought. These results are crucial for better management of price and yield risks, especially for cereal producers. They support the idea of the possibility of implementing revenue insurance in France that takes into account the correlation between prices and yields, and also the impact of other external factors such as weather indicators.

Resumé

L'assurance revenu est un outil de gestion des risques agricoles qui offre aux agriculteurs une couverture conjointe des risques de rendement et de prix. Lors de la conception de ce type de produit d'assurance, il est important de modéliser la variabilité des risques de revenu en évaluant l'interaction entre les rendements et les prix des cultures. Ce chapitre présente une étude sur l'évaluation de la structure de dépendance entre les prix et les rendements dans les secteurs céréaliers et viticoles en France, en utilisant l'outil des copules. Cette étude évalue également la dépendance conditionnelle en fonction d'autres facteurs, et fournit des implications pour l'évaluation de la possibilité de mettre en place un contrat d'assurance revenu. Ce chapitre est présenté comme un article publié dans "2020-Annual Meeting of the Agricultural and Applied Economics Association" (Bousebata et al., 2020). Nous y montrons que la dépendance entre les prix et les rendements est relativement élevée et peut être décrite par la copule de Frank. Nous constatons que cette structure de dépendance est instable pour les céréales (blé et maïs) car ce sont des cultures standard dont les prix suivent les tendances du marché mondial. Le vin montre toujours une corrélation négative, comme ce secteur est structuré en termes de territoire et de qualité et les prix sont contrôlés localement. Cette étude montre également que la production céréalière et viticole française est fortement influencée par des conditions climatiques extrêmes, telles que la sécheresse. Ces résultats sont cruciaux pour une meilleure gestion des risques de prix et de rendement, notamment pour les producteurs de céréales. Ils soutiennent l'idée de la possibilité de mettre en place une assurance revenu en France qui tienne compte de la corrélation entre les prix et les rendements, mais aussi de l'impact d'autres facteurs externes tels que les indicateurs météorologiques.

Yield and price dependence structures: A copula-based model of French farm income

Abstract

This paper aims to assess and model the dependence structure between crop yields and prices, by using a copula approach. The study is conducted on a database of French farms by considering cereal and wine-growing productions for years 2014 to 2016. We find that the dependence between prices and yields can be modelled with the Frank copula. This dependence is relatively high and influenced by high temperatures. The results highlight some implications for the development of revenue insurance policies aimed at improving the hedging of cereal production.

Keywords Farm income · Dependence structure · Copulas · Crop insurance · France

1 Introduction

Agriculture is a sector where income is subject to a wide variety of risks arising from large-scale natural events (Goodwin and Hungerford 2014; Wang et al. 2020) and the variability of agricultural commodity prices (Johnson 1975). Thus, farm income confronts two main types of risks related to yield and price volatility. The risk of poor yield is mainly due to weather events such as drought, frost, insect infestation, diseases and agricultural techniques implemented by farmers (Coble and Knight 2002) and it is increased by climate change (Kapphan et al. 2012). Price risk is rather linked to the deregulation of financial markets (Chavas 2011) and explained by the fact that most European countries have shifted from market-based support to decoupled direct payments. Then, producers are exposed to high price volatilities on world commodity markets (El Benni et al. 2016).

Conforming to this framework, the European Union has defined, particularly during the reform of the Common Agricultural Policy (CAP) in 2013, some risk management tools including subsidised crop insurance, mutual funds and income stabilisation tools (IST) (Hine et al. 2016; Meuwissen et al. 2018; El Benni et al. 2016). However, agricultural insurance plays a limited role with regard to the hedging of price risk and most IST tools have not yet been implemented (DG-AGRI 2017). Therefore, it appears necessary to carry

out evaluations of the European case because the underlying design of the IST proposed in Europe is different from that of the USA. For instance, the Farm Bill comprises several insurance systems which hedge yield and revenue losses. Premiums are affordable for farmers because they are highly subsidised (Smith et al. 2014).

In France, no such tool is available to farmers (El Benni et al. 2016). France has set up only a private crop insurance system, where premiums can be subsidised up to 65%, and a national public fund for the mutualisation of health and environmental risks (FMSE, *Fonds de Mutualisation Sanitaire et Environnementale*). Moreover, these two risk management tools have a limited effect, as French producers continue to receive European payments from the CAP of around 7 billion € per year, disconnected from market and weather trends (Lidsky et al. 2017). Insurance coverage remains quite low despite the important subsidies dedicated to multi-peril crop insurance.

Ongoing reforms, currently being carried out by the Common Organisation of Agriculture and Agro-industry Markets, are encouraging professional bodies and governments to develop new agricultural insurance products taking into account both yield and price risks, which are key determinants of revenue. They aim at increasing the attractiveness of revenue insurance compared to other risk management tools implemented for different reasons. First, public subsidies for revenue insurance seem justified because the risk covered is probably systemic, i.e. many farmers are exposed to the risk at the same time, thus allowing for public transfers (Meuwissen et al. 2003). Second, correlations between prices and yields are implicitly considered by a farm revenue insurance (El Benni et al. 2016), which seems advantageous compared to separate yield or price risk management instruments.

Some studies have considered actuarial assessments of potential revenue insurance, the resulting costs for the government, potential beneficiaries and conceptual studies on adverse selection and moral hazard issues. However, these studies do not focus on the application of revenue insurance and the modelling of its underlying risks (yields and prices) (Meuwissen et al. 2003; Mary et al. 2013). It is therefore relevant to model the dependence of yields and prices.

Modelling this dependence is of great concern as it may have implications for the eventual implementation of revenue insurance that would address the risks of farm production (Ahmed and Serra 2015). Indeed, it should provide an understanding of the distributions of yield and price risks, which interact simultaneously. Ignoring the dependence between these two risk factors could lead to an overestimation of risk for the insurer. For example, in the case of a "natural" hedge, revenue is stabilised due to the negative relationship between crop yields and prices. Conversely, the case of the positive relationship between yields and prices in a low-price market environment may result an under-hedging of revenue for the producer.

In the statistics literature, there exist several models for dependence structure between price and yield risks. The problem of such approaches is that the individual behaviour of each variable has to be represented by the same parametric family of univariate distributions (Genest and Favre 2007). Thus, it becomes necessary to construct new multivariate distributions with fixed margins and fixed dependencies properties (Kazi-Tani and Rullière 2019). In order to develop a multivariate model with given marginals, a copula approach can be used to characterise the joint distribution of different risks, thus offering considerable flexibility in empirical research. Copulas have recently become a part of the toolkit for applied economic research, resulting in an increasing need for the modelling of multivariate risk factors and their interaction (Woodard et al. 2011). For instance, Emmanouilides and Fousekis 2014 studied the structure of price dependence along the beef supply chain in the USA while Fousekis and Grigoriadis 2017 analysed the strength and the pattern of price relationships for the different types of coffee. With regard to farm revenue insurance, joint modelling of price and yield risks using copulas has been the subject of few studies, and those that exist are mainly in the USA. Zhu et al. 2008 used copulas to model the interaction between prices and yields in order to design an efficient whole farm insurance contract.

The novelty of this research is to investigate and model the pattern of price and yield dependence on a real data set of French farm income extracted from the Farm Accountancy Data Network (FADN). Two types of crops are considered: cereals (wheat and maize) and wine growing. The objective of this research is pursued using the statistical tool of copulas. Various copula models are tested for their ability to model yield and price risks. We also model the dependence structure conditionally on other covariates such as crop insurance purchase, insurance claims, temperatures and sunshine, in order to measure the influence of these factors using conditional copulas. Then, an insight related to the potential to establish a farm revenue insurance that would address the risks of cereal and wine productions, is proposed.

The rest of the article is organised as follows: in Section 2, we develop the methodological tools to perform the copula analysis. Once these tools are available, we present the data in Section 3 and the empirical results in Section 4. Finally, Section 5 concludes this study.

2 Empirical framework

2.1 Copulas

The concept of copulas was introduced in 1959 by Abe Sklar. During the financial crisis of 2007 and 2008, copulas have come to the attention of the general public due to their use in the modelling of multidimensional phenomena, mainly in the realm of quantitative risk management. They are a flexible tool that can be used to realistically represent risk dependence.

By definition, a copula is a multivariate distribution function with standard uniform univariate margins. Thus it contains all the information on the dependence structure of the model. For the sake of simplicity, let us focus on the bivariate case, in which we consider a pair of continuous random variables X and Y marginally distributed according to $F(x) = \mathbb{P}(X \leq x)$ and $G(y) = \mathbb{P}(Y \leq y)$. Let $H(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ be their joint distribution function. According to Sklar's theorem (Sklar 1959), there exists a unique function $C : [0, 1]^2 \to [0, 1]$ such that:

$$H(x, y) = C(F(x), G(y)), \qquad \text{for all } (x, y) \in \mathbb{R}^2. \tag{1}$$

The function C is referred to as the copula associated with H. It is the bivariate cumulative distribution function (cdf) of the random vector $(F(X), G(Y))$ with uniform margins on $[0, 1]$. The mappings $X \mapsto U := F(X)$ or $Y \mapsto V := G(Y)$ used in the above representation are usually referred to as the probability-integral transformations (to uniformity) and are standard tools for simulation purposes.

2.2 Dependence measures

Several measures of association between the components of a random pair can be considered, Kendall's Tau (Nelsen 2007) [paragraph 5.1.1], and Spearman's Rho (Nelsen 2007) [paragraph 5.1.2] being the most popular ones. These measures are invariant to strictly increasing functions and can be interpreted as probabilities of concordance minus probabilities of discordance of two random pairs. Both of them can be written only in terms of the copula C:

$$\tau = 4 \int_0^1 \int_0^1 C(u, v)\, dC(u, v) - 1, \tag{2}$$

$$\rho = 12 \int_0^1 \int_0^1 C(u, v)\, du dv - 3. \tag{3}$$

Let us note that ρ coincides with the correlation coefficient between the uniform marginal distributions. Starting from a sample $(U_1, V_1), \ldots, (U_n, V_n)$ of independent observations from C, it can be estimated by its empirical counterpart as

$$\hat{\rho} = \frac{12}{n} \sum_{i=1}^{n} U_i V_i - 3. \tag{4}$$

A similar formula holds for $\hat{\tau}$. Another measure of association based on concordance called *medial correlation coefficient*, was proposed by Blomqvist (Nelsen 2007) [paragraph 5.1.4], and is given by :

$$\beta = 4C\left(\frac{1}{2}, \frac{1}{2}\right) - 1. \tag{5}$$

This parameter quantifies the probability that X and Y would jointly exceed their median value.

2.3 Inference

Three main approaches have been proposed for estimating copulas: parametric, semi-parametric and non-parametric methods (Genest and Favre 2007). In our context, we advocate for the parametric approach which is based on the estimation of the parameters(s) θ of the copula assumed to belong to some parametric family $\{C_\theta,\ \theta \in \Theta\}$.

Numerous parametric families of copulas can be found in the literature. Let us focus on two popular models: Elliptical and Archimedean copulas. Elliptical copulas (Frahm et al. 2003) are built from elliptical distributions thanks to an uniformization of their margins. The level sets of an elliptical distribution density are ellipses whose shape is determined by a (kind of) covariance matrix. Important examples in this family are the Gaussian and the Student copulas. Archimedean copulas (Naifar 2011) are determined by a univariate function, called the generator, whatever the dimension is. A number of generators have been proposed, involving on one or two parameters and tuning the dependence strength between the marginals. In this study, we focus on three Archimedean copulas: Gumbel, Frank, and Clayton (Beck 2015) [pages 17-21]. The estimation of the parameter(s) θ can be done for instance using the maximum likelihood method or the method of moments (Mazo et al. 2014). In the latter case, θ is estimated by minimizing a given distance between the empirical $\hat{\tau}$ and $\hat{\rho}$ computed from Equation (4) and the theoretical ones $\tau(\theta)$ and $\rho(\theta)$ calculated according to Equations (2) and (3) under the model C_θ.

2.4 Goodness-of-fit tests

Two main techniques can be used to select the copula that fits best a dataset. First, one can rely on visual diagnostics. The idea is to use Rosenblatt's transformation (Hofert and Mächler 2014) to transform, under the null hypothesis $(U, V) \sim C_0$, the pairs (U_i, V_i) towards independent observations. Then, the p-value of an independence test is computed and encoded as a color, see Figure 2 for examples, in the empirical results section. Second, one may use a goodness-of-fit test based on Kendall's distribution function K (also called multivariate probability integral transformation) (Genest et al. 2009) defined as

$$K(t) = \mathbb{P}(C(U, V) \leq t) = \int_0^1 \int_0^1 \mathbb{1}_{\{C(u,v)\leq t\}} dC(u, v). \tag{6}$$

The theoretical Kendall distribution under the null hypothesis $(U, V) \sim C_0$ is compared to its sample version thanks to a Cramér–von Mises statistics (Genest and Favre 2007) and the associated p-value is computed.

2.5 Covariates

In some cases, the dependence structure of the random pair (X, Y) may depend on an external (possible multivariate) random variable Z. Conditional copulas were introduced to tackle this issue (Gijbels et al. 2011).

Similarly to Equation (1), one can write the joint and marginal distribution functions of (X, Y) conditionally on $Z = z$, as:

$$H_z(x, y) = \mathbb{P}(X \leq x, Y \leq y | Z = z) = C_z(F_z(x), G_z(y)) \tag{7}$$

where $F_z(x) = \mathbb{P}(X \leq x | Z = z)$ and $G_z(y) = P(Y \leq y | Z = z)$. In this context, C_z is referred to as a conditional copula. Starting from a set of observations $(X_1, Y_1, Z_1), ..., (X_n, Y_n, Z_n)$, a non-parametric estimator of $C_z(u, v)$ can be considered (Gijbels et al. 2011):

$$\hat{C}_z(u, v) = \sum_{i=1}^{n} w_i(z, h) \mathbb{1}_{\{\hat{F}_z(X_i) \leq u, \hat{G}_z(Y_i) \leq v\}}. \tag{8}$$

Here, $w_i(z, h)$ is a sequence of weights selecting the observations (X_i, Y_i) such that the associate covariate Z_i is close to the estimation point z. The range of the selected points is tuned by the parameter h_n called the bandwidth. The margin distributions F_z and G_z are estimated using similar smoothing techniques:

$$\hat{F}_z(x) = \sum_{i=1}^{n} w_i(z, h) \mathbb{1}_{\{X_i \leq x\}}, \ \hat{G}_z(y) = \sum_{i=1}^{n} w_i(z, h) \mathbb{1}_{\{Y_i \leq y\}}. \tag{9}$$

The conditional copula can be used to estimate conditional Spearman's ρ and Kendall's τ providing then association measures depending on the covariate Z. As an example, Spearman's ρ (3) is extended to the covariate framework as

$$\rho(z) = 12 \int_0^1 \int_0^1 C_z(u, v) du dv - 3, \tag{10}$$

and the associated estimator $\hat{\rho}(z)$ is obtained by plugging the estimated conditional copula (8) in the previous Equation (10).

3 Data

3.1 Database

In order to model farm revenue for different types of crops, this work is based on an empirical data extracted from the Farm Accountancy Data Network[1] (FADN). This exhaustive dataset surveys around 7 000 commercial-sized farm holdings every year. It comprises significant accounting and financial information about French professional farms along with individual and structural data. Particular attention is paid to the years 2014 and 2015 when French cereal production reached a high record in a market context with low prices (Rodier et al. 2015). We also pay attention to 2016, which was a year characterised by a decline in harvests, due to spring storms and summer drought (Triquenot et al. 2016).

1. A detailed presentation of the database can be found at: http://agreste.agriculture.gouv.fr

Modelling the pair (price, yield) for a specific type of crops is a relevant way to learn more about farm revenue generation. We also model this pair conditionally to other covariates. The list of variables involved in our study is given in Table 1.

Table 1. Database description

Variable	Definition	Unit
Farm specialisation	3 types (wheat, maize, quality wine-growing)	Class
Gross product	Gross product of the considered crop	Euro
Gross yield	Gross yield of the considered crop	Quintal or Hectolitre
Harvested acreage	Cultivated area of the considered crop	Hectare
Yields	Yields divided by the acreage	Quintal or Hectolitre/Hectare
Price	Gross product divided by quintals or hectolitres sold	Euro/Quintal or Hectolitre
Crop insurance	The Farm purchased or not a crop insurance policy	Yes/No
Insurance claims	The Farm received or not some crop insurance claims	Yes/No
Temperature	Deviation from the average of the last five years	°C
Precipitation	Deviation from the average of the last five years	mm
Sunshine duration	Deviation from the average of the last five years	Hour

Notes: Temperature, precipitation and sunshine are continuous variables. Crop insurance and claims are discrete variables.

In order to overcome any operational error during data collection, it was necessary to preprocess raw data and conduct control and consistency tests to deal with outliers or missing data that could bias our study and then ensure the robustness of our results.

3.2 Choice of considered sectors

We selected three different types of crops : wheat, maize and quality wine-growing. Wheat is the prominent cereal produced in France. It is mostly located in the West of France and around the Parisian basin[2]. France is the first European producer and exporter of wheat and it is ranked fifth largest country in the world in terms of national wheat production (Ben-Ari et al. 2018). This is due to very high yields, about 7.4 t/ha, compared to the world's four largest wheat producers, such as Russia and the United States, which harvest about 5 and 3 t/ha of wheat respectively. Maize is the second largest crop production in France, cultivated on more than 3 million hectares in 2016. Thanks to favourable soil and climate conditions and the performance of producers. France is also the world's largest exporter of maize seeds (Ben-Ari et al. 2018).

Weather conditions in autumn 2014 and summer 2015 had very contrasting effects on cereals (Delort et al. 2014; Rodier et al. 2015). Winter crops such as wheat had high yields, unlike autumn crops such as maize which suffered from drought and summer heat waves. However, the French wheat record harvest occurred within an abundant global context. Thus, wheat price dropped at the same time on global markets. However, the drop of the euro against the dollar supported the prices of agricultural commodities exchanged

2. More details can be found at: https://agriculture.gouv.fr/overview-french-agricultural-diversity

in euros. For maize, despite the decrease in production, global stocks remained high (Rodier et al. 2015). In 2016, cereal production suffered greatly in France due to climatic conditions (bad weather in spring and drought in summer) which led to significant yield decline. Despite the poor harvests in France, cereal prices remained low, due to the abundance of world production (Triquenot et al. 2016).

Wine-growing comes in second place after cereals in terms of yields. The production value amounts to over 12,4 billion euros (among which 79% for quality wines). In terms of wine production, France occupies the first place worldwide along with Italy and Spain, depending on years. French viticulture is a leading production mostly based on family farms. In spite of the slight decrease of wine consumption every year, prices increase regularly thanks to exports. The two main concepts related to French quality wines are the concept of terroir and the controlled designation of origin system (*Appellation d'Origine Contrôlée* - AOC). Appellation rules define which grape varieties and winemaking practices are approved for classification in each of France's geographically defined "appellations".

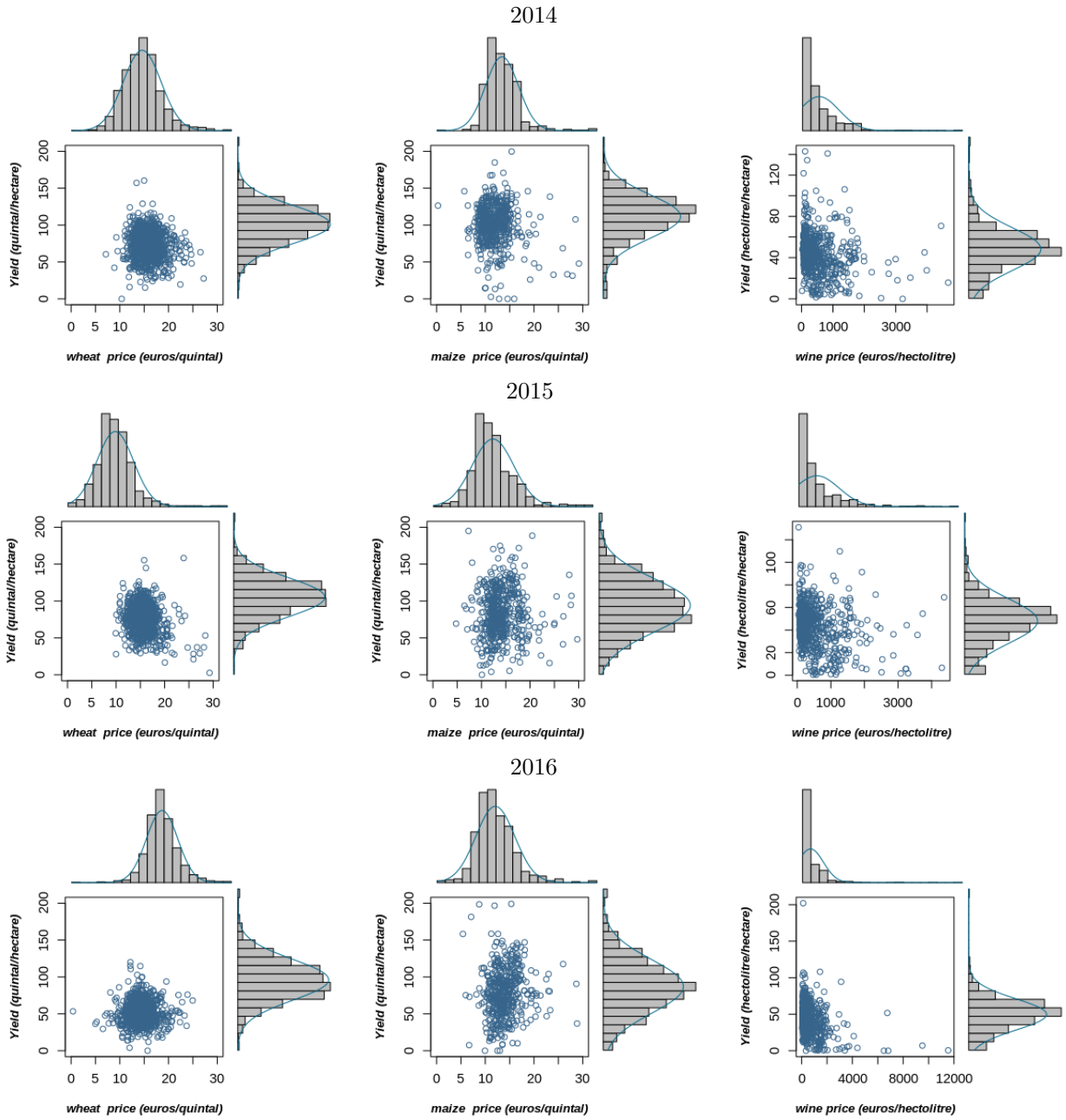
Thanks to mild temperatures in winter and spring 2014, wine production increased by 17% for AOC wine. At the same time, production stocks at the beginning of the 2014/2015 wine year were lower than in the previous year (-10%) for all wine categories. Along with a reduced dynamic of foreign trade, prices of AOC wine felt sharply at the beginning of the year before stabilising, while they increased for other wines (Rodier et al. 2015). Year 2015 was characterised by a slight increase in harvest levels but stable and limited availability, especially for AOC wines (Rodier et al. 2015). Prices increased slightly compared to 2014. In 2016, several vineyards were severely affected by several weather accidents and the impact on harvests was very significant. However, in the first nine months, prices of AOC wine were dynamic (+7.5% year-on-year), and systematically above the 2015 prices (Triquenot et al. 2016).

