UNIVERSIDAD DE SEVILLA

imus
Instituto Universitario de Investigación
de Matemáticas de la Universidad de Sevilla
"Antonio de Castro Brzezicki"

Programa de doctorado "Matemáticas"

PhD Dissertation

Enhancing Robustness and Sparsity via Mathematical Optimization

Author
Alba Victoria Olivares Nadal

Supervisors
Prof. Dr. *Emilio Carrizosa Priego*
Prof. Dr. *Pepa Ramírez Cobo*

June 9, 2016

A mis padres y mis hermanas.

Agradecimientos

En primer lugar me gustaría dar las gracias a todas las personas que han pasado por mi vida y que, a través de buenas y malas experiencias, me han enseñado algo. Todas estas experiencias han contribuido a hacerme crecer como persona y, por tanto, a ellas les debo parte de esta tesis.

Me gustaría continuar agradeciendo a todos los profesores y personas que me han inculcado el amor por las Matemáticas y me han motivado a emprender este duro pero gratificante camino.

Un lugar especial ocupan mis directores de tesis, Pepa y Emilio, que han ido acompañándome y construyendo la persona e investigadora que soy ahora. En particular, me gustaría dar las gracias a Emilio por la oportunidad que me ha brindado, por hacerme partícipe de sus ideas, por las horas interminables en su despacho y por su cercanía. A Pepa, por ser una persona ejemplar y porque más que una directora ha sido una amiga, que me ha aportado nuevas perspectivas tanto en el ámbito profesional como el personal.

Otro lugar especial lo reservo para Víctor, que me ha abierto la mente y enseñado otras formas de investigar y de trabajar, siempre desde la cercanía y la conciliación personal.

Quiero agradecer a todas las compañeras del IMUS, que han conseguido crear un ambiente de trabajo estupendo y se han convertido en mis amigas a lo largo de estos años. Mención especial aquí para Vanesa, por escuchar mis penas y mis alegrías con paciencia, prestarme su ordenador, compartir licencias, y ser el alma de la L2. Estoy también agradecida a Marina, por ser tan alegre y atenta, y a Diego y Asun, por ser los mejores compañeros de despacho. Por supuesto tengo que dar las gracias también a Marithania, por compartir conmigo el camino de la carrera y del doctorado; espero que sigamos acompañándonos. También quiero mencionar a las personas que me hicieron sentir acompañada durante mis estancias en Londres: Mei, Boris, Sandra...

A Isabel, por esos estupendos consejos.

Un hueco en esta página se merecen todas mis amigas, porque son un pilar fundamental en mi vida y porque sin su apoyo y paciencia esta tesis no habría sido posible. Clara, Sara, Irene, Carmen, Maripe... da igual cuándo y cómo nos conociéramos, sois todas muy importantes para mí.

Y cómo no voy a mencionar a la persona que ha convivido y compartido conmigo el último y más difícil periodo de mi tesis. Rob, gracias por esa paciencia casi infinita, por tu comprensión, for comforting me with your hugs and jokes, always holding my hand. Thank you.

Por último, pero lo más importante, gracias a mi familia. A toda. A mis padrinos. A mi abuela, por su amor incondicional y por inculcarme la importancia de la independencia. A mi abuelo, por animarme a estudiar y por creer que yo era capaz de todo. A mi tata. A mis hermanas, Julia y Ana, por estar ahí cuando las necesito, consolándome y ayudándome cada una a su manera. A mi padre, porque sería capaz de cualquier cosa por mí, por todas las veces que me has preguntado la lección cuando era pequeña. A mi madre, por inculcarme mi amor por las Matemáticas pero advertirme de la dureza del camino. Porque sé que tu amor es incondicional y que nunca me vas a fallar.

Gracias.