4 Empirical results

4.1 Joint modelling of yields and prices

Before performing copula fitting techniques, we first examine how wheat, maize and wine productions are distributed. Scatter plot and histograms of yields per hectare and prices for respectively 2014, 2015 and 2016 are given in Figure 1. It appears that prices and yields per hectare for wheat and maize crops are symmetrically distributed, while they are not for wine crops. Scatter plot for wine follows a downward trend in yields when prices rise for the three years. For cereals, a non-linear relationship is identified, hence confirming the need to use copula to model the dependency.

Fig. 1. Histograms and Scatter plot of (prices, yields) for considered productions



It is also interesting to use Kendall's τ or Spearman's ρ statistic to estimate a rank-based measure of association on the pair (price, yield). Table 2 gives a summary of Spearman's ρ coefficient, a rank-based measure of association on the pair (price, yield), and the p-value associated with the correlation tests.

For wheat, we notice a strong negative correlation between yields and prices in 2015 based on Spearman's method, of about -0.19. The test of association between paired samples gives a p-value 4×10^{-9} which means that the null hypothesis (which supposes that prices and yields are independent) is rejected at the 5% level. The negative correlation means that yields per hectare and prices vary in opposite ways. This is consistent

Table 2. Summary of independence tests of (prices, yields)

	Wheat		Maize		Wine	
Year	Spearman's ρ	P.value	Spearman's ρ	P.value	Spearman's ρ	P.value
2014	-0.0853	0.0086	0.0042	0.9220	-0.3051	0.0000
2015	-0.1893	0.0000	0.0747	0.0938	-0.2773	0.0000
2016	0.0631	0.0593	0.1473	0.0013	-0.3394	0.0000

Notes: The null hypothesis H_0 of Spearman rank correlation test assumes that prices and yields are independent and it is considered at the 5% level.

with the hypothesis of efficient markets assuming that prices and yields tend to move in opposite directions. This negative correlation is particularly obvious in 2014 and especially 2015, as abundant wheat harvests and higher yields (Delort et al. 2014; Rodier et al. 2015) led to commensurate market prices decreases. This effect is called the "natural hedge" (Sherrick 2012). The correlation dropped in 2016 because of unfavourable meteorological conditions that penalised the harvesting of wheat in France (Triquenot et al. 2016). In contrast to the French situation, the world cereal harvest reached a record level, putting pressure on prices. Despite low yields in France, prices remained as low as before, hence the positive correlation.

The same reasoning applies for maize, with a positive correlation. For two consecutive years, 2015 and 2016, the lack of rain and summer heat hampered maize development. The production performance declined below the five-year average, and total maize exports and stocks declined due to lower production. However, prices did not rise, as global stocks were high during 2016.

For wine production, Spearman's ρ shows that there exists a very strong dependence between yields and prices. In 2014 and 2015, the harvests were slightly higher with stable and limited stocks, while prices of AOC wine were lower (Rodier et al. 2015). This explains the very strong negative correlation conversely between yield and prices. For year 2016, following the succession of weather hazards, wine production was severely affected (Triquenot et al. 2016). The historical decline in French harvest took place in a context of a decline of world production. Yet, production stocks increased in 2016 for AOC wines as a result of the good harvests in 2014 and 2015. These stocks largely offset the negative impact of a reduced production in 2016.

Elasticity is another interesting economic measure to quantify the yield and price sensitivity from one year to another. It is defined as a ratio of two variations:

$$\text{elasticity} = \frac{\text{price growth rate}}{\text{yield growth rate}} \tag{11}$$

Table 3 shows that the wine price is highly elastic. In 2014/2015, prices changed relatively faster than yields. This is consistent in the case of wine where, despite low harvests, wine lovers are willing to pay high

prices for high quality wine. As a result, we observe a super negative elasticity. For 2015/2016, elasticity is rather positive, which means that yield increases caused similar trends in prices. This may seem rather paradoxical given the decline in harvests in 2016, but as already mentioned above, thanks to the good harvests in 2014-2015, the stocks could compensate this decline. Wine prices seem to be very dependent on quality rather than on quantity, and also on the annual effects of stocks.

Table 3. Elasticity measures

	Elasticity		
year	Wheat	Maize	Wine
2014/2015	-0.42	-0.95	-9.07
2015/2016	0.27	0.18	5.15

Notes: Elasticity measures how much prices change fractionally when yields change fractionally.

Hence, since the pair (prices, yields) is dependent for wheat, maize and wine, copulas are fitted in the next paragraph to model this dependence.

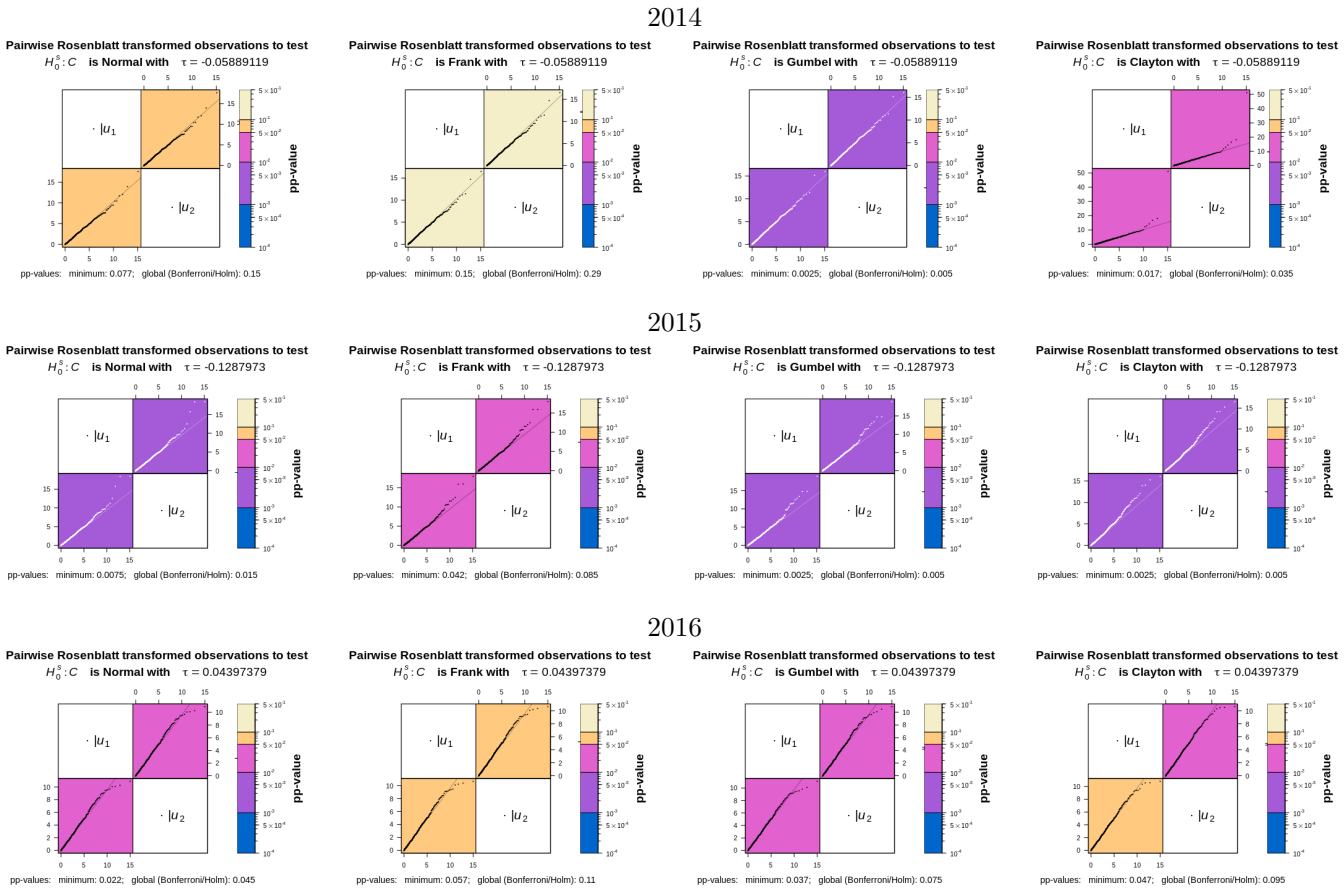
4.2 Price and yield modelling copula

Several goodness-of-fit (GOF) tests are performed with the Gumbel-Hougaard, Clayton, Frank, Normal and Student copulas as candidate families. Here both formal and informal ways to select the modelling copula will be discussed in turn.

First, the informal way relies on graphical diagnostics. We focus on the method based on Rosenblatt's transformation (Hofert and Mächler 2014). The observations are used to compute the pairwise Rosenblatt transformed data under different hypothesis, meaning different copulas. Here we test the GOF with Gumbel, Clayton, Normal and Frank copulas. Afterwards, we apply the pairwise test of independence to compute a matrix of p-values, converted to colours as shown on Figure 2. These figures display the transformation of the pairwise scatter plots to pairwise QQ-plots for the four copulas. We conclude that Frank copula fits best wheat for the three different years (Figure 2). For wine, the pairwise Rosenblatt transformed data under different hypothesis (Figure A2 in Appendix) show that Frank copula does not fit well the model. However, in terms of empirical quantiles, this copula fits almost with the theoretical ones. Figure A1 (see Appendix) shows that apparently Normal, Gumbel and Frank copulas match well the observations of maize. Hence, it is preferable to use a formal test for GOF to come up with a more precise conclusion.

Second, focusing on the formal tests of GOF, Table 4 summarises the resulting p-values using different copulas for each crop. As it can be seen from the summary, for all five tested families, Frank copula is the best one since it has the highest p-value showing that it is not rejected at the 5% significance level. This seems consistent with the graphical approach.

Fig. 2. QQ-plot Pairwise Rosenblatt transformed observations of wheat



Notes: Orange to white colours for p-values $\geq 5\%$ which is the chosen significant level, and blue to red for p-values $\leq 5\%$.

Table 4. Goodness of fit tests p-values of (prices, yields) for considered productions

	Wheat					Maize					Wine				
year	Gumbel	Clayton	Frank	Normal	Student	Gumbel	Clayton	Frank	Normal	Student	Gumbel	Clayton	Frank	Normal	Student
2014	0.00	0.00	0.20	0.01	0.01	0.09	0.07	0.35	0.06	0.05	0.00	0.00	0.01	0.00	0.00
2015	0.00	0.00	0.01	0.00	0.00	0.01	0.03	0.22	0.01	0.03	0.00	0.00	0.00	0.00	0.00
2016	0.00	0.00	0.06	0.00	0.00	0.35	0.02	0.65	0.28	0.14	0.00	0.00	0.00	0.00	0.00

Notes: p-values for the test statistic are obtained by means of a multiplier approach Genest et al. 2009 with 10 000 replications.

4.3 Influence of covariates

As mentioned above, different elements can play a role to explain how these two marginal probability functions are tied together. Here, we consider two types of covariates: discrete (crop insurance purchase and insurance claims) and continuous (temperature and sunshine). Each one will be considered in a separate way.

Crop insurance purchase and claims

Crop insurance and claims are classified into two binary classes : farmers who purchased or not a crop insurance and farmers who received or not claims. First of all, we use a multivariate analysis of variance (MANOVA)[3] to determine whether these different classes, and also their combinations, influence the relationship between prices and yields. In this analysis, the null hypothesis assumes that the class does not affect the pair. The p-value allows to choose classes having a strong effect on the joint distribution. Therefore, we consider a classification according to insurance purchase and claims for each crop and for each year of the study. According to the formal tests of GOF for each class (Table A1 in Appendix), it appears that Frank's copula fits best in most cases, as it has the highest p-value.

To determine whether the classification made according to crop insurance purchase and claims, is efficient or not, the parameter of Frank copula θ and the medial correlation coefficient β are estimated in each case and we notice whether they differ significantly or not.

Table 5. Estimation of the parameters and the dependence measures of the selected Frank Copula

		Wheat			Maize			Wine		
year	class	θ^a	ρ^b	β^c	θ	ρ	β	θ	ρ	β
2014	No class	-0.53	-0.09	-0.07	0.04	0.00	0.00	-1.93	-0.31	-0.23
	No insurance	-0.05	0.00	-0.01		NS			NS	
	With insurance	-0.57	-0.09	-0.07		NS			NS	
2015	No class	-1.18	-0.19	-0.14	0.47	0.07	0.06	-1.73	-0.28	-0.21
	No insurance	-0.93	-0.15	-0.11		NS		-2.11	-0.33	-0.25
	with insurance	-1.17	-0.19	-0.14		NS		-1.29	-0.21	-0.16
	No claims		NS		0.86	0.14	0.11		NS	
	With claims		NS		0.13	0.02	0.02		NS	
	No insurance No claims		NS			N.S		-2.26	-0.35	-0.27
	No insurance With claims		NS			N.S		-1.88	-0.28	-0.23
	With insurance No claims		NS			N.S		-1.57	-0.25	-0.19
	With insurance With claims		NS			N.S		-0.94	-0.16	-0.12
2016	No class	0.40	0.06	0.05	0.90	0.15	0.11	-2.15	-0.34	-0.26
	No insurance	0.13	0.02	0.02		NS		-2.57	-0.40	-0.30
	with insurance	0.46	0.07	0.06		NS		-1.68	-0.27	-0.20
	No claims	0.75	0.12	0.09		NS			NS	
	With claims	0.14	0.02	0.02		NS			NS	

Notes: [a] Theta is Frank copula parameter, [b] rho is Spearman rank correlation, [c] beta is the medial correlation coefficient. NS indicates not significant parameters (for the sake of clarity, only the lines with significant parameters are displayed).

Table 5 shows how parameters θ and β change according to different classes of significant variables. It appears that the insured farms are those that have a strong negative dependence β between prices and yields, for wheat in 2014/2015 and very strong for wine in particular in 2015/2016. This is due to a perfect match between supply and demand. Indeed, on a perfect market, prices are expected to move downwards when yields are better and vice versa, which means the balance between supply and demand determines the

3. Multivariate Analysis of Variance is a statistical test that stands for multivariate analysis of variance for multiple dependent variables.

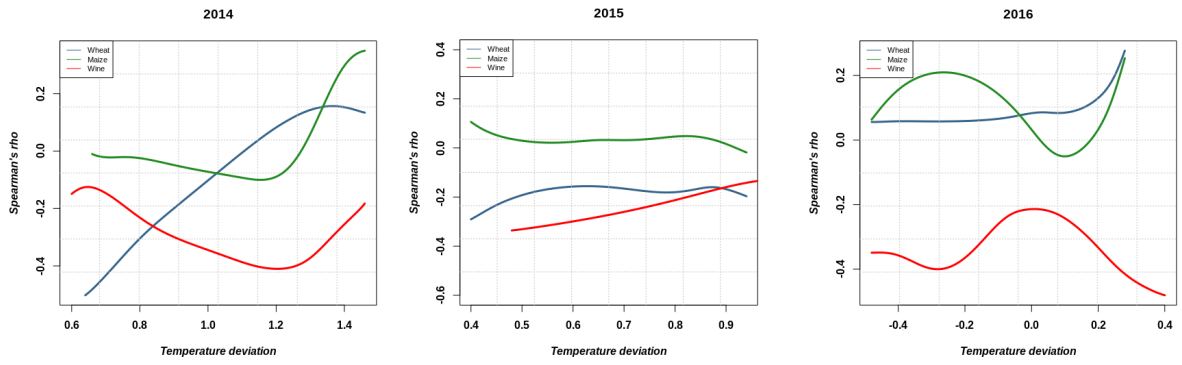
price. This was the case for wheat, as production was at a world record, resulting in lower market prices (as explained earlier). Because wine is a very particular asset, out of global markets, it behaves in a different way, linked to an elastic market impacted by quality and stocks. The correlation becomes positive in 2016 for wheat, while there is no insurance effect for maize.

As a conclusion, we note that insured wheat producers are those with the most volatile pair (price,yield), and conversely for wine. This can result in a slight adverse selection effect, because the farmer observes his yields (over the last 3 years for example) and can anticipate a future potential use of crop insurance. We also note that existing crop insurance contracts (which hedge only yield risks) are more adapted to wine than to field crops, as wine prices are determined on productive regions and follow closely yield trends. Conversely, setting up a revenue insurance appears more conceptually adapted to the hedging of cereal production rather than wine production.

Temperature and sunshine deviations

Figure 3 and 4 display the estimated parameter of the Frank copula according to temperature and sunshine deviations for wheat, maize and wine. We extract a relationship between these covariates and the structure of yield and price dependence using Spearman's ρ. To do so, we use the approach based on the conditional copula estimation described previously.

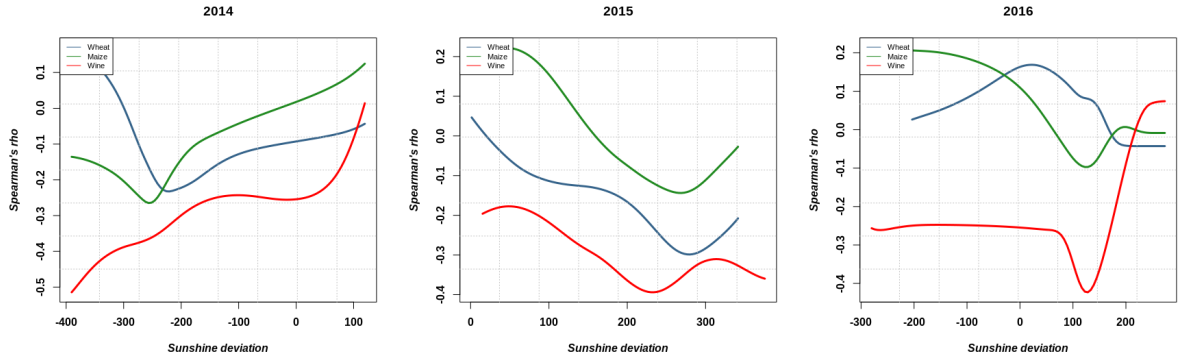
Fig. 3. Spearman's ρ conditionally on temperature deviation



(a) **(b)** **(c)**

For wheat, Figures 3 (a), (b), (c) display respectively an increasing function of Spearman's ρ depending on the temperature deviation of 2014 compared to the average of the last five years, a slight upward trend in 2015 but still a negative correlation, and stabilisation around a very low dependency in 2016. This means that when the gap increases, wheat yields and prices tend to vary in the same direction. A visual

Fig. 4. Spearman's ρ conditionally on sunshine deviation



(a) **(b)** **(c)**

inspection of Figure 4 reveals a downward trend for large number of days of sunshine as in 2015, and an upward trend for little sunshine in 2016 (the difference between this year's average and the average of the last 5 years is lower than 0). The reason behind this trend is that wheat crops are very sensitive to a change in weather conditions. In fact, price volatility and the decrease in agricultural yields are closely linked to natural hazards, such as record temperatures (2016 was an extremely hot year). Existing studies indicate that yields decline with higher temperatures and decreased precipitation (as in 2016) and increase with higher precipitation (as in 2015) (Pirttioja et al. 2015). The drought in 2016 had a strong impact on wheat (Ciais et al. 2005), as a lack of water occurred during the in early autumn, which is a critical period.

For wine, Figures 3 (a), (b) and (c) exhibit different trends : a decreasing one for 2014, an increasing one in 2015 and in 2016 with some fluctuations. They show clearly that the correlation seems to be very strong and negative when the temperature deviation increases until 1.2 C°. Spearman's ρ for sunshine deviation in Figure 4 tends to decrease for long periods of sunshine. Vines requires good sunshine, average high temperatures and regular rainfall for their growth. Solar radiation is an important element of photosynthesis that allows the vine to accumulate reserves (sugars) in its fruits. However, the vine is sensitive to very high temperatures accompanied by long periods of drought because it causes a slowdown or even a halt in the growth of leaves and grapes [4]. For this reason, extreme temperatures reduce the chance to produce a quality wine (White et al. 2006).

4. More details on this topic can be found at: https://www.oenologie.fr/climat-pour-le-vin

5 Conclusion

This paper aimed at modelling the dependence structure between yields and prices using a copula model approach. The study used a real data set of French farms extracted from the Farm Accountancy Data Network (FADN) considering two main productions: cereals (wheat and maize) and wine growing.

The results show that the dependence between prices and yields is relatively high and it can be described with a Frank Copula, regardless the type of crop. Moreover, the two variables show different types of dependence for each crop, according to events related to the year of the study : local climatic change affecting local production and yields, or global fluctuations of commodity markets as a result of global production and other external factors. We showed that the dependence structure between prices and yields is unstable for wheat and maize, while wine-growing has always a negative correlation. This reflects the organisation of the wine-growing sector, which is structured in terms of territory and quality. This is not the case for wheat and maize, which are completely standard cereals whose prices follow world market trends. On perfect markets, prices and yields use to vary in opposite directions.

This study also examined the effectiveness of existing insurance contracts. The empirical analysis showed that existing crop policies are more suited to wine than to cereals because they only hedge directly yield and not price risks. Indeed, ignoring the dependence between price risk and yield risk could lead to an overestimation of the cereal revenue risk, in the case of "natural hedge" where revenue is stabilised due to the negative relationship between crop yields and prices. Conversely, in the case of a positive relationship between yields and prices (for instance, in a low-price market environment), this may lead to an under-hedging of revenue for the producer. This study also shows that French cereal and wine productions are significantly influenced by extreme weather. Crop yields were indeed sensitive to very high temperatures in 2016.

For future studies, this work offers many insights, such as an overview related to the development and pricing of farm revenue insurance that would be more suitable for the protection of cereal . The results of this analysis support the idea of combining price and yield risk hedging into a single revenue insurance policy that would provide increased insurance coverage, especially for cereal producers.

References

Ahmed, Osama, and Teresa Serra. 2015. "Economic analysis of the introduction of agricultural revenue insurance contracts in Spain using statistical copulas." *Agricultural Economics* 46 (1): 69–79. https://doi.org/10.1111/agec.12141.

Beck, Nicholas. 2015. "Multivariate risk measures and a consistent estimator for the orthant based tail value-at-risk." PhD diss., Concordia University. https://doi.org/10.1051/ps/2018015.

Ben-Ari, Tamara, Julien Boé, Philippe Ciais, Remi Lecerf, Marijn Van der Velde, and David Makowski. 2018. "Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France." *Nature communications* 9 (1): 1–10. https://doi.org/10.1038/s41467-018-04087-x.

Chavas, Jean-Paul. 2011. "Agricultural policy in an uncertain world." *European Review of Agricultural Economics* 38 (3): 383–407. https://doi.org/10.1093/erae/jbr023.

Ciais, Ph, M Reichstein, Nicolas Viovy, André Granier, Jérôme Ogée, Vincent Allard, Marc Aubinet, Nina Buchmann, Chr Bernhofer, Arnaud Carrara, et al. 2005. "Europe-wide reduction in primary productivity caused by the heat and drought in 2003." *Nature* 437 (7058): 529. https://doi.org/10.1038/nature03972.

Coble, Keith H, and Thomas O Knight. 2002. "Crop insurance as a tool for price and yield risk management." In *A comprehensive assessment of the role of risk in US agriculture,* edited by Richard E Just and Rulon D Pope, 445–468. Kluwer Academic Press, Boston. https://doi.org/10.1007/978-1-4757-3583-3_20.

Delort, Annie, Olivier Satger, Guillaume Wemelbeke, Brice Edan, Hana Bouhalli, Patrice Arnoux, Laurent Bernadette, Marie-Anne Lapuyade, and Jeanne Gabrysiak. 2014. "Bilan conjoncturel 2014: En 2014, les prix des principales productions végétales et animales sont en recul sur un an."

DG-AGRI. 2017. *Risk management schemes in* EU *agriculture; dealing with risk and volatility.*

El Benni, Nadja, Robert Finger, and Miranda PM Meuwissen. 2016. "Potential effects of the income stabilisation tool (IST) in Swiss agriculture." *European Review of Agricultural Economics* 43 (3): 475–502. https://doi.org/10.1093/erae/jbv023.

Emmanouilides, Christos J, and Panos Fousekis. 2014. "Vertical price dependence structures: copula-based evidence from the beef supply chain in the USA." *European Review of Agricultural Economics* 42 (1): 77–97. https://doi.org/10.1093/erae/jbu006.

Fousekis, Panos, and Vasilis Grigoriadis. 2017. "Joint price dynamics of quality differentiated commodities: copula evidence from coffee varieties." *European Review of Agricultural Economics* 44 (2): 337–358. https://doi.org/10.1093/erae/jbw015.

Frahm, Gabriel, Markus Junker, and Alexander Szimayer. 2003. "Elliptical copulas: applicability and limitations." *Statistics & Probability Letters* 63 (3): 275–286. https://doi.org/10.1016/S0167-7152(03)00092-0.

Genest, Christian, and Anne-Catherine Favre. 2007. "Everything you always wanted to know about copula modeling but were afraid to ask." *Journal of Hydrologic Engineering* 12 (4): 347–368. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:4(347).

Genest, Christian, Bruno Rémillard, and David Beaudoin. 2009. "Goodness-of-fit tests for copulas: A review and a power study." *Insurance: Mathematics and economics* 44 (2): 199–213. https://doi.org/10.1016/j.insmatheco.2007.10.005.

Gijbels, Irène, Noël Veraverbeke, and Marel Omelka. 2011. "Conditional copulas, association measures and their applications." *Computational Statistics & Data Analysis* 55 (5): 1919–1932. https://doi.org/10.1016/j.csda.2010.11.010.

Goodwin, Barry K, and Ashley Hungerford. 2014. "Copula-based models of systemic risk in US agriculture: implications for crop insurance and reinsurance contracts." *American Journal of Agricultural Economics* 97 (3): 879–896. https://doi.org/10.1093/ajae/aau086.

Hine, RC, Kenneth Arthur Ingersent, and AJ Rayner. 2016. *The reform of the common agricultural policy.* Springer.

Hofert, Marius, and Martin Mächler. 2014. "A graphical goodness-of-fit test for dependence models in higher dimensions." *Journal of computational and graphical statistics* 23 (3): 700–716. https://doi.org/10.1080/10618600.2013.812518.

Johnson, D Gale. 1975. "World agriculture, commodity policy, and price variability." *American Journal of Agricultural Economics* 57 (5): 823–828. https://doi.org/10.2307/1239087.

Kapphan, Ines, Pierluigi Calanca, and Annelie Holzkaemper. 2012. "Climate change, weather insurance design and hedging effectiveness." *The Geneva Papers on Risk and Insurance-Issues and Practice* 37 (2): 286–317.

Kazi-Tani, Nabil, and Didier Rullière. 2019. "On a construction of multivariate distributions given some multidimensional marginals." *Advances in Applied Probability* 51 (2): 487–513. https://doi.org/10.1017/apr.2019.14.

Lidsky, Vincent, Carole Maudet, Georges-Pierre Malpel, François Gerster, Michel Helfter, Hervé Lejeune, and François-Gilles Le Theule. 2017. "Les outils de gestion des risques en agriculture." *Inspection générale des finances.*

Mary, Sébastien, Fabien Santini, and Pierre Boulanger. 2013. *An ex-ante assessment of CAP Income Stabilisation Payments using a farm household model.* Technical report. https://doi.org/10.22004/ag.econ.158860.

Mazo, Gildas, Stephane Girard, and Florence Forbes. 2014. "Weighted least-squares inference based on dependence coefficients for multivariate copulas." In *Compstat, 21st symposium of the IASC, Geneva, Switzerland.* https://doi.org/10.1051/ps/2015014.

Meuwissen, Miranda PM, Ruud BM Huirne, and Jerry R Skees. 2003. "Income insurance in European agriculture." *EuroChoices* 2 (1): 12–17. https://doi.org/10.1111/j.1746-692X.2003.tb00037.x.

Meuwissen, Miranda PM, Yann de Mey, and Marcel van Asseldonk. 2018. "Prospects for agricultural insurance in Europe." *Agricultural Finance Review* 78 (2): 174–182. https://doi.org/10.1108/AFR-04-2018-093.

Naifar, Nader. 2011. "Modelling dependence structure with Archimedean copulas and applications to the iTraxx CDS index." *Journal of Computational and Applied Mathematics* 235 (8): 2459–2466. https://doi.org/10.1016/j.cam.2010.10.047.

Nelsen, Roger B. 2007. *An introduction to copulas.* Springer Science & Business Media.

Pirttioja, Nina, Timothy R Carter, Stefan Fronzek, Marco Bindi, Holger Hoffmann, Taru Palosuo, Margarita Ruiz-Ramos, Fulu Tao, Mirek Trnka, Marco Acutis, et al. 2015. "Temperature and precipitation effects on wheat yield across a European transect: a crop model ensemble analysis using impact response surfaces." *Climate Research* 65:87–105. https://doi.org/10.3354/cr01322.

Rodier, Daniel, Olivier Satger, Gérard Thomas, Patrice Arnoux, Laurent Bernadette, Annie Delort, Hana Bouhalli, Marie-Anne Lapuyade, Aurélien Lavergne, and Christian Pendaries. 2015. "Bilan conjoncturel 2015: En 2015, des marchés agricoles sous la pression des excédents mondiaux et communautaires."

Sherrick, Bruce J. 2012. "Relative Importance of Price vs. Yield variability in Crop Revenue Risk." *farmdoc daily* 2.

Sklar, Abe. 1959. "Fonctions de répartition à n dimensions et leurs marges." *Publications de l'Institut de Statistique de l'Université de Paris* 8:229–231.

Smith, Vincent H, James B Johnson, and John P Hewlett. 2014. "New Farm Programs in the 2014 Farm Bill: Price Loss Coverage, Agricultural Risk Coverage and the Supplemental Coverage Agricultural Insurance Option for Wyoming Farms and Ranches."

Triquenot, Alice, Michelle Le Turdu, Thibaut Champagnol, Sylvie Bernadet, Laurent Bernadette, Annie Delort, Mélanie Kuhn-Le Braz, Marie-Anne Lapuyade, Aurélien Lavergne, Christian Pendaries, et al. 2016. "Bilan conjoncturel 2016: 2016, une année marquée par la baisse des récoltes, sous l'effet des intempéries printanières et de la sécheresse estivale, et un début d'amélioration de la conjoncture pour certains secteurs de l'élevage (porcs et lait)."

Wang, H Holly, Jesse B Tack, and Keith H Coble. 2020. *Frontier studies in agricultural insurance.* https://doi.org/10.1057/s41288-019-00156-4.

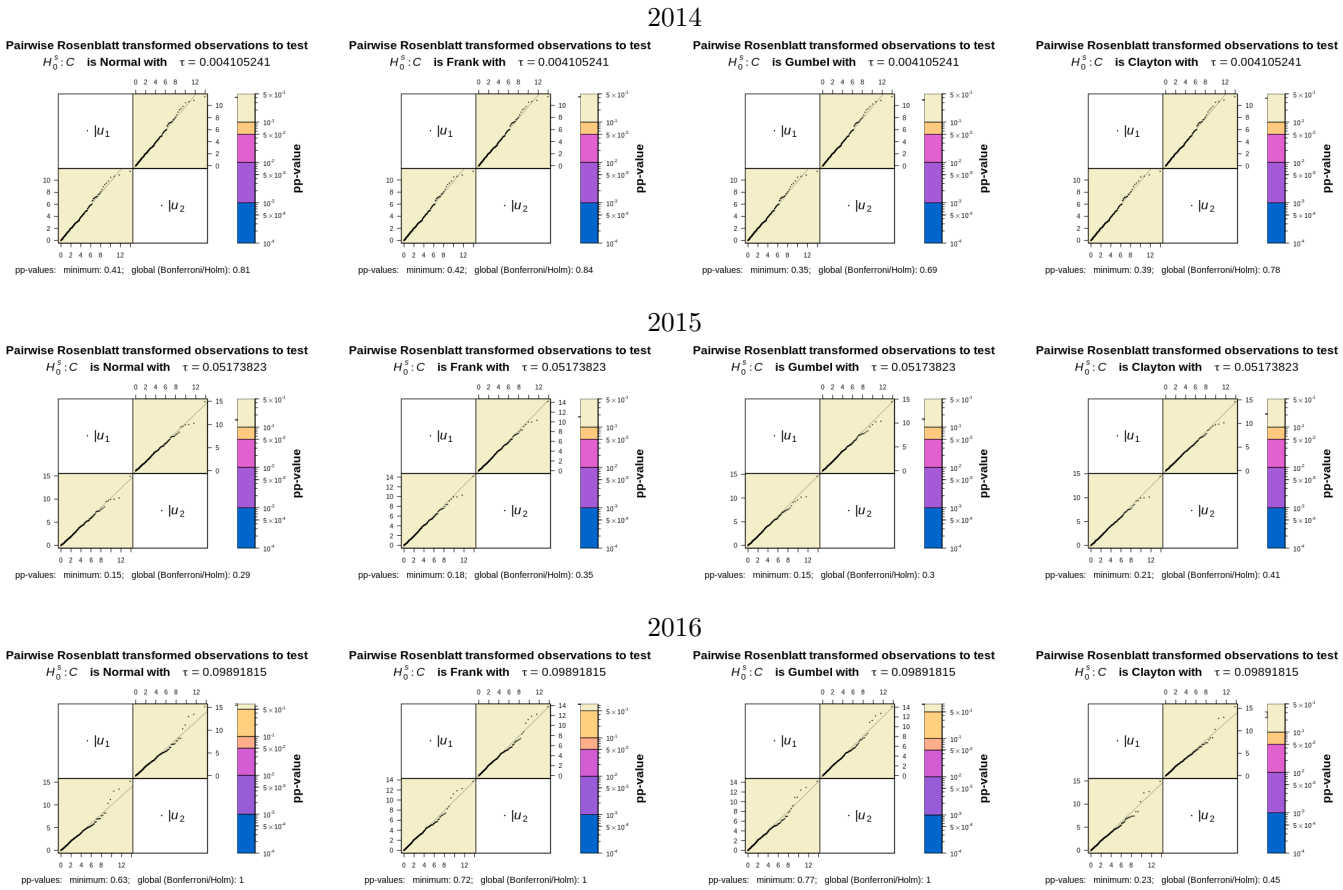
White, Michael A, NS Diffenbaugh, Gregory V Jones, JS Pal, and F Giorgi. 2006. "Extreme heat reduces and shifts United States premium wine production in the 21st century." *Proceedings of the National Academy of Sciences* 103 (30): 11217–11222. https://doi.org/10.1073/pnas.0603230103.

Woodard, Joshua D, Nicholas D Paulson, Dmitry Vedenov, and Gabriel J Power. 2011. "Impact of copula choice on the modeling of crop yield basis risk." *Agricultural Economics* 42:101–112. https://doi.org/10.1111/j.1574-0862.2011.00555.x.

Zhu, Y, S Ghosh, and B Goodwin. 2008. "Modeling dependence in the design of whole farm insurance contract a copula-based approach, Selected paper at the Annual Meeting of the Agricultural & Applied Economics Association (AAEA) 2008." https://doi.org/10.22004/ag.econ.6282.

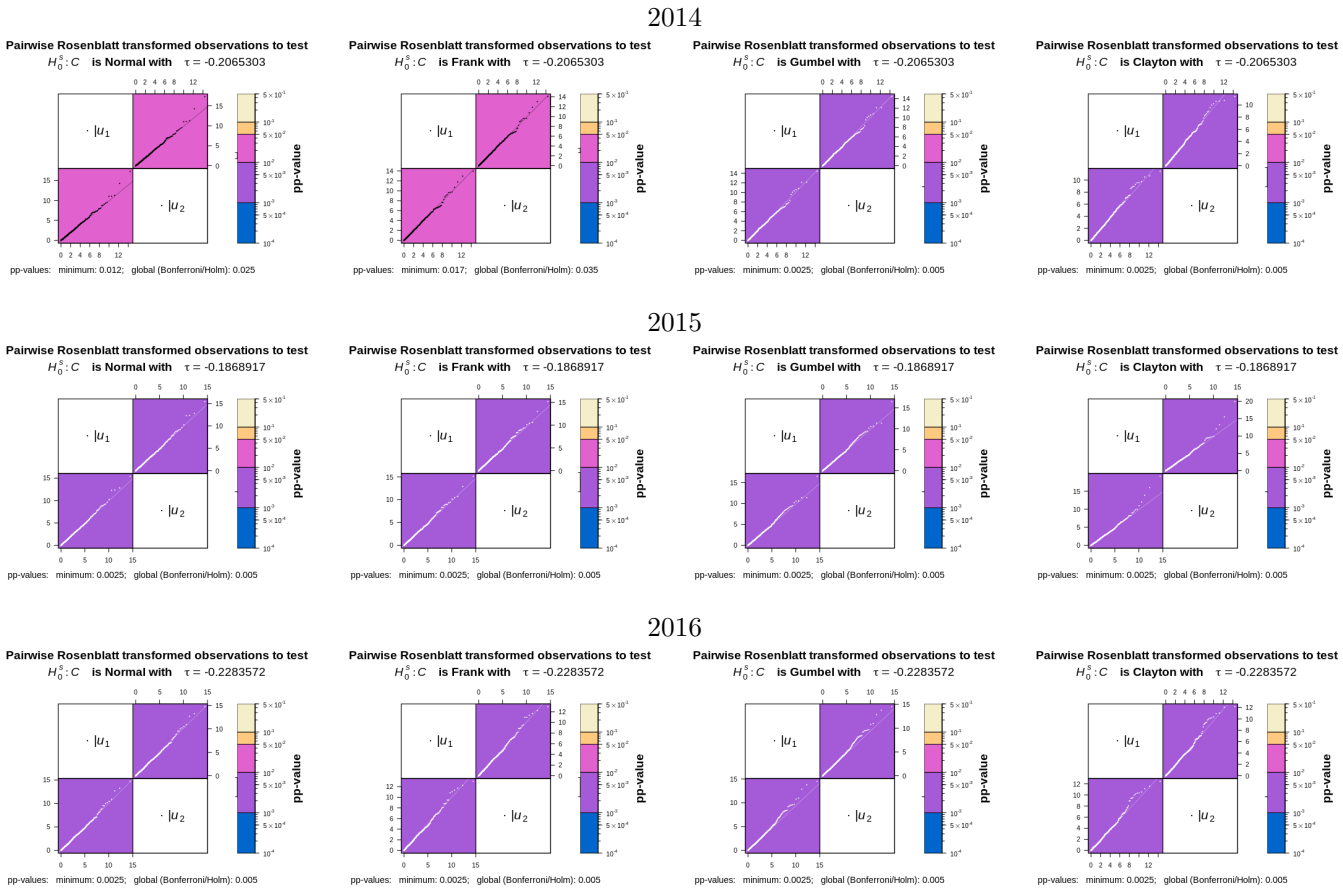
Appendix

Fig. A1. QQ-plot Pairwise Rosenblatt transformed observations of maize



Notes: Orange to white colours for p-values $\geq 5\%$ which is the chosen significant level, and blue to red for p-values $\leq 5\%$.

Fig. A2. QQ-plot Pairwise Rosenblatt transformed observations of wine



Notes: Orange to white colours for p-values $\geq$ 5% which is the chosen significant level, and blue to red for p-values $\leq$ 5%.