Resumen

Esta tesis se centra en derivar métodos robustos o dispersos bajo la perspectiva de la optimización para problemas que tradicionalmente se engloban en los campos de la Estadística o de la Investigación Operativa. Concretamente, el objetivo de esta tesis doctoral es fusionar técnicas de optimización con conceptos estadísticos para desarrollar metodologías innovadorass que puedan mejorar a los métodos ya existentes y que aúnen las matemáticas teóricas con los problemas de la vida real.

Por una parte, los métodos robustos propuestos facilitarán un nuevo enfoque sobre el modelado y la interpretación de problemas clásicos del área de la Investigación Operativa, produciendo soluciones que sean resistentes a varios tipos de incertidumbre. Por otra parte, las estrategias dispersas desarrolladas para resolver problemas notorios del área de Estadística tendrán forma de Problemas No Lineales Mixtos (es decir, problemas de optimización con algunas variables enteras o binarias y función objetivo no lineal, denotados MINLP a partir de ahora). Se mostrará que los métodos propuestos no solamente son manejables computacionalmente, sino que además realzan la interpretabilidad y obtienen una buena calidad de predicción.

Específicamente, el Capítulo 2 se centra en descubrir causalidades potenciales en series temporales multivariantes. Esto se lleva a cabo formulando el problema como un MINLP donde las restricciones modelan distintos aspectos de la dispersión, incluyendo restricciones que no permiten la aparición de relaciones espúreas en el modelo. El método muestra un buen rendimiento en términos de poder de predicción y de recuperación del modelo original.

Análogamente, el objetivo del Capítulo 3 es descubrir cuáles son los predictores relevantes en un problema de regresión lineal, sin llevar a cabo tests de significación ya que éstos pueden fallar si existe multicolinealidad. Para ello, se formulan MINLPs que restringen los métodos de estimación seleccionados, añadiendo restricciones que miden la importancia de los predictores y que están diseñadas para evitar los problemas que produce la multicolinearidad en los datos. Los modelos restringidos muestran un buen equilibrio entre interpretabilidad y precisión.

Por otra parte, en el Capítulo 4 se generaliza el problema clásico del vendedor de periódicos, asumiendo demandas correladas. En particular, una estrategia de inventario

robusta, donde no se asumen hipótesis distribucionales sobre la demanda, se formula como un problema de optimización. Para el modelado de dicho problema se hace uso de técnicas que ligan conceptos estadísticos con conjuntos de incertidumbre. Las soluciones obtenidas son robustas ante la presencia de ruido con alta variabilidad en los datos, mientras evitan el exceso de conservadurismo.

En el Capítulo 5 se extiende esta formulación para series temporales multivariantes. El escenario es, además, más complejo: no solamente se busca fijar los niveles de producción, sino que se quiere determinar la localización de instalaciones y la asignación de clientes a las mismas. Empíricamente se muestra que, para diseñar una cadena de suministros eficiente, es importante tener en cuenta la correlación y la variabilidad de los datos multivariantes, desarrollando técnicas basadas en los datos que hagan uso de métodos de predicción robustos.

Un examen más exhaustivo de las características específicas del problema y de los conjuntos de incertidumbre se lleva a cabo en el Capítulo 6, donde se estudia el problema de selección de portfolios con costes de transacción. En este capítulo se obtienen resultados teóricos que relacionan los costes de transacción con diferentes maneras de protección ante la incertidumbre de los retornos. Como consecuencia, los resultados numéricos muestran que calibrar la penalización de los costes de transacción produce resultados que son resistentes a los errores de estimación.

Abstract

This thesis is focused on deriving robust or sparse approaches under an optimization perspective for problems that have traditionally fell into the Operations Research or the Statisics fields. In particular, the aim of this Ph.D. dissertation is to merge optimization techniques with statistical concepts, leading to novel methods that may outperform the classic approaches and bridge theoretical mathematics with real life problems.

On one hand, the proposed robust approaches will provide new insights into the modelling and interpretation of classic problems in the Operations Research area, yielding solutions that are resilient to uncertainty of various kinds. On the other hand, the sparse approaches derived to address some up-to-the-minute topics in Statistics will take the form of Mixed Integer Non-Linear Programs (i.e. optimization problems with some integer or binary variables and non linear objective function, denoted as MINLP thereafter). The proposed methods will be shown to be computationally tractable and to enhance interpretability while attaining a good predictive quality.