Table A1. Goodness of fit tests p-values for each classification according to crop insurance purchase and claim classes

		Wheat					Maize					Wine				
year	class	Gumbel	Clayton	Frank	Normal	Student	Gumbel	Clayton	Frank	Normal	Student	Gumbel	Clayton	Frank	Normal	Student
2014	No class	0.0000	0.0000	0.2032	0.0103	0.0076	0.0871	0.0748	0.3452	0.0643	0.0466	0.0000	0.0001	0.0114	0.0000	0.0000
	No ins	0.0905*	0.1206*	0.2951*	0.0624*	0.0449*	0.5509	0.8241	0.7632	0.5221	0.3977	0.0000	0.0007	0.0202	0.0000	0.0000
	With ins	0.0000*	0.0037*	0.4299*	0.0872*	0.0773*	0.0423	0.0362	0.2408	0.0267	0.0388	0.0000	0.0012	0.0588	0.0009	0.0009
	No indem	0.0001	0.0026	0.2283	0.0260	0.0214	0.2583	0.1881	0.5341	0.1959	0.2070	0.0000	0.0000	0.0169	0.0000	0.0000
	With indem	0.0071	0.0314	0.5350	0.1611	0.2140	0.3880	0.3540	0.5431	0.2347	0.2143	0.0000	0.0423	0.1638	0.0234	0.0230
	No ins No indem	0.2892	0.1844	0.3996	0.1298	0.0894	0.2694	0.5580	0.5443	0.2414	0.2227	0.0000	0.0001	0.0149	0.0001	0.0000
	No ins With indem	0.0951	0.9368	0.3374	0.1451	0.1406	0.1218	0.8390	0.2104	0.1074	0.0832	0.0518	0.5658	0.3630	0.1446	0.1701
	With ins No indem	0.0061	0.0381	0.4416	0.1513	0.1484	0.1492	0.1002	0.4065	0.1104	0.1658	0.0000	0.0023	0.1479	0.0111	0.0111
	With ins With indem	0.0024	0.0490	0.6792	0.2679	0.3147	0.2800	0.2477	0.4352	0.1421	0.1648	0.0083	0.0915	0.2282	0.0403	0.0375
2015	No class	0.0000	0.0000	0.0143	0.0000	0.0000	0.0111	0.0270	0.2206	0.0135	0.0307	0.0000	0.0000	0.0006	0.0000	0.0000
	No ins	0.1144*	0.0723*	0.4403*	0.1487*	0.1822*	0.2605	0.3910	0.4853	0.1994	0.1720	0.0000*	0.0025*	0.0103*	0.0000*	0.0000*
	With ins	0.0000*	0.0000*	0.0105*	0.0000*	0.0000*	0.0145	0.0218	0.1872	0.0131	0.0372	0.0000*	0.0036*	0.0439*	0.0004*	0.0004*
	No indem	0.0000	0.0000	0.0422	0.0006	0.0008	0.0758*	0.3275*	0.4733*	0.1495*	0.2405*	0.0000	0.0002	0.0036	0.0000	0.0000
	With indem	0.0000	0.0004	0.1048	0.0014	0.0030	0.2132*	0.2216*	0.5167*	0.1803*	0.2736*	0.0000	0.0036	0.0628	0.0014	0.0017
	No ins No indem	0.1264	0.1061	0.3857	0.1252	0.1412	0.4647	0.9012	0.7132	0.4412	0.4318	0.0000*	0.0143*	0.0176*	0.0001*	0.0000*
	No ins With indem	0.2296	0.3085	0.2738	0.0547	0.0579	0.1857	0.5005	0.2401	0.0810	0.0780	0.0004*	0.1903*	0.4104*	0.1083*	0.0943*
	With ins No indem	0.0000	0.0000	0.0115	0.0001	0.0000	0.0796	0.3099	0.4614	0.1423	0.2564	0.0000*	0.0124*	0.1235*	0.0072*	0.0053*
	With ins With indem	0.0000	0.0041	0.1805	0.0093	0.0120	0.1992	0.2180	0.5045	0.1679	0.2967	0.0125*	0.0289*	0.1337*	0.0099*	0.0173*
2016	No class	0.0000	0.0014	0.0605	0.0002	0.0001	0.3481	0.0232	0.6513	0.2760	0.1395	0.0000	0.0000	0.0001	0.0000	0.0000
	No ins	0.3174*	0.5113*	0.6218*	0.2994*	0.4244*	0.4497	0.7311	0.7713	0.4591	0.5029	0.0000*	0.0000*	0.0026*	0.0000*	0.0000*
	With ins	0.0005*	0.0029*	0.0810*	0.0006*	0.0001*	0.5143	0.0180	0.6424	0.3166	0.1782	0.0000*	0.0001*	0.0291*	0.0001*	0.0003*
	No indem	0.0088*	0.0954*	0.2521*	0.0281*	0.0213*	0.5583	0.0200	0.6286	0.2660	0.2464	0.0000	0.0000	0.0017	0.0000	0.0000
	With indem	0.0049*	0.0071*	0.1261*	0.0036*	0.0020*	0.6079	0.5453	0.8567	0.6587	0.5063	0.0000	0.0000	0.0105	0.0000	0.0000
	No ins No indem	0.1825	0.5466	0.5863	0.2508	0.3112	0.3216	0.4576	0.5907	0.2845	0.3085	0.0000	0.0000	0.0073	0.0001	0.0000
	No ins With indem	0.0272	0.3485	0.5126	0.2115	0.2098	0.2002	0.7098	0.4368	0.1788	0.2068	0.0007	0.0109	0.0851	0.0051	0.0046
	With ins No indem	0.0244	0.1261	0.2877	0.0418	0.0352	0.7333	0.0211	0.6710	0.3360	0.3081	0.0000	0.0005	0.1225	0.0125	0.0145
	With ins With indem	0.0004	0.0016	0.0690	0.0006	0.0002	0.6262	0.3512	0.8250	0.6051	0.4063	0.0001	0.0023	0.0648	0.0007	0.0011

Notes: p-values for the test statistic are obtained by means of a multiplier approach Genest et al. 2009 with 10 000 replications. Classes that have an influence are marked with $*$ at the 5% level or less and the analysis studies will be performed on them.

Chapter 5

The effects of natural hedge on revenue stability and implications for pricing the revenue insurance contract using copulas

Contents

Abstract

The natural hedge mechanism occurs when crop yields and prices are negatively correlated. An effective implementation of a revenue insurance policy requires to take into account this mechanism. In this chapter, we aim to empirically characterise the degree of the natural hedge in the wheat, maize and wine sectors, using copulas. We also aim to analyse its effect on the value of actuarially fair premiums of a revenue insurance policy. We start by introducing the context of the study in Section 5.1. The methodological approach and the description of the variables used in the analyses are presented in Section 5.2. Section 5.3 is devoted to the analysis of the natural hedge effects on income stability and crop insurance participation. Finally, we present in section 5.4 the implications of the results for the implementation of a revenue insurance policy in France. We conclude with some final remarks in Section 5.5.

Resumé

Le mécanisme de couverture naturelle se produit lorsque les rendements et les prix des cultures sont négativement corrélés. Une mise en œuvre efficace d'un système d'assurance revenu nécessite de prendre en compte ce mécanisme. Dans ce chapitre, nous visons à caractériser empiriquement le degré de couverture naturelle dans les secteurs du blé, du maïs et du vin, en utilisant des copules. Nous visons également à analyser son effet sur la valeur des primes actuariellement justes d'une police d'assurance revenu. Nous commençons par présenter le contexte de l'étude dans la Partie 5.1. L'approche méthodologique et la description des variables utilisées dans les analyses sont présentées dans la Partie 5.2. La partie 5.3 est consacrée à l'analyse des effets de la couverture naturelle sur la stabilité du revenu et la participation à l'assurance récolte. Enfin, nous présentons dans la Partie 5.4 les implications des résultats pour la mise en place d'une assurance revenu en France. Nous concluons par quelques remarques finales dans la Partie 5.5.

5.1 Introduction

Agriculture is generally subject to multiple and increasing risks. The two main risks are the risk of low yields, linked to climate change and natural perils (Anandhi et al., 2016; Moschini and Hennessy, 2001; Ullah et al., 2016), and the risk of low prices, related to the financial markets deregulation (Ortmann et al., 1992; Ullah et al., 2016). Protection against natural, climatic and market risks falls within the realm of good risk management to better cover farmers (Hardaker et al., 2015). A non-exhaustive list of risk management tools includes diversification, irrigation, agricultural inputs, futures hedging, forward contracting or insurance contracts (see Section 1.1.2 of the State of the Art). The latter is the subject of our study. There are several types of crop insurance schemes, such as yield and revenue insurance (Kang, 2007). Revenue insurance policy is more effective as it covers the overall farm income, whereas crop insurance policy only covers crop yields. It provides better protection against both yield and price risks than individual crop insurance contracts. In addition, revenue insurance also takes into account the "natural hedge" effect, when prices and yields are inversely correlated, which moderates revenue variability and has some implication on the premium calculation (Feng et al., 2014).

Natural hedge has received much attention in the analysis of agricultural risk management (Kimura et al., 2010; Finger, 2012; Ramsey et al., 2019). It is defined as a negative correlation between prices and yields. This inverse dependence could smooth out the revenue variations and thus reduce farmers demand for insurance solutions (Kimura et al., 2010). Furthermore, the degree of this dependence is of great practical importance for the revenue insurance design. Indeed, pricing these contracts requires calculating the probability of loss, which depends on the joint distribution function of yields and prices. This probability is then used to calculate the actuarially fair premium rate which is important for an efficient insurance program. Therefore, it must be evaluated accurately as an under- or over-evaluation could lead to several adverse consequences such as distorting the insurance demand and supply and harming the business of insurers. Thus, the dependence between yield and price risks must be taken into account when designing revenue insurance products. In the case of negative dependence, natural hedge would have the potential to make premiums less expensive for producers, on the one hand, and less costly for insurers, on the other hand, as the number of claims would be lower.

Most of existing studies estimate the natural hedge, more generally the correlation between yield and price, in terms of Pearson correlation coefficient (Finger, 2012; Embrechts et al., 2002; Coble et al., 2000). However, there is little evidence that the dependence between these two risks is linear. Alternatively, the correlation was estimated with non-parametric measures

of association such as the Spearman rank correlation (Finger, 2012; Ramsey et al., 2019). The latter is invariant to monotonic transformations and does not rely on a linearity assumption. However, these correlations are limited as they do not characterise the dependence of risks in the distribution. Insurance contracts are often affected by the occurrence of extreme events, such as the severe drought that happened in France in 2016 (Ben-Ari et al., 2018). Moreover, the joint distribution between yields and prices must be calculated for the insurance pricing issue discussed earlier. This joint distribution and the associated dependence structure can be characterised by copula functions. Copulas present a framework for considering both linear and non-linear dependence as well as the dependence of the distribution tails. They have been widely used to analyse financial series and risk factors. In the case of revenue insurance, a small number of studies have been developed such as Duarte and Ozaki (2019); Ramsey et al. (2019); Rusyda et al. (2021); Ahmed and Serra (2015).

Based on this background, the objective of this study is to estimate the natural hedge using copulas. We examine different parametric copulas for modelling the joint distribution between crop yields and prices. Then, to highlight the relevance of using natural hedge in the design of a revenue insurance, we provide empirical analyses to analyse the effect of natural hedge on revenue variability, insurance premiums and claims. Finally, we propose an insurance solution by integrating yields and prices correlations in the pricing, through simulations.

Our analysis is based on a French farm income dataset extracted from the Farm Accountancy Data Network (FADN). This choice is motivated by the fact that there is currently no revenue insurance scheme in France for farmers. In addition, existing insurance policies, such as multi-peril crop insurance, provide low levels of coverage despite the significant subsidies (Lidsky et al., 2017). A better understanding of the dependence structure between prices and yields, especially of natural hedge effects, would improve analyses for the potential implementation of a revenue insurance contract for French farmers. Our empirical analyses are based on the wheat, maize and wine productions, which represent the three most important crops in French agriculture.

5.2 Methodology

This section provides a brief introduction to the copula tool used to model the dependence structure between yields and prices (see State of the Art Section 1.2 for more details). It then presents the procedure for modelling indemnities and setting premiums in a revenue insurance scheme.

5.2.1 Measures of dependence and Copulas

By definition, a copula is a multivariate distribution function with standard uniform univariate margins. Thus it contains all the information on the dependence structure of the model. For the sake of simplicity, let us focus on the bivariate case, in which we consider a pair of continuous random variables X and Y marginally distributed according to $F(x) = \mathbb{P}(X \leq x)$ and $G(y) = \mathbb{P}(Y \leq y)$. Let $H(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ be their joint distribution function. According to Sklar's theorem, Sklar (1959), there exists a unique function $C : [0, 1]^2 \to [0, 1]$ such that:

$$H(x, y) = C(F(x), G(y)), \qquad \text{for all } (x, y) \in \mathbb{R}^2. \tag{5.1}$$

The function C is referred to as the copula associated with H. It is the bivariate cumulative distribution function (cdf) of the random vector $(F(X), G(Y))$ with uniform margins on $[0, 1]$. The mappings $X \mapsto U := F(X)$ or $Y \mapsto V := G(Y)$ used in the above representation are usually referred to as the probability-integral transformations (to uniformity) and are standard tools for simulation purposes.

Several measures of association between the components of a random pair can be considered, Kendall's Tau (Nelsen, 2007) [paragraph 5.1.1], and Spearman's Rho (Nelsen, 2007) [paragraph 5.1.2] being the most popular ones. These measures are invariant to strictly increasing functions and can be interpreted as probabilities of concordance minus probabilities of discordance of two random pairs. Both of them can be written only in terms of the copula C:

$$\tau = 4 \int_0^1 \int_0^1 C(u, v)\, dC(u, v) - 1, \tag{5.2}$$

$$\rho = 12 \int_0^1 \int_0^1 C(u, v)\, du dv - 3. \tag{5.3}$$

Let us note that ρ coincides with the correlation coefficient between the uniform marginal distributions. Starting from a sample $(U_1, V_1), \ldots, (U_n, V_n)$ of independent observations from C, it can be estimated by its empirical counterpart as

$$\hat{\rho} = \frac{12}{n} \sum_{i=1}^{n} U_i V_i - 3. \tag{5.4}$$

A similar formula holds for $\hat{\tau}$.

Three main approaches have been proposed for estimating copulas: parametric, semi-parametric and non-parametric methods (Genest and Favre, 2007). In our context, we advocate for the parametric approach which is based on the estimation of the parameters(s) θ of

the copula assumed to belong to some parametric family $\{C_\theta,\ \theta \in \Theta\}$.

Numerous parametric families of copulas can be found in the literature. Let us focus on two popular models: Elliptical and Archimedean copulas. Elliptical copulas (Frahm et al., 2003) are built from elliptical distributions thanks to an uniformisation of their margins. The level sets of an elliptical distribution density are ellipses whose shape is determined by a (kind of) covariance matrix. Important examples in this family are the Gaussian and the Student copulas. Archimedean copulas (Naifar, 2011) are determined by a univariate function, called the generator, whatever the dimension is. A number of generators have been proposed, involving on one or two parameters and tuning the dependence strength between the marginals. In this study, we focus on three Archimedean copulas: Gumbel, Frank, and Clayton (Beck, 2015) [pages 17-21]. The estimation of the parameter(s) θ can be done for instance using the maximum likelihood method or the method of moments (Mazo et al., 2015). In the latter case, θ is estimated by minimising a given distance between the empirical $\hat{\tau}$ and $\hat{\rho}$ computed from Equation (1.17) and the theoretical ones $\tau(\theta)$ and $\rho(\theta)$ calculated according to Equations (1.3) and (1.4) under the model C_θ.

One technique can be used to select the copula that fits best a dataset is the goodness-of-fit test based on Kendall's distribution function K (also called multivariate probability integral transformation) (Genest et al., 2009) defined as

$$K(t) = \mathbb{P}(C(U,V) \leq t) = \int_0^1 \int_0^1 \mathbb{1}_{\{C(u,v)\leq t\}} dC(u,v). \tag{5.5}$$

The theoretical Kendall distribution under the null hypothesis $(U,V) \sim C_0$ is compared to its sample version thanks to a Cramér–von Mises statistics (Genest and Favre, 2007) and the associated p-value is computed.

5.2.2 Indemnity modelling and pricing in revenue insurance

In the literature, revenue insurance rate calculation adopts three main steps: fitting marginal distribution functions of yields and prices, selecting the best fitting Copula model, and Monte Carlo simulation (Zhu et al., 2008; Walters and Preston, 2018; Ahmed and Serra, 2015). These three steps are explained below.

First, modelling observed crop yields and prices over time involves adjusting for trends. Indeed, yields and prices series can exhibit deterministic changes, which can influence the modelling of the dependence structure. For instance, in the case where decreasing price levels accompany increasing yield levels, the natural hedge could be overestimated. The most common approach in agricultural economics in practice for estimating the trend is the "ad hoc" detrending method (Finger, 2012; Zhu et al., 2008). The approach consists first of

detrending the time series of yields and prices data. To this end, we regress yields and prices on a quadratic time trend given by $y_t = c + a_1 t + a_t t^2 + \epsilon_t$, where y_t is the observation data in year t. Then, residuals $\hat{\epsilon}_t$ and predicted observations for the year of reference 2017, $\hat{y}_{2017}$, can be calculated and detrended data are given by the following:

$$\tilde{y}_t = \hat{y}_{2017} \left(1 + \frac{\hat{\epsilon}_t}{\hat{y}_t}\right). \tag{5.6}$$

In our study, we apply a robust regression, MM estimation, to detrend yields and price (Finger, 2012). The latter is an important tool for the analysis of data affected by outliers or when the distribution of residual is not normal. One of the robust regression estimation methods is M estimation (Huber, 1973), which is an extension of the maximum likelihood estimation method and a robust estimate. S-estimation (Rousseeuw and Yohai, 1984), based on the scale of the residuals of the M-estimate, is also provided in regression robust. Then it comes MM estimation procedure which combines both S estimation and M estimation procedure (Yohai, 1987).

The second stage of the "ad hoc" approach is to estimate the marginal distribution of yield and prices using detrended data as a reference. There are several approaches, parametric and non-parametric, to crop modelling in the agricultural economics literature (see Section 1.1.4 for an overview of these methods). We focus here on the parametric approach and select the main distributions used in practice. Then we estimate the parameters and assess the goodness of fit. The candidate distributions are: Normal (Ozaki et al., 2008), Lognormal (Stokes, 2000), Skewnormal (Duarte and Ozaki, 2019), Weibull (Chen and Miranda, 2004) and Gamma (Gallagher, 1986) for yield modelling and Normal, Lognormal (Samuelson, 2015) and Burr (Tejeda and Goodwin, 2008) for price modelling.

The second step consists of modelling the dependence structure between crop yields and prices. Indeed, the pricing of revenue insurance policies is an actuarial problem that consists in calculating a joint distribution of yields and prices. This joint distribution corresponds to the revenue distribution and can be calculated using the copulas presented above.

In the third step, a Monte Carlo simulation method is used to derive the actuarially fair premium rate for the revenue insurance contracts (Zhu et al., 2008; Walters and Preston, 2018; Ahmed and Serra, 2015). To this end, we use the United States revenue insurance programme (RA) as a reference in the insurance model used hereafter. Let consider the revenue R as a function of two variables, the yield Y and the price P, whose expression is $R = YP$. The RA indemnity is given by (Zhu et al., 2008):

$$\max\left\{(\lambda_i R^e - R_i), 0\right\}, \tag{5.7}$$

where R_i is total annual revenue with $i \in \{wh, ma, wi\}$ (wheat, maize, wine), $R^e = \mathbb{E}(R_i)$ is the expected revenue and $\lambda_i \in (0, 1)$ is the coverage level percentage. If the actual revenue $R_i \leq \lambda_i R^e$, then the farmer is compensated from the insurer with the amount of $(\lambda_i R^e - R_i)$. Then, the actuarially fair premium, which is equal to the expected loss of this contract (Musshoff et al., 2011), is given by:

$$EL(R_i) = \mathbb{E}\left[(\lambda_i R^e - R_i)\mathbb{1}_{\{R_i \leq \lambda_i R^e\}}\right], \tag{5.8}$$

where $\mathbb{1}_{\{R_i \leq \lambda_i R^e\}}$ is an indicator equal to one if farmer is indemnified, and zero otherwise. Finally, we perform a Monte Carlo method to simulate the yields and prices based on the selected copula. From the parameters estimate of this copula, two sets of random sequences U and V are generated as the distribution of the unit yield and price. Then, using the inverse transformation of the estimated cumulative distribution function $F_y^{-1}(U)$ and $F_p^{-1}(V)$, we obtain the yield and price series. The average of the revenue series, obtained by multiplying yields and prices, is used as the expected revenue value R^e. Thus, the expected loss is derived from equation (5.8). We repeat this process 10,000 times to derive the revenue insurance premium using Monte Carlo simulations.

5.2.3 Data

Our analysis is based on French farm income dataset extracted from the FADN (See Section 1.1.5) over the period 2000-2017. We focus on wheat, maize and wine (excluding champagne) productions, which represent the three most important crops in French agriculture. Within the original dataset, we select only the farms that appeared for at least 10 years in the sample during the period 2000-2017. Our sample finally includes 615 wheat producers (with a total of 8480 annual observations), 299 maize producers and 410 wine producers (with a total of 5472 annual observations). The main variables used into the analysis are given in the Table 5.1. We note that the variability of revenue (as well as price and yield) is calculated by considering the growth rate Δ_R between each year, defined as:

$$\Delta_R = (R_N - R_{N-1})/R_{N-1}.$$

To remove the effect of farm size, most variables are standardised by dividing them by the cultivated area. We also lagged insurance premiums and claims to avoid endogeneity issues (Goodwin, 1993). Finally, the price series for each crop were deflated using the agricultural producer price index (IPPAP, IPAMPA - Base 2015) available at INSEE[1].

[1] https://www.insee.fr/fr/statistiques/series/109144301

Variable	Unit	Description
Farm specialisation	-	3 types (wheat, maize, quality wine-growing)
Gross product	Euro	Gross product of the considered crop
Gross yield	Quintal or hL	Gross yield of the considered crop
Harvested acreage	ha	Cultivated area of the considered crop
Yield	Quintal or hL/ha	Yield/Acreage
Yield variability	%	Annual variation of yields
Price	Euro/Quintal or hL	Gross product/Quintals or hectolitres sold
Price variability	%	Annual variation of price
Revenue	Euro/ha	Revenue of the farm
Revenue variability	%	Annual variation of revenue
Premiums	Euro/ha	Crop insurance premiums/Cultivated area
Insurance claims	Euro/ha	Crop insurance claims received/Cultivated area

Table 5.1. List of variables

5.3 Natural hedge effects

Intuition suggests that natural hedge has a stabilising effect on farm revenues and a reducing effect on the demand for insurance solutions. However, few empirical studies are showing such an effect. In this study, we investigate the effect of natural hedge on revenue variability and participation in crop insurance.

Yield, price and revenue variability. Annual variations in yield, price, and farm revenue for the three crops, wheat, maize and wine, are presented in Figure 1. This graphical representation shows that revenue variability has broadly the same trend as price variability, which asserts that revenue is strongly affected by financial market risk. Moreover, revenue variability is much higher for cereals than for wine. Since wheat and maize are standard crops, they are highly affected by financial market trends. We can see a peak around 2007-2008 where prices increased by about 176% for wheat and 95% for maize. Indeed, the spike in international cereal prices over 2007 and 2008 constituted a global food crisis. Between 2003 and their peak in mid-2008, international maize and wheat prices roughly doubled (Headey, 2011). Since 2008, price level variations have been more persistent. We can see a significant drop in yields in 2016, accompanied by a decline in revenues of approximately 31% for wheat, 20% for maize and 11% for wine. Indeed, it was a year of severe drought that penalised France's wheat harvest. Contrary to the French situation, the World cereal harvest reached a record level, leading to a drop in prices. Therefore, farmers suffered double losses, low yields and low prices. Wine production was also affected by drought. However, wine prices follow closely the evolution of wine production since it is a local market. Indeed, the wine industry

is a particular market, driven by the concept of "*terroir*"[2] and quality rather than quantity.