More specifically, Chapter 2 is focused on discovering potential causalities in multivariate time series. This is undertaken by formulating the estimation problem as a MINLP in which the constraints model different aspects of the sparsity, including constraints that do not allow spurious relationships to appear. The method shows a good performance in terms of forecasting power and recovery of the original model.

Analogously, in Chapter 3 the aim is to discover the relevant predictors in a linear regression context without carrying out significance tests, since they may fail in the presence of strong collinearity. To this aim, the preferred estimation method is tightened, deriving MINLPs in which the constraints measure the significance of the predictors and are designed to avoid collinearity issues. The tightened approaches attain a good trade-off between interpretability and accuracy.

In contrast, in Chapter 4 the classic newsvendor problem is generalized by assuming correlated demands. In particular, a robust inventory approach with distribution-free autoregressive demand is formulated as an optimization problem, using techniques that merge statistical concepts with uncertainty sets. The obtained solutions are robust against the presence of noises with high variability in the data while avoiding overconservativeness.

In Chapter 5 this formulation is extended to multivariate time series in a more complex setting, where decisions over the location-allocation of facilities and their production levels are sought. Empirically, we illustrate that, in order to design an efficient supply chain or to improve an existent one, it is important to take into account the correlation and variability of the multivariate data, developing data-driven techniques which make use of robust forecasting methods.

A closer examination of the specific characteristics of the problem and the uncertainty sets is undertaken in Chapter 6, where the portfolio selection problem with transaction costs is considered. In this chapter, theoretical results that relate transaction costs with different ways of protection against uncertainty of the returns are derived. As a consequence, the numerical experiments show that calibrating the transaction costs term yields to results that are resilient to estimation error.

Contents

1 Introduction **2**

1.1 Sparsity 3

1.2 Robustness 7

1.3 Problems to be treated 10

1.3.1 Sparsity in Statistical problems 10

1.3.2 Robustness in Operations Research problems 13

2 A Sparsity-Controlled Vector Autoregressive Model **22**

2.1 Introduction 23

2.2 Preliminaries 25

2.3 The SC-VAR. 28

2.3.1 Different aspects of the sparsity 28

2.3.2 Mathematical Programming formulation 29

2.3.3 Choice of parameters 31

2.4 Numerical illustrations 32

2.4.1 Comparison methodology 32

2.4.2 Simulation study 34

2.4.3 Real data sets 40

2.4.4 Tuning of the parameters 44

2.5 Concluding remarks and extensions 46

3 Enhancing Sparsity by Tightening Linear Regression Models **50**

3.1 Introduction 51

3.2 New constrains for linear regression approaches 52

3.2.1 Sparsity 52

3.2.2 Significance 53

3.2.3 Collinearity 55

3.3 Numerical illustrations 57

3.3.1 Methods under comparison 58

3.3.2 Design of experiments 61

3.3.3 Real datasets 62
3.3.4 Simulated data 64
3.4 Concluding remarks 66

4 Robust Newsvendor Problem with Autoregressive Demand 68
4.1 Introduction 69
4.2 The model 71
4.2.1 The single-product case 71
4.2.2 The multi-period case 80
4.2.3 The multi-product case 81
4.3 Numerical illustrations 82
4.3.1 Benchmark methods 83
4.3.2 Synthetic data generation and experiments design 83
4.3.3 Results 85
4.4 Concluding remarks and extensions 89

5 On the Effect of the Demand Forecasting Technique on Location-Allocation Decisions 90
5.1 Introduction 91
5.2 Robust location problems with Vector Autoregressive demands 93
5.2.1 Illustration of demands' feasible paths 97
5.3 A data-driven approach for a single-period location-production model . . 100
5.4 Description of the experiments 103
5.4.1 Data sets 103
5.4.2 Stochastic models for the demand 103
5.4.3 Forecasting techniques for the single-period inventory policy . . 105
5.4.4 Profitability structure 105
5.5 Numerical results 105
5.5.1 Results under the maximum profit criterion 106
5.5.2 The effect of the allocation 110
5.6 Concluding remarks and extensions 116

6 A Robust Perspective on Transaction Costs in Portfolio Optimization118
6.1 Introduction 119
6.2 The robust optimization problem 122
6.3 The robust regression problem 125
6.4 The Bayesian portfolio problem 131
6.5 Data-driven portfolios 135
6.5.1 Description of the portfolios 136
6.5.2 Comparison methodology 139

6.5.3 Discussion . . . 139
6.6 Concluding remarks . . . 146

References **156**

Chapter 1

Introduction

High dimensional real data are becoming customary to solve any real application problem in the fields of Statistics and Operations Research (Friedman et al., 2001; Hastie et al., 2015). As a consequence, two important properties are lately sought to derive competitive methodologies for and from Data Analysis: sparsity and robustness. Both properties are quite valuable since they enhance the standard outputs in many aspects. First, a solution is said to be robust against uncertainty on the data if its performance does not deteriorate much when perturbations on the data arise or estimation error exists (Bertsimas et al., 2011). This is quite important since historical data are often affected by measurement errors, and future data are unknown and thus need to be estimated. Second, a solution is said to be sparse if it consists of a large number of zeroes and only a few non-zero coefficients (Fan and Lv, 2010). Sparsity helps visualization, which is a trending topic nowadays that we are in the so called big data era (Hastie et al., 2015). The popularity of the sparsity does not only lie on easing the interpretability of a process, but requiring sparsity is also essential because the use of all the available attributes to predict a variable of interest may yield a poorer predictive quality due to an overfit to the training data. Furthermore it may imply computational as well as financial costs.

The aim of this thesis is to solve some trending problems in Operations Research and Statistics under a robust and a sparse perspective, respectively, by making use of mathematical programming techniques. The combination of optimization tools with statistical concepts will lead to approaches that outperform the current methods, as our numerical experiments show.

1.1 Sparsity

A plethora of real world data, such as e.g. air pollution measures, brain functional connections, gene expressions...etc, involve multiple features mutually interacting. However, the consideration of a large amount of features for prediction may not only imply high computational costs but may also lead to other problematic issues, such as a poor predictive quality or the uninterpretability of the underlying process (Rish and Grabarnik, 2014; Bickel et al., 2009). To illustrate this, consider a multivariate time series model as the one depicted on the left panel of Figure 1.1. Each node corresponds with a feature, and an arrow from node j to node i means that the model uses feature j in the forecast of feature i. Data following this multivariate time series model was simulated. In the right panel of Figure 1.1 the model estimated from this data by a conventional estimation procedure, namely Ordinary Least Squares (OLS thereafter), is depicted. As observed, the estimated graph contains a lot of spurious arrows that complicate the interpretability of the real process and, as it will be shown in Chapters 2 and 3, it may deteriorate the forecasting power.