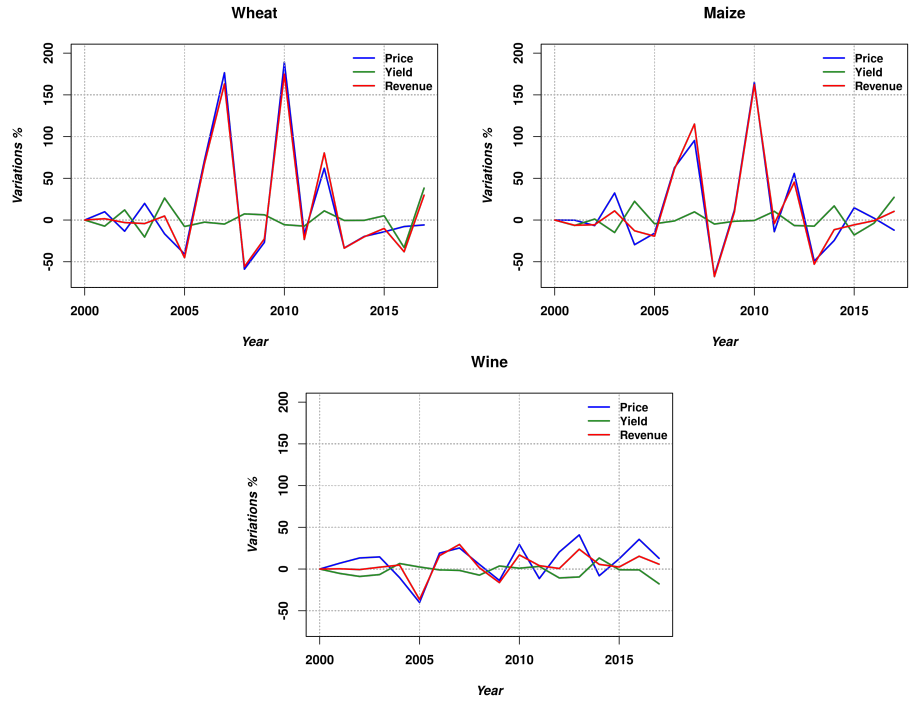


Figure 1 Yield, price and revenue variations for wheat, maize and wine.

Natural Hedge estimate per year and regions. We aim to estimate the correlation between crop yields and prices from 2000 to 2017 to identify the presence of a potential natural hedge over the years. Then we analyse the natural hedge effects according to the regions. We use copulas to model this dependence, and we run a Goodness-of-fit test using Cramér–von Mises statistics, within the time frame 2000-2017, to select the suitable copula model. Table 5.9 summarises the p-values obtained using five parametric copulas (Frank, Clayton, Gumbel, Normal, Student) for each crop. The test shows that Frank copula is the common model for all years for wheat and maize crops since it has the highest p-value. Wine production can also be modelled by a Frank copula for 2000, 2005 and 2014, whereas none of the five copula models shows a matching for the other years. Notwithstanding the insignificant p-value, the Frank copula is assumed to remain suitable for wine since this model allows any dependence structure ranging from complete negative rank correlation to complete positive rank correlation. The parameter estimates of the Frank copula are presented in Table 5.2. This table also shows the empirical Spearman coefficient with a significance test at the 5% level. It can be seen that the Spearman coefficient is consistent with the one computed through the Frank copula parameter. Figure 2 displays the evolution of the natural hedge

[2]The concept of "*terroir*" was first developed in the 14th century in the Bourgogne region of France to identify the qualities of wines in terms of geoclimatic origin and authentic production methods (Whalen, 2009).

over the years. The results show that prices and yields are negatively correlated for wheat production for all years. In contrast, the correlation for maize is variable over the years. Regarding wine, the correlation between yields and prices is mainly negative and particularly strong in 2015-2016. This is due to the slightly high harvest in 2015 with stable stocks, accompanied by low prices for quality wines. For 2016, adverse weather conditions have strongly affected wine and cereal productions.

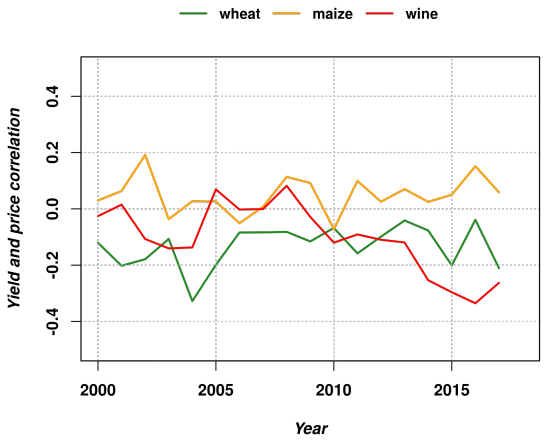


Figure 2 Annual correlations between yields and prices using Frank copula

Figures 4 and 5 present the correlations between crop yields and prices at the regional level with a Frank copula. The latter is used as it covers a wide range of dependence possibilities, including the lower and upper Fréchet bounds and the independent copula. The correlation is represented by the Spearman coefficient computed through the Frank copula. The Figure describes the correlation variation $\hat{\rho_C}$ in French regions where the strong negative (resp. high positives) correlations are in dark green and ranging between $[-1, 0.3[$ (resp. dark red and ranging between $[0.3, 1[$). For example, for maize crop in 2017, the regions that show negative values of $\hat{\rho_C}$ are mainly in Nord-Pas-de-Calais, Provence-Alpes-Côte d'Azur, Lorraine, Ile-de-France and Bourgogne. Indeed, agriculture occupies 53.2% of the Metropolitan France territory and almost 75% in regions such as Nord-Pas-de-Calais and Centre. As these areas represent a significant part of the production, the correlation between prices and yields is expected to be strong. However, the correlation between cereals prices and yields at the regional level is very volatile over the years. Indeed, cereal prices and production quantities move in very different ways. Cereal prices are driven by global market trends, whereas production evolves locally and is strongly influenced by regional weather conditions. Regarding the wine sector, it is shown that natural hedge is more widespread and stable over the years than in the cereal sector.

The occurrence of natural hedge in the graphical representations is consistent with wine

	Wheat				Maize				Wine			
Year	θ	$\hat{\rho}_C$	$\hat{\rho}$	p-value	θ	$\hat{\rho}_C$	$\hat{\rho}$	p-value	θ	$\hat{\rho}_C$	$\hat{\rho}$	p-value
2000	−0.73	−0.12	−0.12*	0.02*	0.18	0.03	0.03	0.74	−0.15	−0.03	−0.03	0.70
2001	−1.24	−0.20	−0.20*	0.00*	0.38	0.06	0.07	0.34	0.09	0.02	0.01	0.94
2002	−1.09	−0.18	−0.17*	0.00*	1.17	0.19	0.19*	0.01*	−0.65	−0.11	−0.12	0.06
2003	−0.64	−0.11	−0.10*	0.03*	−0.22	−0.04	−0.04	0.59	−0.85	−0.14	−0.15*	0.01*
2004	−2.08	−0.33	−0.32*	0.00*	0.16	0.03	0.03	0.66	−0.83	−0.14	−0.15*	0.01*
2005	−1.22	−0.20	−0.20*	0.00*	0.16	0.03	0.03	0.70	0.42	0.07	0.06	0.31
2006	−0.51	−0.08	−0.08	0.07	−0.31	−0.05	−0.05	0.43	−0.01	−0.00	−0.01	0.81
2007	−0.50	−0.08	−0.08	0.06	0.04	0.01	0.01	0.91	−0.00	−0.00	−0.01	0.83
2008	−0.49	−0.08	−0.08*	0.05*	0.69	0.11	0.12*	0.05*	0.49	0.08	0.08	0.11
2009	−0.70	−0.12	−0.12*	0.00*	0.55	0.09	0.09	0.13	−0.17	−0.03	−0.03	0.53
2010	−0.41	−0.07	−0.07	0.11	−0.43	−0.07	−0.07	0.25	−0.72	−0.12	−0.13*	0.02*
2011	−0.96	−0.16	−0.16*	0.00*	0.60	0.10	0.10	0.13	−0.55	−0.09	−0.09	0.10
2012	−0.59	−0.10	−0.10*	0.03*	0.15	0.03	0.02	0.81	−0.66	−0.11	−0.12*	0.02*
2013	−0.25	−0.04	−0.04	0.36	0.42	0.07	0.07	0.32	−0.72	−0.12	−0.13*	0.02*
2014	−0.46	−0.08	−0.07	0.11	0.15	0.02	0.02	0.74	−1.57	−0.25	−0.25*	0.00*
2015	−1.23	−0.20	−0.20*	0.00*	0.30	0.05	0.04	0.54	−1.86	−0.30	−0.30*	0.00*
2016	−0.23	−0.04	−0.04	0.42	0.92	0.15	0.15*	0.04*	−2.13	−0.34	−0.35*	0.00*
2017	−1.29	−0.21	−0.21*	0.00*	0.35	0.06	0.05	0.51	−1.63	−0.26	−0.27*	0.00*

Table 5.2. Yield and price correlations estimates through a Frank copula. θ is the Frank copula parameter, $\hat{\rho}_C$ is the Spearman rank correlation estimated using Frank copula, $\hat{\rho}$ is the Spearman rank correlation estimated empirically and p-value is its associated test significance measure. The null hypothesis H_0 of Spearman rank correlation test assumes that prices and yields are independent and it is considered at the 5% level.

production areas. As the most prominent French wine-growing region, Aquitaine represents more than 30% of the French wine production, with nearly 140 000 hectares of vineyards and a production of about 8 million hectolitres per year. This is most probably the reason why the Spearman correlation coefficient in this region has a high negative value. Alsace, Champagne-Ardenne, Burgundy and Franche-Comté are the regions where wine production is the second most important agricultural product in France. Finally, the results show a strong evidence of natural hedge in the main wine producing regions. This is due to the perfect balance between supply and demand, as the wine market is elastic and influenced by quality and stocks. However, the same cannot be said for wheat and maize, whose prices are closely tied to the evolution of the international market, and yields are related to local production conditions. Consequently, the correlation between prices and yields is erratic. Hence, the need to to devote a special attention to the dependence structure between prices and yields.

Natural hedge effect on revenue variability and crop insurance To study the effect of natural hedge on revenue variability, insurance premiums and claims, we calculate the correlation between prices and yields at the farm level. First, we detrend the yield and price series to remove the time bias (see Section 5.2.2). To do this, we use the estimated trends at the aggregate level, presented in Table 5.4, to detrend the price and yield series at the individual

farm level (Finger, 2012). Then, the correlation estimation is performed for each farm using Frank copula. Figure 3 displays the distribution of the estimated correlation between yields and prices for each crop. As can be seen, the correlation between yields and prices is slightly shifted to the left for wheat and centred around -0.16, while it is symmetric for maize crop. The curve slopes steeply to the left for wine and shows a high concentration in the natural hedge side around -0.58, and a mild one on the right side 0.25. The estimated correlation at the farm level is then used as an explanatory variable for the variability of revenue, insurance premiums and indemnities and the area covered. Since each farm corresponds to a single observation in the dataset, the variables of interest are calculated as the average of the available observations for the period 2000 to 2017 (Finger, 2012). Then, we use several regression models which are presented in the Table 5.3. To reduce the potential influence of outliers, we use a robust regression with an MM estimator. The Natural hedge is negatively correlated with measures of insurance participation (insurance premiums, claims and area covered). The natural hedge does not have the expected effect on insurance premiums. In general, the concept of a natural hedge is expected to reduce insurance premiums, which is not the case here. This can be explained by the inefficiency of existing insurance products, as they do not consider the correlation between prices and yields. Indeed, the current insurance market offers crop insurance products that only cover yield risk. As for the revenue variability, the latter is positively correlated with the natural hedge of wheat and maize and negatively correlated with wine. The negative relationship implies that the higher the natural hedge, the lower the variability. Thus, the natural hedge measure has a moderating effect on the volatility of wine revenues.

These results highlight the organisation of the wine sector again and, conversely, the volatility of the cereal sector, which is influenced by the global financial markets. Then, it seems essential to set up an insurance scheme adapted to each sector's characteristics and consider the natural hedge effect.

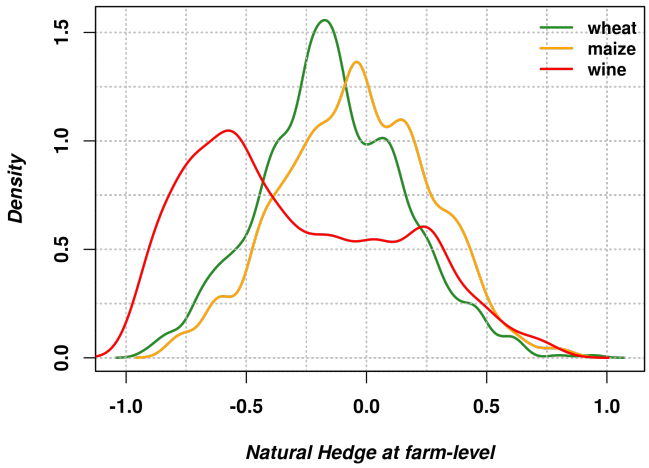


Figure 3 Yield and price correlations at the farm level for wheat, maize and wine-growing.

		Revenue variability (euros/ha)	Log premiums (euros)	Log indemnities (euros)	Log area covered (ha)
Wheat	NH(price, yield)	0.24^{**}	-0.61^{***}	-0.24^{***}	-0.26^{*}
	Constant	0.05^{***}	-3.18^{***}	2.02^{***}	3.91^{***}
	State of influence	Yes	Yes	Yes	Yes
	Observations	615	615	615	615
	R^2	0.30	0.29	0.34	0.36
Maize	NH(price, yield)	0.16^{***}	-0.68^{***}	-0.59^{***}	0.28^{***}
	Constant	0.06^{**}	3.11^{***}	2.25^{***}	3.37^{***}
	State of influence	Yes	Yes	Yes	Yes
	Observations	299	299	299	299
	R^2	0.33	0.34	0.31	0.32
Wine	NH(price, yield)	-0.07^{***}	-0.87^{***}	-0.2^{***}	-0.77^{***}
	Constant	0.04	2.35^{***}	3.82^{***}	4.98^{***}
	State of influence	Yes	Yes	Yes	Yes
	Observations	410	410	410	410
	R^2	0.57	0.48	0.47	0.43

Table 5.3. Regression models explaining observed price and yield correlations at farm level for wheat, maize and wine. Statistical significance at the 10%, 5%, and 1% are denoted by $*$, $**$, and $***$.

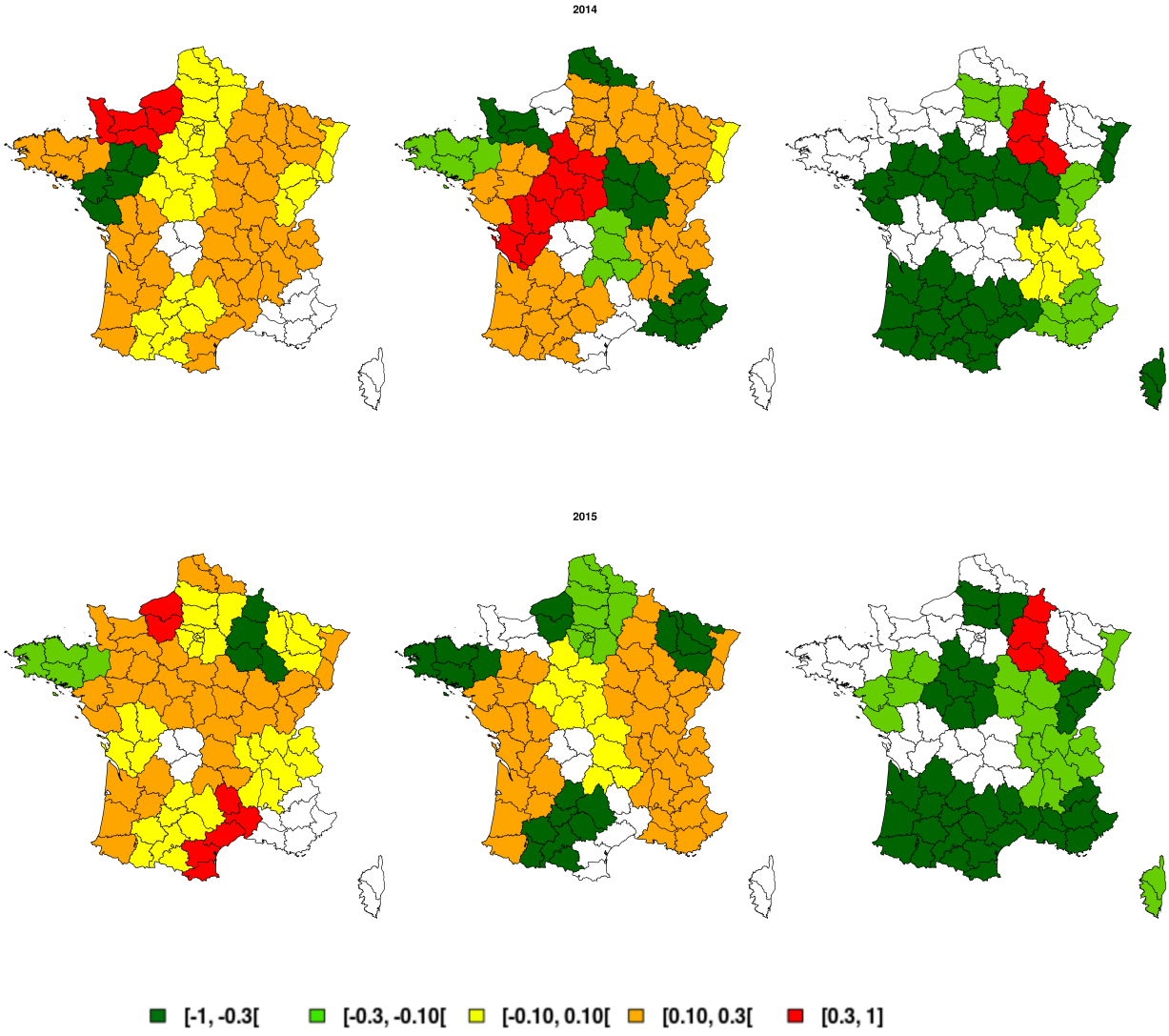


Figure 4 Yield and price correlations at the region level for wheat (first column), maize (second column), wine-growing (third column) for years **2014** (first row), **2015** (second row). Dark green represents large negative correlations. Dark red represents large positive correlations. Regions depicted in white reflect no production or lack of data.

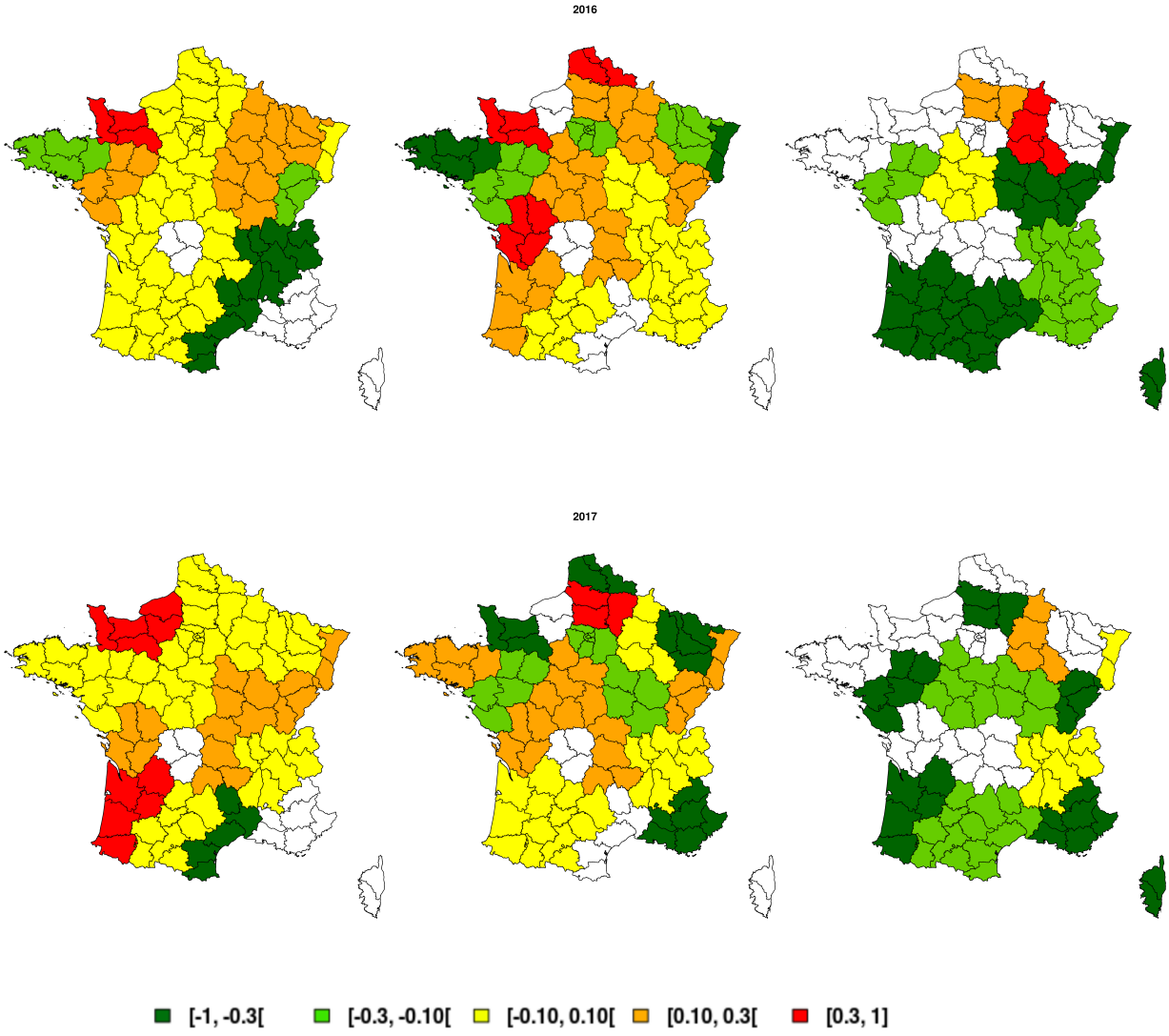


Figure 5 Yield and price correlations at the region level for wheat (first column), maize (second column), wine-growing (third column) for years **2016** (first row), **2017** (second row). Dark green represents large negative correlations. Dark red represents large positive correlations. Regions depicted in white reflect no production or lack of data.