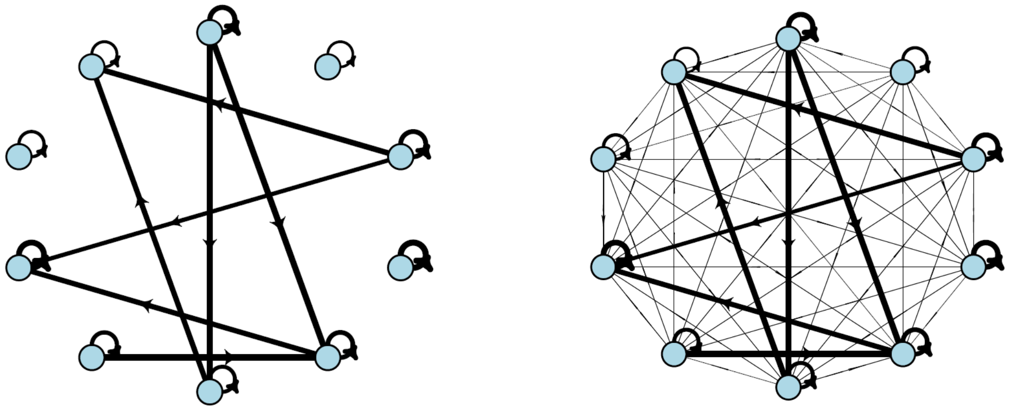


Figure 1.1: A simulated graph (left panel) together with its approximated counterpart (right panel), estimated by OLS.

Since the data collection technologies are improving altogether with communication systems and computers memories and processors, the dimension of the data sets is increasing drastically (Hastie et al., 2015), emphasizing even more the phenomenon shown in Figure 1.1. Since we are not likely to easily understand procedures in which more than a handful of variables interact, nowadays researchers are frequently given high dimensional databases and asked to return an interpretable output which explains what is truly happening behind that pile of data; i.e., they are asked to choose which features are relevant for the prediction (Friedman et al., 2001; Hastie et al., 2015). One way to decide whether a predictor is meaningful or not is by obtaining a solution with a large number of zeroes, in which only the most significant features are associated with the non-zero coefficients. In other words, a sparse solution is sought.

Despite of easing the interpretability of our data, sparsity might not seem a property to be sought if we take into account that the fit to the training data improves when the number of non-zero coefficients of a model increases. However, if the model is fitted *too strictly* to the training data, the prediction errors can be large. This phenomenon is called overfitting, and can happen frequently if all the available variables are used for the prediction. Therefore sparsity becomes a desirable property because it may not only enhance interpretability but it may also avoid the overfitting.

Although sparsity has traditionally been attained as a consequence of variable selection in regression problems, sparse models are in demand in many other fields, such as Support Vector Machines, Principal Components Analysis, Matrix Decompositions..., etc (Hastie et al., 2015; Carrizosa et al., 2016a; Carrizosa and Guerrero, 2014). Since very inaccurate predictions may be obtained if multivariate time series are analyzed separately by repeatedly using one-dimensional time series forecasting tools (Peña and Sánchez, 2007), multivariate models stand as a popular tool amongst staticians to study time series. However, the number of parameters rapidly increases with the dimension of the multivariate time series (Peña and Poncela, 2006). In order to avoid the numerical issues derived from this fact, sparsity has also been studied recently in a time series context. As a consequence, a different type of approaches, called graphical models, have arisen (Arnold et al., 2007). In particular, sparsity has been widely studied in order to unravel potential causalities in biological systems. For instance, graphical models have been developed to deal with genetic networks (Abegaz and Wit, 2013; Shojaie and Michailidis, 2010; Lozano et al., 2009; Hu and Hu, 2009; Dobra et al., 2004), which includes *E. coli* and *Arabidopsis thaliana* regulatory networks, or human cancer cell data. Sparse models have also been developed to enhance visualization of brain functional connectivity or to understand the effect of air pollution and exposure over human health (Dominici et al., 2000; Gorrostieta et al., 2013; Valdés-Sosa et al., 2005).

Traditionally linear regression has caught most of the attention due to its simplicity and applicability. In linear regression a variable is expressed as a linear combination of some attributes. In practice, it has been applied to a wide range of real problems. For example, its applications range from the prediction of the development of illnesses, such as cancer or diabetes, to the estimation of the overall reported crime rate (Hastie et al., 2015). Despite OLS is the conventional procedure to estimate the coefficients of linear regression models, dense solutions are to be obtained (see, for instance, Figure 1.1).

However, a plethora of sparse approaches have been proposed in this context Hesterberg et al. (2008). In fact, the importance of variable selection in the Statistics field led to the early birth of many methods around the 60s and 70s. *Forward stepwise regression* and *all-subsets regression* (Furnival and Wilson, 1974) were found amongst the most popular ones. However, these methods are focused on picking the meaningful predictors, and they do not address the task of estimating their associated coefficients. This is normally done by the user on a second step using a standard criterion such as maximum-likelihood or OLS. As a consequence, the solutions obtained by these approaches may be quite unstable, since small perturbations in the data can change the set of selected features.

An attempt to overcome such instability was undertaken by the *ridge regression* (Draper et al., 1998; Miller, 2002), which consists on adding to the objective function an l_2-norm penalization term over the coefficients to be estimated. In contrast to the previous methods, this approach includes all the variables in the model but performs a shrinkage on their associated coefficients. A more interesting approach is the *Lasso* (Tibshirani, 1996), which implies adding an l_1-norm penalization term instead. Although this approach is also subject to a shrinkage of the coefficients, it attains different levels of sparsity depending on the value of the penalty added to the objective function. Therefore, this method performs simultaneously variable selection and parameter estimation. The output of the *Lasso* consists on a path of solutions with different degrees of sparsity. The final solution to be used is usually chosen ad-hoc using a modeling validation technique such as cross-validation.

The popularity of the Lasso relies mostly on its tractability, which was enhanced by the publication of the Least Angle Regression method (LAR thereafter) by Efron et al. (2004). This algorithm progressively adds to the model the variable most correlated with the current residuals. Although the LAR may seem quite different to the *Lasso*, this algorithm can be modified to efficiently calculate the Lasso path of solutions and has led to a large number of handy functions implemented in standard mathematical packages such as Sjöstrand et al. (2012); Hastie and Efron (2013). In order to illustrate the performance of this approach, Figure 1.2 shows the solution obtained by the LAR for the graph depicted at the beginning of this section. Only one arrow is allowed in the model per each node. As it can be observed, the obtained graph is more interpretable than that obtained by the OLS, although thinner arrows than the original graph show the effect of the shrinkage over the coefficients. Such a shrinkage is caused by the l_1-norm penalization term, and it will be shown in Chapters 2 and 3 that this may also deteriorate the predictive power. Furthermore, the estimation of the coefficients using the OLS and the Lasso may be highly unstable when there exists strong collinearity on the data (Hesterberg et al., 2008). In fact, although the Lasso is a robust approach rather than a sparse model (Bertsimas and Mazumder, 2014; Caramanis et al., 2012) and thus reduces the variance of the estimates, it is still subject to strong instability and produces biased estimates (Hesterberg et al., 2008).

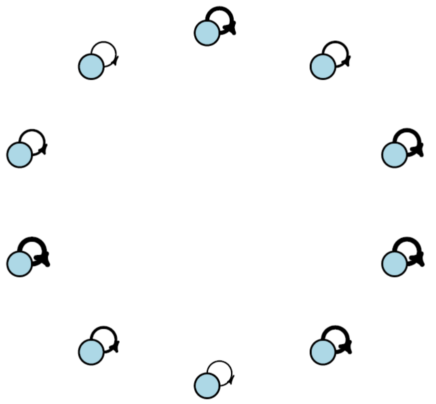


Figure 1.2: Graph estimated using the LAR methodology.

As the aim of this Ph.D. thesis is to merge statistical concepts with optimization techniques, Chapters 2 and 3 will be devoted to formulate sparse approaches in terms of MINLPs. In particular, sparse methods will be modelled in a multivariate time series and a linear regression context with the aim of providing alternatives to the

Lasso which could help avoiding its drawbacks: deterioration of the predictive quality due to strong shrinkage over the coefficients, and misbehaviour in the presence of high correlations. The tractability of the proposed approaches will be supported by the recent improvements of the computer processors and the MINLP solvers (Bertsimas and King, 2015), making it feasible to address problems of large size, as our numerical experiments report.

1.2 Robustness

In a classic optimization problem the parameters are assumed to be known, yielding a solution that will remain optimal along the time. However, in practice parameters are often not deterministic; for example, future demand in supply chain management is unknown. Since in the presence of uncertain parameters the original optimal solution may become suboptimal or unfeasible, many branches of study have arisen to understand and/or alleviate the impact of parameter uncertainty in mathematical programming.