5.4 Design of a revenue insurance

For each crop, we use area-weighted averages for all farms to obtain annual yields and prices. Table 5.4 presents the descriptive statistics for crop yields and prices series and the robust regression model estimates for detrending. The mean of annual wine production prices (360 euros/hectare) and standard deviation (134) implicitly reveal a quite significant fluctuation. Indeed, some wines may meet quality criteria, such as the production region, appellation, vintage, producer reputation and the number of bottles produced, which influence prices and explain the fluctuations. Wheat and maize yields are characterised by a negative value of skewness which indicates that the data are tilted to the left. Except for wheat yields, the normality test confirms the normality of the data. The regression results show a first-order linear trend for all yield and price series except for wheat and maize yields. There is also a second order trend for the wine price series. Although the trends for wheat and maize yields are not significant, they are removed to ensure that we do not overestimate the negative correlations between yields and prices. Figure 6 shows the original and detrended yield and price series. We also test the dependence and heterogeneity of variances over time, using the Ljung Box test, for the detrended series and their squares. It turns out that the series, at the 1% level, do not exhibit these two properties.

		Wheat		Maize		Wine	
		Yields	Prices	Yields	Prices	Yields	Prices
Descriptive statistics	Mean	68.88	15.37	93.64	14.54	46.24	360.17
	Standard deviation	5.22	7.34	7.17	6.73	4.21	134.44
	Skewness	-1.43	1.00	-0.37	1.30	0.88	1.41
	Kurtosis	5.27	3.05	4.15	3.88	1.33	2.53
	Jarque–Bera test	0.01	0.22	0.49	0.06	0.07	0.05
Regression model	Intercept	69.86***	14.25***	94.23***	12.60**	46.26***	358.99
	Trend order 1	1.318	10.26*	8.39	6.84*	-12.58*	392.01***
	Trend order 2	3.863	-5.64	2.25	-1.9	1.45	305.54***
Ljung-Box test	Detrended series	0.24	0.61	0.66	0.57	0.71	0.54
	Detrended series^2	0.21	0.88	0.66	0.68	0.71	0.52

Table 5.4. Descriptive statistics, regression model results and Ljung-Box test for detrended crop yields and prices series. Statistical significance at the 10%, 5%, and 1% are denoted by $*$, $**$, and $***$. The Jarque-Bera test is a normality test based on skewness and kurtosis of the data.

Dependence modelling. Once the time bias in the data has been removed, we estimate the dependence structure between yields and prices using copulas. Table 5.5 summarises the p-values of Goodness of fit for different copulas. As can be seen in the summary, for the five families we tested, Frank copula fits best to wine as it has the highest p-value and is

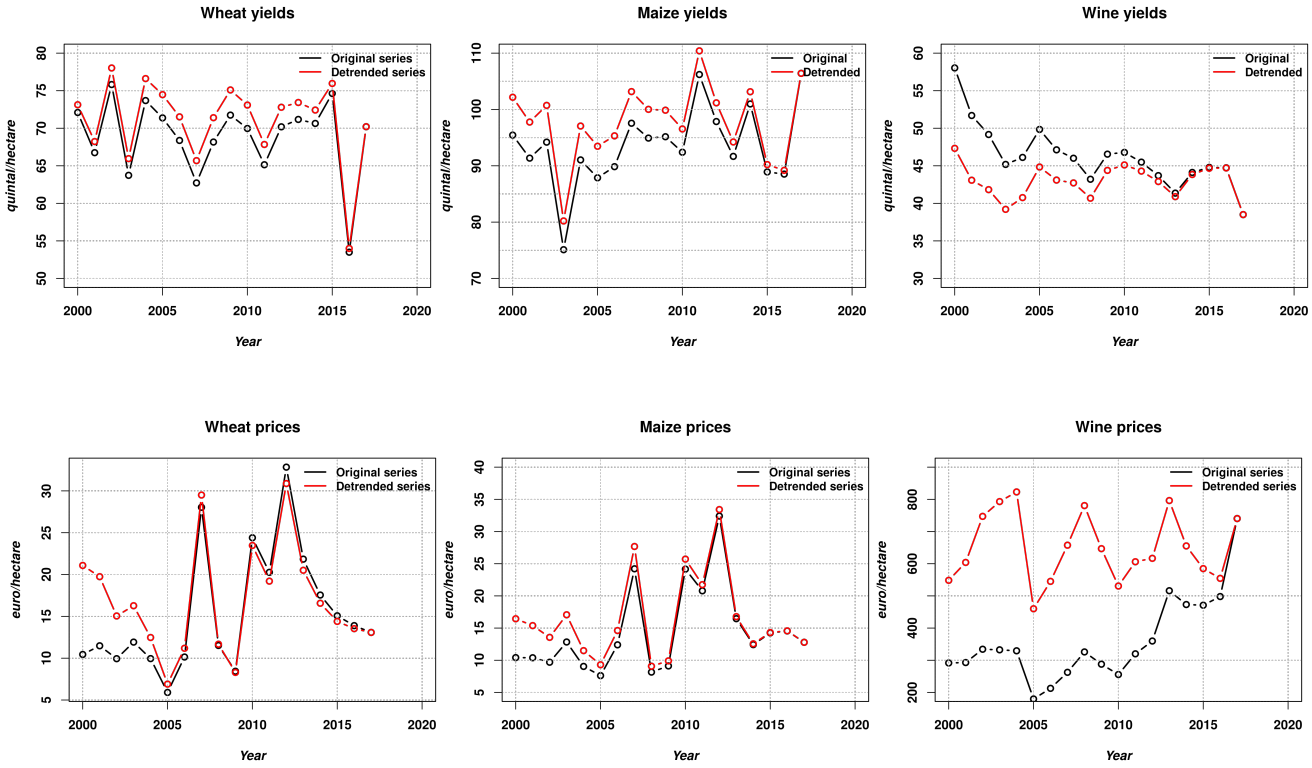


Figure 6 Original and detrended yield and price series

above 5%, while Clayton copula fits best to wheat and maize. The Clayton copula can be an intuitive choice at the aggregate level modelling (area-weighted averages of yields and prices over all farms) as extreme events such as widespread droughts or other natural perils reducing yields would lead to a lower tail dependence. This is in line with the fact that the Clayton copula is an asymmetric archimedean model, with a higher dependence in the negative tail than in the positive tail. Furthermore, the Clayton copula is preferred in the agricultural economics literature when dealing with the aggregate level (regional or national) (Goodwin and Hungerford, 2014). The parameter estimates of the selected copulas models are given in Table 5.6. The estimation results show that prices and yields are negatively correlated for wheat and wine. The correlation is rather positive for maize but not significant at the 5% level.

Crops	Gumbel	Clayton	Frank	Normal	Student
Wheat	0.10	0.78	0.44	0.05	0.04
Maize	0.09	0.23	0.19	0.07	0.07
Wine	0.00	0.13	0.14	0.03	0.02

Table 5.5. Goodness of fit tests p-values of yields and prices correlations. p-values for the test statistic are obtained by means of a multiplier approach (Genest et al., 2009) with 10 000 replications.

Crops	Copula model	Parameter	$\hat{\rho}_c$	ρ	p-value
Wheat	Clayton	-0.28	-0.24	-0.24	0.04
Maize	Clayton	0.13	0.09	0.11	0.07
Wine	Frank	-0.81	-0.80	-0.82	0.00

Table 5.6. Yield and price correlation estimates. ρ is Spearman rank correlation and p-value is its associated test significance measure. The null hypothesis H_0 of Spearman rank correlation test assumes that prices and yields are independent and it is considered at the 5% level. $\hat{\rho}_C$ is Spearman rank correlation estimated using the selected copula model.

Marginal distribution of yields and prices. Table 5.7 presents the fitting procedures for selecting the yield and price models for wheat, maize and wine. To select the model that best fits the data, we used two statistical criteria: Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). For yield, Weibull(k, δ) distribution is chosen for wheat and maize while Normal $\mathcal{N}(\mu, \sigma)$ distribution is selected for wine, for having presented the lowest value for the AIC and BIC statistics. Following the same criteria, the model chosen for the price series is the Lognormal Log-$\mathcal{N}(\mu_i, \sigma_i)$ distribution for wheat $(i = we)$, maize $(i = m)$ and wine $(i = wi)$. The maximum likelihood estimates of the parameters of the yield and price distributions for wheat, maize and wine are respectively: $Y_{we} \sim$ Weibull$(19.44, 73.23)$, $Y_m \sim$ Weibull$(17.01, 100.83)$, $Y_{wi} \sim \mathcal{N}(42.18, 2.12)$, $P_{we} \sim$ Log-$\mathcal{N}(2.75, 0.39)$, $P_m \sim$ Log-$\mathcal{N}(2.73, 0.35)$ and $P_{wi} \sim$ Log-$\mathcal{N}(6.73, 0.15)$.

		Wheat		Maize		Wine	
	Model	AIC	BIC	AIC	BIC	AIC	BIC
Yields	Normal	115.48	117.27	123.818	125.59	83.54	85.32
	Log Normal	118.11	119.89	124.82	126.59	83.84	85.62
	Skew Normal	110.56	111.76	124.68	125.88	85.13	86.33
	Gamma	117.18	118.96	124.45	126.23	83.73	85.51
	Weibull	109.38	111.16	122.87	124.65	83.88	85.66
Prices	Normal	121.76	123.54	122.22	124.01	222.46	224.24
	Log Normal	119.91	121.7	116.22	118.00	222.25	224.03
	Burr	120.96	122.74	120.99	122.77	223.11	224.89

Table 5.7. Information criteria AIC and BIC for marginal distribution of yields and prices

Simulation study and policy implication. In order to estimate the actuarially fair premiums, we perform a Monte Carlo simulation method (presented in Section 5.2). Two scenarios are considered to compare premiums: natural hedge is not taken into account, natural hedge is taken into account. The dependence structure is described by the Frank copula for wine and the Clayton copula for wheat. We use these copulas to draw 10 000 revenues series. The

expected loss, given in equation 5.8, is calculated at different coverage levels. Then we derive the actuarially fair premium, which is the expected loss ratio to the liability. The level of coverage considered is $\lambda_i \in \{70, 80, 90\}\%$ and the year 2017 is taken as a reference for the simulation.

Table 5.8 presents the actuarially fair premium values of a revenue insurance contract for wheat, maize and wine crops in France. The results show that if the price and yield correlation is ignored, the premiums values of wheat revenue insurance at levels 70%, 80% and 90% are respectively 7.45, 14.70 and 32.08 euros/hectare. In contrast, in the case where correlations are used, the fair insurance premiums are 5.99, 12.66 and 25.48 euros/hectare corresponding respectively to the coverage levels of 70%, 80% and 90%. These results show that ignoring the natural hedge leads to an overestimation of the risks by the insurer and thus the premiums will be increased by +20%. Therefore, the demand for insurance solutions would be low. For a coverage level of 80%, the actuarially fair premium for a maize revenue insurance programme is 28.25% euros/hectare when dependence between yields and prices is taken into account, otherwise it is 30.49% euros/hectare. Although the correlation is positive (Spearman coefficient about 0.09), the premiums are nevertheless reduced by 7.35%. This result is suspicious since the statistical test is not significant (see Table5.6). For wine, the actuarially fair premiums are reduced when the natural hedge effect is taken into account. At the 80% coverage level, the premium of revenue insurance is 41% less expensive than an insurance contract without the natural hedge consideration. This is due to the fact that the natural hedge level of wine estimated by Frank copula is very high. The expected losses for wine are quite high because of the strong regional differences. Indeed, the areas of production are more or less expensive depending on the "*terroir*". In case of a renowned terroir, where the cultivation of vines is favourable and qualitative, then the price per hectare will quickly rise.

		Coverage 70%			Coverage 80%			Coverage 90%		
Crops	Revenue insurance	EL	Premium %	Premium	EL	Premium %	Premium	EL	Premium %	Premium
Wheat	With NH	135.26	4.43	5.99	168.48	7.51	12.66	216.17	11.79	25.48
	Without NH	148.24	5.03	7.45	181.36	8.10	14.70	243.20	13.19	32.08
Maize	With NH	245.05	5.74	14.06	288.49	9.79	28.25	340.53	14.41	49.06
	Without NH	249.32	6.36	15.87	295.61	10.32	30.49	352.71	14.56	51.37
Wine	With NH	10803.73	1.26	136.13	12374.17	1.26	156.49	14671.98	1.43	209.81
	Without NH	10815.42	2.02	218.01	17203.47	1.48	427.55	19705.03	1.56	504.14

Table 5.8. Actuarially fair premium rate of a revenue insurance contract. NH is the natural hedge and EL is the expected loss.

5.5 Discussion

In this study, we used several copulas to empirically characterise natural hedging for wheat, maize and wine from 2000 to 2017. We first investigated the impact of natural hedge on revenue variability and insurance participation. The results indicate that the natural hedge has a moderating effect on revenue variability in the wine sector. However, it does not have the expected effect on insurance premiums and claims as well as income variability for wheat and maize, which may denote a lack of efficiency of existing contracts. Second, we assessed whether farm revenue insurance contracts are likely to reduce the cost of insurance purchase when the natural hedge is taken into account. To do this, we modelled the joint distribution of price and yield risks using copulas as well as the marginal distributions. Then we used Monte Carlo simulations that allowed us to obtain the expected revenue, expected loss and then the premium rates. It turns out that revenue insurance provides lower premiums - a decrease of about 13.9% for wheat and 63% for wine with a 80% coverage - than an insurance contract without natural hedge consideration. However, the variation appeared to be smaller for maize, in line with the non- significance of statistical tests on copulas. These results provide an interesting starting point for developing a revenue insurance policy in France that offers better yield and price risk management, particularly for wheat and wine-growing farmers.

Nevertheless, the measurement of the natural hedge degree may differ depending on the level of aggregation, i.e. national, regional or farm level. At a national level, the effect of natural hedge could be overestimated, and the variability of income at the farm level would therefore be underestimated. To overcome the potential problems of an overestimation, the natural hedge should also be quantified at the farm level and incorporated into pricing. This is the subject of our ongoing work.

5.A Appendix

	Wheat					Maize					Wine				
Year	Gumbel	Clayton	Frank	Normal	Student	Gumbel	Clayton	Frank	Normal	Student	Gumbel	Clayton	Frank	Normal	Student
2000	0.0021	0.0231	0.1893*	0.0343	0.0202	0.3834	0.4490	0.6700*	0.3638	0.3041	0.0093	0.0055	0.1014*	0.0032	0.0035
2001	0.0000	0.0027	0.2506*	0.0613	0.0960	0.6065	0.5506	0.8274*	0.5817	0.3496	0.0000	0.0000	0.0036	0.0000	0.0000
2002	0.0000	0.0001	0.1153*	0.0074	0.0375	0.0345	0.2433	0.4622*	0.1073	0.1564	0.0000	0.0000	0.0030	0.0000	0.0000
2003	0.0004	0.0040	0.3331*	0.0709	0.0689	0.0798	0.0382	0.2300*	0.0340	0.0194	0.0000	0.0000	0.0035	0.0000	0.0000
2004	0.0000	0.0000	0.2949*	0.0635	0.0314	0.1270	0.1485	0.4279*	0.1259	0.1434	0.0000	0.0000	0.0018	0.0000	0.0000
2005	0.0000	0.0028	0.1973*	0.0267	0.0311	0.5491	0.6653	0.7999*	0.5292	0.5993	0.0000	0.0000	0.0419*	0.0000	0.0000
2006	0.0009	0.0011	0.2014*	0.0195	0.0267	0.3946	0.2963	0.5422*	0.2291	0.2190	0.0000	0.0000	0.0073	0.0000	0.0000
2007	0.0042	0.1107	0.6356*	0.3264	0.0596	0.5153	0.5686	0.7635*	0.4949	0.4570	0.0000	0.0000	0.0034	0.0000	0.0000
2008	0.0059	0.1373	0.7690*	0.4852	0.5286	0.5011	0.6130	0.9347*	0.7152	0.2975	0.0000	0.0000	0.0336	0.0000	0.0000
2009	0.0000	0.0007	0.0739*	0.0016	0.0023	0.4164	0.1399	0.6270*	0.2899	0.2434	0.0000	0.0000	0.0131	0.0000	0.0000
2010	0.0389	0.2703	0.8128*	0.5552	0.4201	0.0741	0.0773	0.2471*	0.0417	0.0338	0.0000	0.0000	0.0043	0.0000	0.0000
2011	0.0000	0.0587	0.3053*	0.0789	0.1305	0.3847	0.0503	0.5457*	0.1820	0.1922	0.0000	0.0001	0.0113	0.0000	0.0000
2012	0.0045	0.0262	0.4337*	0.1637	0.1800	0.0009	0.0013	0.0800*	0.0006	0.0007	0.0000	0.0000	0.0028	0.0000	0.0000
2013	0.2140	0.2013	0.5198*	0.2065	0.1058	0.2851	0.1985	0.5824*	0.2122	0.4343	0.0000	0.0000	0.0025	0.0000	0.0000
2014	0.0047	0.0468	0.5069*	0.1328	0.1638	0.3775	0.3135	0.6475*	0.3056	0.4882	0.0000	0.0367	0.0525*	0.0005	0.0002
2015	0.0000	0.0000	0.0902*	0.0020	0.0022	0.3030	0.4331	0.6327*	0.3140	0.2532	0.0000	0.0000	0.0062	0.0000	0.0000
2016	0.8339	0.6029	0.8476*	0.6677	0.4307	0.9491	0.4705	0.9343*	0.8545	0.6416	0.0000	0.0000	0.0038	0.0002	0.0000
2017	0.0000	0.0019	0.3060*	0.0848	0.1335	0.4729	0.4086	0.7166*	0.4006	0.5625	0.0000	0.0000	0.0078	0.0000	0.0001

Table 5.9. Goodness of fit tests p-values of yields and prices correlations for considered productions. p-values for the test statistic are obtained by means of a multiplier approach (Genest et al., 2009) with 10 000 replications.

Conclusion and perspectives

The conclusion is organised around the two main topics that have been addressed throughout this thesis: conditional extreme values in high dimensions and dependence modelling in the design of revenue insurance. For each topic, we present a summary of the work carried out and the research perspectives open to us.

Conditional extreme values in high dimensions

Summary of contributions

In the context of conditional extremes, this thesis first introduced a new approach, called Extreme-PLS in Chapter 2. This model combines the partial least squares (PLS) dimension reduction method and distributions tails modelling, in a non-linear inverse regression framework. The main advantage of using Extreme-PLS and inverse regression is to circumvent the well-known curse of dimensionality. Regressing X against Y leads to a one-dimensional regression problem. The latter allows to quantify the effect of covariates X on the extreme values of Y in a simple and interpretable way. Moreover, a visual presentation of the conditional extreme quantiles can be provided. From the theoretical point of view, the asymptotic properties of the Extreme-PLS estimator are established under an inverse single-index model and a heavy tail assumption, without recourse to linearity nor independence assumptions. In addition, the rate of convergence of the proposed estimator is $1/\sqrt{k}$, where k is the number of exceedances (observations considered as extreme), thanks to the dimension reduction. We also considered an iterative procedure in a multi-index setting to estimate other directions of dimension reduction space. Numerical simulations of the Extreme-PLS estimator provides promising results in a high-dimensional setting and also outperforms the proposed estimator in Xu et al. (2020), as soon as the independence assumption is not satisfied. Our method provides a practical and flexible tool for analysing the lowest crop yields by considering the effects of multiple factors whose dimensionality is moderately high.

Second, we have proposed an extension of the Extreme-PLS model in the Bayesian framework in Chapter 3. We retain the Extreme-PLS framework to identify the direction of dimension reduction β by introducing some prior information on it. Our method proposed a Bayesian formulation to compute the posterior distribution of β. We adapted the von Mises-Fisher distribution over the unit sphere (Mardia and Jupp, 2009) to hyperballs for use in characterising the likelihood function of X given Y and β. Then we derived the posterior distribution of β from the likelihood function and a desired prior distribution. We have chosen three possible prior distributions, namely conjugate, hierarchical and sparse. The maximum a posteriori estimator of β is explicit for the conjugate and sparse priors, while it is not for the hierarchical prior. Numerical examples showed that the proposed method is efficient when the sample size is very small. It allowed us to integrate some relevant prior information, provided by experts in the field of agricultural economics, on the French farm income data.

Perspectives

These two new lines of research offer many perspectives in the theoretical and application frameworks. Regarding the first contribution, we could estimate classical risk indicators, such as the conditional tail index and extreme conditional quantiles, thanks to the dimension reduction that overcome the curse of dimensionality. Working on the pair $(\hat{\beta}^t X, Y)$ should yield improved results for most estimators dealing with conditional extreme values. It would then be very interesting to quantify the gain in terms of convergence rate. The Extreme-PLS model can also be extended to general cases with $\gamma \in \mathbb{R}$, i.e. expand to all domains of attraction (Gumbel and Weibull). This extension would offer a wide range of statistical tools for estimating risk measures in the presence of a high dimensional covariate X. For example, we could refer to the results of Daouia et al. (2013), who generalise conditional extreme value estimators for heavy-tailed distributions in any domain of attraction, to introduce the dimension reduction space proposed by our model.

In the multi-index framework, the selection of the number of the most relevant directions is of great importance and deserves further investigation. Another direction to consider is the optimal choice of the threshold y, which is a crucial point. The choice of an optimal threshold boils down to choosing the number of exceedances k. This choice is a difficult task in practice. However, it is has been discussed in several articles in the literature (Guillou and Hall, 2001; Caeiro and Gomes, 2015; Lee et al., 2015).

Within this model, one can also investigate the extreme behaviour of X, i.e. study how the extreme values of Y may depend on the extreme values of X. For example, low yields are often linked to a series of simultaneous extreme events, such as floods and drought accompanied by a price spike in the financial markets. Thus, identifying the joint extreme directions in the

data is crucial for assessing risk. We could rely on the recent work of Meyer and Wintenberger (2020) who deals with tail dependence in a high dimensional context by identifying the most relevant extreme directions.

For the second contribution, we are currently considering the computation of the posterior distribution associated with hierarchical prior of β directly by importance sampling or Monte Carlo Markov chain (Besag, 2001; Gamerman and Lopes, 2006; Hastings, 1970; Metropolis et al., 1953). We are also considering introducing other priors such as an uninformative distribution (Jeffreys, 1946) or other shrinkage priors such as the ridge or elastic-net (Van Erp et al., 2019). As before, it would be interesting to estimate the extreme conditional quantiles and to investigate how introducing the prior information on the dimension reduction direction could sharpen the estimators on small samples. Another important issue to be addressed is the optimal choice of both the concentration parameter of the prior distribution and the number of exceedances. This optimal dual choice is a challenging task in practice and deserves to be investigated (Rootzén and Tajvidi, 2006).

From a practical point of view, we can apply these two models in the analysis of the highest/lowest crop yields depending on other financial and weather variables, in other production sectors such as wheat, maize, wine and over additional years. One could select a wide range of potential factors from the large database of French farm income (FADN) presented in Section 1.1.5. In our application, the developed models allow us to understand and determine the factors that generally influence the smallest crop yields of wheat. We can enhance this application and perform more refined analyses. As an example, in view of climate disruption and its impact on reducing agricultural production, a fruitful line of work would be to collect enough climate data to design a model of the influence of weather parameters on small yields. Using monthly weather indicators to analyse wheat yields, particularly in the fall and winter, could be useful. Another example is to study the effect of chemical input use, such as fertilisers and pesticides, on high yields. Such analysis is important to assess the trade-off between high yields and the potential environmental damage resulting from farmers excessive use of these inputs. This analysis would have direct implications for strengthening the insurance system by providing better coverage. Another important application is to quantify the right tail of cereal prices distributions as function of other variables to tackle the global food security issue effectively. Finally, it would also be interesting to apply these two approaches to a new dataset and to estimate risk measures in other application fields, such as actuarial and financial sciences.

Dependence modelling in the design of revenue insurance

Summary of contributions

In the context of the design of a revenue insurance scheme, we have first focused, in Chapter 4, on evaluating and modelling the risks of wheat, maize and wine production in France over the period 2014-2016. The growing need to model the joint distribution of yield and price risk and their dependence structure motivated this study to use the copula approach. We used various parametric copula models (Normal, Student, Gumbel, Clayton and Frank) and then performed goodness-of-fit tests to select the most suitable one. Frank Copula seems to describe well the dependence between yields and prices, which is relatively high. Then, we analysed the impact of crop insurance purchase and weather indicators, such as temperature and drought, on the correlation between prices and yields using conditional copulas. This correlation is more volatile for wheat, for both insured and uninsured, while it is stable and always negative for wine. While wine prices are determined locally and closely follow yields evolution, global markets drive wheat prices. Therefore, we have shown that the existing crop insurance contracts are more adapted to wine than cereals crops. It is also shown that both cereal and wine production are significantly impacted by extreme weather conditions, especially by the severe drought of 2016. Finally, the results highlight the importance of developing revenue insurance policies in France to improve the hedging of cereal production.

Second, in Chapter 5, we have used Frank copula to characterise empirically the natural hedge in the wheat, maize and wine sector from 2000 to 2017. Then, we studied the impact of natural hedge on revenue variability, insurance premiums and claims. The results showed that the natural hedge measure has a moderating effect on wine but not on wheat and maize. Furthermore, it does not have a reducing impact on insurance participation for all crops. This is due to the inefficiency of existing yield insurance policies that do not hedge prices and therefore do not consider the consequent correlation. Finally, we analysed the natural hedge effect on the value of the actuarially fair premium for pricing a revenue insurance contract. The results indicated that revenue insurance is likely to reduce agricultural insurance premiums in France. For example, at a 80% coverage, revenue insurance offers lower premiums, such a decrease of about 13.9% for wheat, compared to an insurance policy that does not take natural hedge into account.

Perspectives

The first contribution offers many insights, such as an overview of the development and pricing of farm revenue insurance contracts. The latter would allow for better insurance

coverage, especially for cereal producers. The study also provides an overview of developing an income insurance scheme that covers agricultural costs in addition to prices and yields. The analysis carried out could be extended to other years and by considering other explanatory variables derived from weather (temperature, precipitation, drought at monthly scale) or risk management tools (chemical inputs, technology used, insurance measures, etc.), using the multivariate conditional copula. For example, one could analyse the effect of chemical inputs use, such as fertilisers and pesticides, on the interaction between prices and yields regarding the development of organic agriculture. This would provide a better understanding of the income generation of organic farms. This analysis could be used to conduct more in-depth studies on crop insurance policies, including small-scale and organic farmers. One critical point about the data used here is that it is counted annually from a representative sample of farms, which can be considered commercial in size. Therefore, small farms and organic production are neglected as well as some farms that do not appear continuously in the database. The FADN data represents the only opportunity we had at this stage, as the data held by insurers is not freely available.

For the second contribution, our ongoing work aims at improving the design of revenue insurance by pricing policies at the farm level. This is very important to overcome a potential natural hedge overestimation issue and avoid insurance market failures. In the next steps, we could improve the model by integrating other information that could potentially affect the natural hedge, such as cultivated area, financial or weather indicators. We would also like to improve marginal distribution fitting models using other parametric techniques and the non-parametric kernel estimator. Fitting agricultural data with distributions that adequately represent the tails of the low yield and high price distributions deserves to be investigated. Another research topic to be pursued is to model the correlation in the tails of price and yield distributions using extreme copulas and study their effects on revenue insurance pricing.

Bibliography

Aarssen, K. and de Haan, L. (1994). On the maximal life span of humans. *Mathematical Population Studies*, 4(4):259–281. 47

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. Dover Publications. 111

Aghbalou, A., Portier, F., Sabourin, A., and Zhou, C. (2021). Tail inverse regression for dimension reduction with extreme response. *arXiv preprint arXiv:2108.01432*. 5, 58

Ahmad, A. A., Diop, A., and Girard, S. (2019). Estimation of the tail-index in a conditional location-scale family of heavy-tailed distributions. *Dependence Modeling*, 7(1):394–417. 4, 46

Ahmad, A. A., Diop, A., Girard, S., and Usseglio-Carleve, A. (2020). Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model. *Electronic Journal of Statistics*, 14(2):4421–4456. 56, 65

Ahmed, O. and Serra, T. (2015). Economic analysis of the introduction of agricultural revenue insurance contracts in spain using statistical copulas. *Agricultural Economics*, 46(1):69–79. 3, 15, 17, 171, 173, 174

Anandhi, A., Steiner, J. L., and Bailey, N. (2016). A system's approach to assess the exposure of agricultural production to climate change and variability. *Climatic Change*, 136(3-4):647–659. 11, 170

Anderson, P. L. and Meerschaert, M. M. (1998). Modeling river flows with heavy tails. *Water Resources Research*, 34(9):2271–2280. 4, 46

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804. 42

Beale, E., Kendall, M., and Mann, D. (1967). The discarding of variables in multivariate analysis. *Biometrika*, 54(3-4):357–366. 29

Beck, N. (2015). *Multivariate risk measures and a consistent estimator for the orthant based tail value-at-risk*. PhD thesis, Concordia University. 26, 173

Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244. 60

Beirlant, J., Bouquiaux, C., and Werker, B. J. (2006). Semiparametric lower bounds for tail index estimation. *Journal of Statistical Planning and Inference*, 136(3):705–729. 54

Beirlant, J., Broniatowski, M., Teugels, J. L., and Vynckier, P. (1995). The mean residual life function at great age: Applications to tail estimation. *Journal of Statistical Planning and Inference*, 45(1-2):21–48. 54

Beirlant, J., Dierckx, G., Guillou, A., and Staăricaă, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180. 57

Beirlant, J. and Goegebeur, Y. (2003). Regression with response distributions of Pareto-type. *Computational Statistics & Data Analysis*, 42(4):595–619. 56

Beirlant, J. and Goegebeur, Y. (2004). Local polynomial maximum likelihood estimation for Pareto-type distributions. *Journal of Multivariate Analysis*, 89(1):97–118. 56, 64

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons. 69, 77

Beirlant, J. and Teugels, J. L. (1992). Modeling large claims in non-life insurance. *Insurance: Mathematics and Economics*, 11(1):17–29. 48

Beirlant, J., Teugels, J. L., and Vynckier, P. (1996). *Practical analysis of extreme values*, volume 50. Leuven University Press. 53, 54

Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., Van der Velde, M., and Makowski, D. (2018). Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nature Communications*, 9(1):1–10. 2, 18, 19, 171

Bernard-Michel, C., Gardes, L., and Girard, S. (2009). Gaussian regularized sliced inverse regression. *Statistics and Computing*, 19(1):85–98. 68

Besag, J. (2001). Markov chain monte carlo for statistical inference. *Center for Statistics and the Social Sciences*, 9:24–25. 38, 116, 194

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, 10:3–41. 38

Bhattacharya, P. K. and Gangopadhyay, A. K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, 18(3):1400–1415. 64

Billingsley, P. (1995). *Measure and Probability (third edition)*. John Wiley & Sons. 94

Bingham, N., Goldie, C., and Teugels, J. (1987). *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its application*. Cambridge University Press. 43, 44, 68, 83

Bottolo, L., Consonni, G., Dellaportas, P., and Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6(1):25–47. 59

Bousebata, M., Enjolras, G., and Girard, S. (2020). The dependence structure between yields and prices: A copula-based model of French farm income. In *Annual Meeting of the Agricultural and Applied Economics Association*, pages 1–15. AAEA. 142, 143

Bousebata, M., Enjolras, G., and Girard, S. (2021). Extreme partial least-squares regression. 62, 63

Bozic, M., Newton, J., Thraen, C. S., and Gould, B. W. (2014). Tails curtailed: accounting for nonlinear dependence in pricing margin insurance for dairy farmers. *American Journal of Agricultural Economics*, 96(4):1117–1135. 3

Broniatowski, M. (1993). On the estimation of the Weibull tail coefficient. *Journal of Statistical Planning and Inference*, 35(3):349–365. 54

Caeiro, F. and Gomes, M. I. (2015). Threshold selection in extreme value analysis. *Extreme value modeling and risk analysis: Methods and applications*, pages 69–82. 193

Cai, X., Lin, G., and Li, J. (2021). Bayesian inverse regression for supervised dimension reduction with small datasets. *Journal of Statistical Computation and Simulation*, 91:1–16. 109

Castellanos, M. E. and Cabras, S. (2007). A default Bayesian procedure for the generalized pareto distribution. *Journal of Statistical Planning and Inference*, 137(2):473–483. 60

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276. 73

Chatelain, F. and Le Bihan, N. (2013). von Mises-Fisher approximation of multiple scattering process on the hypersphere. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6461–6465. IEEE. 111

Chavas, J. P. (2011). Agricultural policy in an uncertain world. *European Review of Agricultural Economics*, 38(3):383–407. 12

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C*, 54(1):207–222. 55, 56, 64

Chen, S.-L. and Miranda, M. J. (2004). Modeling multivariate crop yield densities with frequent extreme events. In *Annual Meeting of the American Agricultural Economics Association*. AAEA. 174

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics*, 33(2):806–839. 55, 64

Chiancone, A., Forbes, F., and Girard, S. (2017). Student sliced inverse regression. *Computational Statistics & Data Analysis*, 113:441–456. 32, 65, 109

Chuangchid, K., Sriboonchitta, S., Rahman, S., and Wiboonpongse, A. (2013). Predicting Malaysian palm oil price using extreme value theory. *International Journal of Agricultural Management*, 2(2):91–99. 4

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B*, 72(1):3–25. 31, 65, 66, 109, 117

Coble, K. H., Heifner, R. G., and Zuniga, M. (2000). Implications of crop yield and revenue insurance for producer hedging. *Journal of Agricultural and Resource Economics*, pages 432–452. 3, 16, 170

Coble, K. H. and Knight, T. O. (2002). Crop insurance as a tool for price and yield risk management. In Just, R. E. and Pope, R. D., editors, *A comprehensive assessment of the role of risk in US agriculture*, pages 445–468. Springer, Boston, MA. 12

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer. 60

Coles, S., Pericchi, L. R., and Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, 273(1-4):35–50. 4, 49

Coles, S. G. and Powell, E. A. (1996). Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review/Revue Internationale de Statistique*, pages 119–136. 5

Coles, S. G. and Tawn, J. A. (1990). Statistics of coastal flood prevention. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 332(1627):457–476. 59

Coles, S. G. and Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(4):463–478. 60

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, 32(3):1062–1092. 32

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26. 68

Cook, R. D. and Forzani, L. (2018). Big data and partial least-squares prediction. *Canadian Journal of Statistics*, 46(1):62–78. 66

Cook, R. D., Helland, I., and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B*, 75(5):851–877. 31, 65, 109

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100(470):410–428. 30, 58

Cook, R. D. and Ni, L. (2006). Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, 93(1):65–74. 32

Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604. 59, 66

Coudret, R., Girard, S., and Saracco, J. (2014). A new sliced inverse regression method for multivariate response. *Computational Statistics & Data Analysis*, 77:285–299. 32, 65, 109

Daouia, A., Gardes, L., and Girard, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli*, 19(5B):2557–2589. 4, 57, 64, 193

Daouia, A., Gardes, L., Girard, S., and Lekina, A. (2011). Kernel estimators of extreme level curves. *Test*, 20(2):311–333. 4, 57, 64, 79

Davison, A. C. and Ramesh, N. (2000). Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society: Series B*, 62(1):191–208. 4, 56, 64

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B*, 52(3):393–425. 4, 52, 55, 64

de Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer Science & Business Media. 44, 48, 50, 53, 69, 82

de Haan, L. and Resnick, S. (1998). On asymptotic normality of the Hill estimator. *Stochastic Models*, 14(4):849–866. 53

de Mey, Y., Wauters, E., Schmid, D., Lips, M., Vancauteren, M., and Van Passel, S. (2016). Farm household risk balancing: empirical evidence from Switzerland. *European Review of Agricultural Economics*, 43(4):637–662. 12

De Michele, C., Salvadori, G., Canossi, M., Petaccia, A., and Rosso, R. (2005). Bivariate statistical approach to check adequacy of dam spillway. *Journal of Hydrologic Engineering*, 10(1):50–57. 22

de Zea Bermudez, P. and Turkman, M. A. (2003). Bayesian approach to parameter estimation of the generalized Pareto distribution. *Test*, 12(1):259–277. 60

de Zea Bermudez, P., Turkman, M. A., and Turkman, K. (2001). A predictive approach to tail probability estimation. *Extremes*, 4(4):295–314. 60

Deheuvels, P. (1981). A nonparametric test for independence. *Publications de l'Institut de Statistique de l'Université de Paris*, 26:29–50. 27

Deheuvels, P., Haeusler, E., and Mason, D. M. (1988). Almost sure convergence of the Hill estimator. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 104, pages 371–381. Cambridge University Press. 53

Diaz-Caneja, M. B., Conze, C. G., Dittmann, C., Pinilla, F. J. G., and Stroblmair, J. (2008). *Agricultural insurance schemes*. Office for Official Publications of the European Union. 2, 14

Diebolt, J., El-Aroui, M.-A., Garrido, M., and Girard, S. (2005). Quasi-conjugate bayes estimates for gpd parameters and application to heavy tails modelling. *Extremes*, 8(1):57–78. 60

Diebolt, J., Gardes, L., Girard, S., and Guillou, A. (2008). Bias-reduced estimators of the Weibull tail-coefficient. *Test*, 17(2):311–331. 48

Diers, D., Eling, M., and Marek, S. D. (2012). Dependence modeling in non-life insurance using the Bernstein copula. *Insurance: Mathematics and Economics*, 50(3):430–436. 22

Ditlevsen, O. (1994). Distribution arbitrariness in structural reliability. *Structural Safety and Reliability*, pages 1241–1247. 4

Draper, N. R. and Smith, H. (1998). *Applied regression analysis*, volume 326. John Wiley & Sons. 29

Drees, H. and Sabourin, A. (2021). Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908–943. 5, 59, 66

Duarte, G. V. and Ozaki, V. A. (2019). Pricing crop revenue insurance using parametric copulas. *Revista Brasileira de Economia*, 73:325–343. 3, 171, 174

Duong, T. T., Brewer, T., Luck, J., and Zander, K. (2019). A global review of farmers' perceptions of agricultural risks and risk management strategies. *Agriculture*, 9(1):10. 11

Durrans, S. R. (1996). Low-flow analysis with a conditional Weibull tail model. *Water Resources Research*, 32(6):1749–1760. 47

Eidman, V. T. (1990). Quantifying and managing risk in agriculture. *Agrekon*, 29(1):11–23. 11, 13

El Benni, N., Finger, R., and Meuwissen, M. P. (2016). Potential effects of the income stabilisation tool (IST) in Swiss agriculture. *European Review of Agricultural Economics*, 43(3):475–502. 12, 14, 15

El Methni, J., Gardes, L., and Girard, S. (2014). Non-parametric estimation of extreme risk measures from conditional heavy-tailed distributions. *Scandinavian Journal of Statistics*, 41(4):988–1012. 46

El Methni, J., Gardes, L., Girard, S., and Guillou, A. (2012). Estimation of extreme quantiles from heavy and light tailed distributions. *Journal of Statistical Planning and Inference*, 142(10):2735–2747. 4, 46

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media. 4, 41, 53

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk Management: Value at Risk and Beyond*, 1:176–223. 3, 16, 170

Emura, T. and Michimae, H. (2017). A copula-based inference to piecewise exponential models under dependent censoring, with application to time to metamorphosis of salamander larvae. *Environmental and Ecological Statistics*, 24(1):151–173. 22

Engelund, S. and Rackwitz, R. (1992). On predictive distribution functions for the three asymptotic extreme value distributions. *Structural Safety*, 11(3-4):255–258. 59

Enjolras, G. and Sentis, P. (2011). Crop insurance policies and purchases in France. *Agricultural Economics*, 42(4):475–486. 14

Falk, M. (1995). Some best parameter estimates for distributions with finite endpoint. *Statistics: A Journal of Theoretical and Applied Statistics*, 27(1-2):115–125. 47

Feng, S., Patton, M., Binfield, J., and Davis, J. (2014). Uneven natural hedge effects in the wheat sector and implications for risk management tools. *EuroChoices*, 13(3):19–25. 170

Ferrez, J., Davison, A., and Rebetez, M. (2011). Extreme temperature analysis under forest cover compared to an open field. *Agricultural and Forest Meteorology*, 151(7):992–1001. 55

Finger, R. (2012). Effects of crop acreage and aggregation level on price-yield correlations. *Agricultural Finance Review*, 72(3):436–455. 2, 170, 171, 173, 174, 180

Finger, R. and El Benni, N. (2014). Alternative specifications of reference income levels in the income stabilization tool. In *Agricultural Cooperative Management and Policy*, pages 65–85. Springer. 14

Finocchio, R. and Esposti, R. (2008). Determinants of farm diversification and interaction with the CAP. An application to FADN of Marche region (Italy). In *2008 International Congress*. European Association of Agricultural Economists. 13

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press. 39, 41

Foudi, S. and Erdlenbruch, K. (2012). The role of irrigation in farmers' risk management strategies in France. *European Review of Agricultural Economics*, 39(3):439–457. 13

Fousekis, P. and Grigoriadis, V. (2017). Joint price dynamics of quality differentiated commodities: copula evidence from coffee varieties. *European Review of Agricultural Economics*, 44(2):337–358. 3

Frahm, G., Junker, M., and Szimayer, A. (2003). Elliptical copulas: applicability and limitations. *Statistics & Probability Letters*, 63(3):275–286. 25, 173

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135. 31, 65

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. In *Annales de la société Polonaise de Mathématique*, volume 6, pages 93–116. 41

Fretheim, T. and Kristiansen, G. (2015). Commodity market risk from 1995 to 2013: an extreme value theory approach. *Applied Economics*, 47(26):2768–2782. 17

Gabriel, S. C. and Baker, C. B. (1980). Concepts of business and financial risk. *American Journal of Agricultural Economics*, 62(3):560–564. 12

Galambos, J. (1977). The asymptotic theory of extreme order statistics. In *The Theory and Applications of Reliability with Emphasis on Bayesian and Nonparametric Methods*, pages 151–164. Elsevier. 48

Gallagher, P. (1986). US corn yield capacity and probability: estimation and forecasting with nonsymmetric disturbances. *North Central Journal of Agricultural Economics*, 8(1):109–122. 17, 174

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press. 38, 116, 194

García Azcárate, T., Sumpsi, J. M., Capitanio, F., Garrido, A., Felis, A., Blanco, I., Enjolras, G., and Bardají, I. (2016). State of play of risk management tools implemented by Member States during the period 2014-2020: national and European frameworks. Technical report. 13, 15

Gardes, L. (2018). Tail dimension reduction for extreme quantile estimation. *Extremes*, 21(1):57–95. 5, 58, 66, 67, 68, 109

Gardes, L. and Girard, S. (2008a). Estimation of the Weibull tail-coefficient with linear combination of upper order statistics. *Journal of Statistical Planning and Inference*, 138(5):1416–1427. 54

Gardes, L. and Girard, S. (2008b). A moving window approach for nonparametric estimation of the conditional tail index. *Journal of Multivariate Analysis*, 99(10):2368–2388. 4, 57

Gardes, L. and Girard, S. (2010). Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2):177–204. 4, 46, 56, 64

Gardes, L. and Girard, S. (2011). Comparison of Weibull tail-coefficient estimators. *Revstat – Statistical Journal*, 4(2):163–188. 54

Gardes, L. and Girard, S. (2012). Functional kernel estimators of large conditional quantiles. *Electronic Journal of Statistics*, 6:1715–1744. 58, 64

Gardes, L. and Girard, S. (2013). Estimation de quantiles extrêmes pour les lois à queue de type Weibull: une synthèse bibliographique. *Journal de la Société Française de Statistique*, 154(2):98–118. 48

Gardes, L., Girard, S., and Guillou, A. (2011). Weibull tail-distributions revisited: a new look at some tail estimators. *Journal of Statistical Planning and Inference*, 141(1):429–444. 48

Gardes, L., Girard, S., and Lekina, A. (2010). Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis*, 101(2):419–433. 58

Gardes, L. and Stupfler, G. (2014). Estimation of the conditional tail index using a smoothed local Hill estimator. *Extremes*, 17(1):45–75. 57

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409. 38

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6(6):721–741. 38

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74(3):549–555. 26

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368. 22, 27, 28, 172, 173

Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213. 28, 173, 185, 190

Gijbels, I., Veraverbeke, N., and Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis*, 55(5):1919–1932. 28

Girard, S. (2004). A Hill type estimator of the Weibull tail-coefficient. *Communications in Statistics-Theory and Methods*, 33(2):205–234. 48, 54

Girard, S., Guillou, A., and Stupfler, G. (2012). Estimating an endpoint with high order moments in the Weibull domain of attraction. *Statistics & Probability Letters*, 82(12):2136–2144. 47

Girard, S., Lorenzo, H., and Saracco, J. (2022). Advanced topics in sliced inverse regression. *Journal of Multivariate Analysis*, 188:104852. 65

Girard, S., Stupfler, G., and Usseglio-Carleve, A. (2021). Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models. *The Annals of Statistics*, 49(6):3358–3382. 82

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3):423–453. 39, 41

Goegebeur, Y., Guillou, A., and Schorgen, A. (2014). Nonparametric regression estimation of conditional tails: the random covariate case. *Statistics*, 48(4):732–755. 4, 57, 64

Goodwin, B. K. (1993). An empirical analysis of the demand for multiple peril crop insurance. *American Journal of Agricultural Economics*, 75(2):425–434. 175

Goodwin, B. K. and Hungerford, A. (2014). Copula-based models of systemic risk in US agriculture: implications for crop insurance and reinsurance contracts. *American Journal of Agricultural Economics*, 97(3):879–896. 3, 12, 185

Goodwin, B. K. and Schroeder, T. C. (1994). Human capital, producer education programs, and the adoption of forward-pricing methods. *American Journal of Agricultural Economics*, 76(4):936–947. 13

Guillou, A. and Hall, P. (2001). A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B*, 63(2):293–305. 193

Gumbel, E. J. (1941). The return period of flood flows. *The Annals of Mathematical Statistics*, 12(2):163–190. 48

Gumbel, E. J. (1958). Statistics of extremes. *Columbia university press*. 41, 51

Haeusler, E. and Teugels, J. L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *The Annals of Statistics*, 13(2):743–756. 53

Hall, P. and Park, B. U. (2002). New methods for bias correction at endpoints and boundaries. *The Annals of Statistics*, 30(5):1460–1479. 47

Hall, P. and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, 15(2):153–167. 4, 56, 65

Hardaker, J. B., Lien, G., Anderson, J. R., and Huirne, R. B. (2015). *Coping with risk in agriculture: Applied decision analysis*. 11, 12, 13, 170

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995. 32, 65

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning. Springer series in statistics*. Springer. 29, 31, 66

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109. 38, 116, 194

He, X., Ng, P., and Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B*, 60(3):537–550. 64

Headey, D. (2011). Rethinking the global food crisis: The role of trade shocks. *Food Policy*, 36(2):136–146. 1, 11, 176

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 17(2):97–114. 31, 65, 73

Hennessy, D. A., Babcock, B. A., and Hayes, D. J. (1997). Budgetary and producer welfare effects of revenue insurance. *American Journal of Agricultural Economics*, 79(3):1024–1034. 2, 15

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174. 53, 77

Hine, R., Ingersent, K. A., and Rayner, A. (2016). *The reform of the common agricultural policy*. Springer. 14

Hocking, R. R. and Leslie, R. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540. 29

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82. 30, 66

Hofert, M. and Mächler, M. (2014). A graphical goodness-of-fit test for dependence models in higher dimensions. *Journal of Computational and Graphical Statistics*, 23(3):700–716. 27

Horowitz, J. L. (2009a). *Semiparametric and nonparametric methods in econometrics*. New York: Springer. 32

Horowitz, J. L. (2009b). *Semiparametric and nonparametric methods in econometrics*, volume 12. Springer. 65

Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3):339–349. 52

Hosking, J. R., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261. 52

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821. 174

Jagger, T. H. and Elsner, J. B. (2006). Climatology models for extreme hurricane winds near the United States. *Journal of Climate*, 19(13):3220–3236. 49

Jagger, T. H. and Elsner, J. B. (2009). Modeling tropical cyclone intensity with quantile regression. *International Journal of Climatology*, 29(10):1351–1361. 64

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461. 35, 122, 194

Johnson, D. G. (1975). World agriculture, commodity policy, and price variability. *American Journal of Agricultural Economics*, 57(5):823–828. 12

Kang, M. G. (2007). *Innovative agricultural insurance products and schemes*, volume 12. Food & Agriculture Org. 2, 170

Katz, R. W. (1999). Extreme value theory for precipitation: sensitivity analysis for climate change. *Advances in Water Resources*, 23(2):133–139. 4

Ker, A. P. and Coble, K. (2003). Modeling conditional yield densities. *American Journal of Agricultural Economics*, 85(2):291–304. 17

Ker, A. P. and Goodwin, B. K. (2000). Nonparametric estimation of crop insurance rates revisited. *American Journal of Agricultural Economics*, 82(2):463–478. 17

Kim, K. and Chavas, J.-P. (2003). Technological change and risk management: an application to the economics of corn production. *Agricultural Economics*, 29(2):125–142. 13

Kimura, S., Antón, J., and LeThi, C. (2010). Farm level analysis of risk and risk management strategies and policies: cross country analysis. 170

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 46(1):33–50. 64

Koenker, R., Chesher, A., Society, E., and Jackson, M. (2005). *Quantile Regression*. Cambridge University Press. 59

Komarek, A. M., De Pinto, A., and Smith, V. H. (2020). A review of types of risks in agriculture: What we know and what we need to know. *Agricultural Systems*, 178:102738. 11

Kong, E. and Xia, Y. (2012). A single-index quantile regression model and its estimation. *Econometric Theory*, 28(4):730–768. 32, 65

Kyung-Joon, C. and Schucany, W. R. (1998). Nonparametric kernel regression estimation near endpoints. *Journal of statistical Planning and Inference*, 66(2):289–304. 79

Lee, J., Fan, Y., and Sisson, S. A. (2015). Bayesian threshold selection for extremal models using measures of surprise. *Computational Statistics & Data Analysis*, 85:84–99. 193

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327. 31, 58, 65, 109

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613. 32

Li, L., Cook, R., and Nachtsheim, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society: Series B*, 67(2):285–299. 32

Li, L., Cook, R. D., and Tsai, C.-L. (2007). Partial inverse regression. *Biometrika*, 94(3):615–625. 32, 65, 109

Lidsky, V., Maudet, C., Malpel, G., Gerster, F., Helfter, M., Lejeune, H., and Theule, F. (2017). Les outils de gestion des risques en agriculture. Technical report. 3, 15, 171

Makus, L. D., Lin, B.-H., Carlson, J., and Krebill-Prather, R. (1990). Factors influencing farm level use of futures and options in commodity marketing. *Agribusiness*, 6(6):621–631. 13

Mao, K., Liang, F., and Mukherjee, S. (2010). Supervised dimension reduction using Bayesian mixture modeling. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, volume 9, pages 501–508. JMLR Workshop and Conference Proceedings. 109

Mardia, K. (1975). Distribution theory for the von Mises-Fisher distribution and its application. In *A Modern Course on Statistical Distributions in Scientific Work*, pages 113–130. Springer. 110

Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*, volume 494. John Wiley & Sons. 110, 193

Martens, H. and Naes, T. (1992). *Multivariate calibration*. John Wiley & Sons. 31, 65

Mary, S., Santini, F., and Boulanger, P. (2013). An ex-ante assessment of cap income stabilisation payments using a farm household model. Agricultural Economics Society. 14

Mason, D. M. (1982). Laws of large numbers for sums of extreme values. *The Annals of Probability*, 10(3):754–764. 53

Matthys, G., Delafosse, E., Guillou, A., and Beirlant, J. (2004). Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance: Mathematics and Economics*, 34(3):517–537. 46

Mazo, G., Girard, S., and Forbes, F. (2015). Weighted least-squares inference for multivariate copulas based on dependence coefficients. *ESAIM: Probability and Statistics*, 19:746–765. 173

McNeil, A. J., Frey, R., Embrechts, P., et al. (2005). *Quantitative risk management: Concepts, techniques and tools*, volume 3. Princeton university press. 49

Meligkotsidou, L., Vrontos, I. D., and Vrontos, S. D. (2009). Quantile regression analysis of hedge fund strategies. *Journal of Empirical Finance*, 16(2):264–279. 55, 64

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092. 38, 116, 194

Meuwissen, M. P., Huirne, R. B., and Skees, J. R. (2003). Income insurance in European agriculture. *EuroChoices*, 2(1):12–17. 2, 14, 15

Meuwissen, M. P., Mey, Y., and van Asseldonk, M. (2018). Prospects for agricultural insurance in Europe. *Agricultural Finance Review*, 78(2):174–182. 14

Meuwissen, M. P., Van Asseldonk, M. A., Pietola, K., Hardaker, J. B., and Huirne, R. (2011). Income insurance as a risk management tool after 2013 CAP reforms? Technical report. 14

Meyer, N. and Wintenberger, O. (2020). Tail inference for high-dimensional data. *arXiv e-prints*, pages arXiv–2007. 194

Mitchell, E. G., Crout, N. M., Wilson, P., Wood, A. T., and Stupfler, G. (2020). Operating at the extreme: estimating the upper yield boundary of winter wheat production in commercial practice. *Royal Society Open Science*, 7(4):191919. 4, 17

Morgan, W., Cotter, J., and Dowd, K. (2012). Extreme measures of agricultural financial risk. *Journal of Agricultural Economics*, 63(1):65–82. 4, 17

Moschini, G. and Hennessy, D. A. (2001). Uncertainty, risk aversion, and risk management for agricultural producers. *Handbook of Agricultural Economics*, 1:87–153. 1, 11, 170

Musshoff, O., Odening, M., and Xu, W. (2011). Management of climate risks in agriculture–will weather derivatives permeate? *Applied economics*, 43(9):1067–1077. 175

Naifar, N. (2011). Modelling dependence structure with Archimedean copulas and applications to the iTraxx CDS index. *Journal of Computational and Applied Mathematics*, 235(8):2459–2466. 26, 173

Naik, P. and Tsai, C.-L. (2000). Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B*, 62(4):763–771. 66

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media. 23, 24, 74, 118, 172

Northrop, P. J. and Attalides, N. (2016). Posterior propriety in Bayesian extreme value analyses using reference priors. *Statistica Sinica*, pages 721–743. 60

Nunez-Antonio, G. and Gutiérrez-Pena, E. (2005). A Bayesian analysis of directional data using the von mises-fisher distribution. *Communications in Statistics-Simulation and Computation*, 34(4):989–999. 114

Ortmann, G., Patrick, G., Musser, W., and Doster, D. (1992). Sources and management of risk: Evidence from leading cornbelt farmers in the USA. *Agrekon*, 31(4):216–221. 1, 12, 170

Ozaki, V. A., Goodwin, B. K., and Shirota, R. (2008). Parametric and nonparametric statistical modelling of crop yield: implications for pricing crop insurance contracts. *Applied Economics*, 40(9):1151–1164. 17, 174

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131. 39, 42

Pickands, J. (1994). Bayes quantile estimation and threshold selection for the generalized Pareto family. In *Extreme value theory and applications*, pages 123–138. Springer. 60

Pisarenko, V. and Sornette, D. (2003). Characterization of the frequency of extreme earthquake events by the generalized Pareto distribution. *Pure and Applied Geophysics*, 160(12):2343–2364. 55

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430. 32, 65

Prescott, P. and Walden, A. (1980). Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, 67(3):723–724. 52

Prescott, P. and Walden, A. (1983). Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples. *Journal of Statistical Computation and Simulation*, 16(3-4):241–250. 52

Purcell, W. D., Koontz, S. R., et al. (1991). *Agricultural futures and options: principles and strategies*. New York: Maxwell Macmillan International. 14

Ramirez, O. A., Misra, S., and Field, J. (2003). Crop-yield distributions revisited. *American Journal of Agricultural Economics*, 85(1):108–120. 17

Ramsey, A. F., Goodwin, B. K., and Ghosh, S. K. (2019). How high the hedge: relationships between prices and yields in the federal crop insurance program. *Journal of Agricultural and Resource Economics*, 44(1835-2019-1539):227–245. 2, 3, 16, 170, 171

Reich, B. J., Bondell, H. D., and Li, L. (2011). Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics*, 67(3):886–895. 109

Rejda, G. E. (2011). *Principles of risk management and insurance*. Pearson Education India. 13

Resnick, S. (1987). *Extreme values, regular variation, and point processes*. Springer, New York. 48

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer. 33

Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930. 194

Rostami, M. and Adam, M. B. (2013). Analyses of prior selections for gumbel distribution. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 29:95–107. 60

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis*, volume 26, pages 256–272. Springer. 174

Rusyda, H., Noviyanti, L., Soleh, A., Chadidjah, A., and Indrayatna, F. (2021). The design of multiple crop insurance in Indonesia based on revenue risk using the copula model approach. *Journal of Applied Statistics*, 48:1–11. 17, 171

Salmon, M. and Schleicher, C. (2006). Pricing multivariate currency options with copulas. *Copulas: From Theory to Application in Finance, Risk Books, London*. 22

Samuelson, P. A. (2015). Rational theory of warrant pricing. In *Henry P. McKean Jr. Selecta*, pages 195–232. Springer. 17, 174

Saracco, J. (1997). An asymptotic theory for sliced inverse regression. *Communications in Statistics-Theory and Methods*, 26(9):2141–2171. 68, 72

Schaffnit-Chatterjee, C., Schneider, S., Peter, M., and Mayer, T. (2010). Risk management in agriculture. *Deutsche Bank Reseach. Sept*. 13

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):879–885. 16

Serra, T., Goodwin, B. K., and Featherstone, A. M. (2003). Modeling changes in the US demand for crop insurance during the 1990s. *Agricultural Finance Review*, 63(2):109–125. 12

Shapiro, B. and Brorsen, B. W. (1988). Factors affecting farmers' hedging decisions. *Applied Economic Perspectives and Policy*, 10(2):145–153. 13

Sharkey, P. and Tawn, J. A. (2017). A poisson process reparameterisation for Bayesian inference for extremes. *Extremes*, 20(2):239–263. 60

Skees, J. R., Black, J. R., and Barnett, B. J. (1997). Designing and rating an area yield crop insurance contract. *American Journal of Agricultural Economics*, 79(2):430–438. 2

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231. 3, 16, 23, 172

Smith, R. and Goodman, D. (2000). Bayesian risk analysis. *Extremes and Integrated Risk Management*, pages 235–251. 59

Smith, R. L. (1987). Estimating tails of probability distributions. *The Annals of Statistics*, 15(3):1174–1207. 52, 53

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4(4):367–377. 4, 55, 64

Smith, R. L. (1998). Bayesian and frequentist approaches to parametric predictive inference. In *Bayesian Statistics*, pages 589–612. Oxford University press. 59

Smith, V. H. and Goodwin, B. K. (1996). Crop insurance, moral hazard, and agricultural chemical use. *American Journal of Agricultural Economics*, 78(2):428–438. 79

Stokes, J. R. (2000). A derivative security approach to setting crop revenue coverage insurance premiums. *Journal of Agricultural and Resource Economics*, 25(1):159–176. 17, 174

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society: Series B*, 52(2):237–258. 31, 65

Taghia, J., Ma, Z., and Leijon, A. (2014). Bayesian estimation of the von-Mises Fisher mixture model with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1701–1715. 116

Tejeda, H. A. and Goodwin, B. K. (2008). Modeling crop prices through a Burr distribution and analysis of correlation between crop prices and yields using a copula method. In *Annual Meeting of the American Agricultural Economics Association*. AAEA. 17, 174

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288. 30, 66, 117

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108. 31, 66

Trestini, S., Szathvary, S., Pomarici, E., and Boatto, V. (2018). Assessing the risk profile of dairy farms: application of the Income Stabilisation Tool in Italy. *Agricultural Finance Review*, 78(2):195–208. 14

Ullah, R., Shivakoti, G. P., Zulfiqar, F., and Kamran, M. A. (2016). Farm risks and uncertainties: Sources, impacts and management. *Outlook on Agriculture*, 45(3):199–205. 1, 11, 12, 13, 170

Van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50. 114, 122, 194

van Oordt, M. R., Stork, P. A., and de Vries, C. G. (2021). On agricultural commodities' extreme price risk. *Extremes*, 24:1–33. 4, 17

Velandia, M., Rejesus, R. M., Knight, T. O., and Sherrick, B. J. (2009). Factors affecting farmers' utilization of agricultural risk management tools: the case of crop insurance, forward contracting, and spreading sales. *Journal of Agricultural and Applied Economics*, 41(1):107–123. 13

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. (2020). Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318. 48

Wali, B., Greene, D. L., Khattak, A. J., and Liu, J. (2018). Analyzing within garage fuel economy gaps to support vehicle purchasing decisions–a copula-based modeling & forecasting approach. *Transportation Research Part D: Transport and Environment*, 63:186–208. 22

Walshaw, D. (2000). Modelling extreme wind speeds in regions prone to hurricanes. *Journal of the Royal Statistical Society*, 49(1):51–62. 59

Walters, C. and Preston, R. (2018). Net income risk, crop insurance and hedging. *Agricultural Finance Review*, 78(1). 173, 174

Wang, H. H., Tack, J. B., and Coble, K. H. (2020). Frontier studies in agricultural insurance. *The Geneva Papers on Risk and Insurance – Issues and Practice*, 45:1–4. 12

Wang, H. J. and Li, D. (2013). Estimation of extreme conditional quantiles through power transformation. *Journal of the American Statistical Association*, 108(503):1062–1074. 55

Wang, L. and Yang, L. (2009). Spline estimation of single-index models. *Statistica Sinica*, pages 765–783. 32, 65

Watson, G. S. and Williams, E. J. (1956). On the construction of significance tests on the circle and the sphere. *Biometrika*, 43(3/4):344–352. 110

Weibull, W. (1951). Wide applicability. *Journal of Applied Mechanics*, 103(730):293–297. 41

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815. 53

Whalen, P. (2009). 'insofar as the ruby wine seduces them': cultural strategies for selling wine in inter-war Burgundy. *Contemporary European History*, 18(1):67–98. 19, 177

Whitson, R. E., Barry, P. J., and Lacewell, R. D. (1976). Vertical integration for risk management: An application to a cattle ranch. *Journal of Agricultural and Applied Economics*, 8(2):45–50. 13

Wing, I. S., De Cian, E., and Mistry, M. N. (2021). Global vulnerability of crop yields to climate change. *Journal of Environmental Economics and Management*, 109:102462. 11

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12(S1):117–142. 31, 65, 109

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610. 32, 65, 109

Wu, T. Z., Yu, K., and Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis*, 101(7):1607–1621. 32, 65

Xu, W., Li, D., and Wang, H. (2020). Extreme quantile estimation based on the tail single-index model. *Statistica Sinica*. 5, 6, 58, 62, 63, 66, 67, 68, 74, 76, 109, 192

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656. 174

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054. 32, 65

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67. 31

Zhu, L., Huang, M., and Li, R. (2012). Semiparametric quantile regression with high-dimensional covariates. *Statistica Sinica*, 22(4):1379. 74

Zhu, Y., Ghosh, S., and Goodwin, B. (2008). Modeling dependence in the design of whole farm insurance contract a copula-based approach. In *Annual Meeting of the American Agricultural Economics Association*. AAEA. 3, 16, 173, 174

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429. 31

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: series B*, 67(2):301–320. 30, 66