First, *sensitivity analysis* focuses on analyzing how the objective value is affected by small changes in the parameters Saltelli et al. (2000). This approach is interesting to study the a posteriori consequences of clinging to the original solution, but does not grant protection against uncertainty when solving the optimization problem.

On the contrary, *stochastic optimization* expresses the parameter uncertainty using probabilistic elements Heyman and Sobel (2003). This generalizes the classical perspective on optimization, yielding new types of constraints and objective functions. Nevertheless, the stochastic modelling of the uncertainty may be quite restrictive as well as computationally costly. On top of that, stochastic optimization may rely on distributional assumptions, which may be unrealistic in practice.

Robust optimization is an alternative methodology which bounds the uncertainty of the parameters by requiring them to belong to a set, which is expressed in a deterministic way Ben-Tal and Nemirovski (1998); Ben-Tal et al. (2009). That is to say, parameters are not known but they are assumed to vary within the so-called uncertainty set. To illustrate this, consider Figure 1.3, where an uncertainty set has been depicted for an ambiguous parameter $\mathbf{u}$.

In order to obtain a solution that performs well under any possible realization, a robustness measure must be optimized in the objective function. Specifically, a classic robust optimization problem can be formulated as follows:

$$\begin{array}{lll} \min \max\limits_{\mathbf{u}_0 \in \mathcal{U}_0} & g_0(\mathbf{x}, \mathbf{u}_0) & \\ \text{s.t} & g_j(\mathbf{x}, \mathbf{u}_j) \leq 0 \quad \forall \mathbf{u}_j \in \mathcal{U}_j, & j = 1, .., m \end{array} \tag{1.1}$$

where $\mathbf{x} \in \mathbb{R}^r$ and $\mathbf{u}_j \in \mathbb{R}^k$ are vectors of variables and uncertain parameters, respectively, and $g_0, g_j : \mathbb{R}^r \to \mathbb{R}$ are functions. The uncertainty sets $\mathcal{U}_j \subseteq \mathbb{R}^k$ determine the

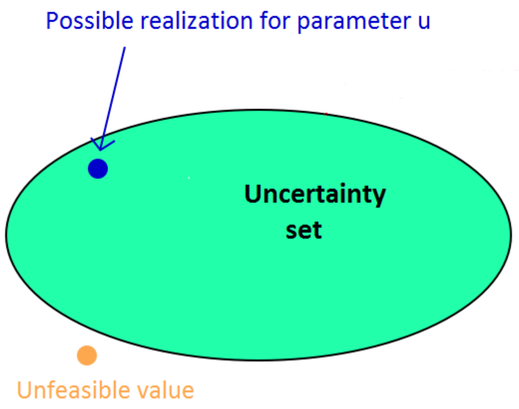


Figure 1.3: Example of an uncertainty set for a non-deterministic parameter $\mathbf{u}$.

possible realizations for the parameters $\mathbf{u}_j$, while $\max\limits_{\mathbf{u}_0 \in \mathcal{U}_0} g_0(\mathbf{x}, \mathbf{u}_0)$ expresses the robust measure included in the problem so as to alleviate the impact of the uncertainty over the objective value.

The computational costs of solving Problem (1.1) may sometimes be considerably higher than those of its deterministic version. However, a deeper study of the properties of a problem can allow to determine uncertainty sets that endow their robust version with good performance and tractability. For a deeper review in recent advances in these topics, the reader is refereed to Bertsimas et al. (2011); Gabrel et al. (2014). Specifically, Bertsimas et al. (2011) provides guidelines on how to construct uncertainty sets with good properties for some type of general problems. Furthermore, many stochastic problems possess properties that allow for a robust expression that performs equivalently while implying less computational costs. These advances have upgraded the robust optimization into a realistic and handy tool to address parameter uncertainty in mathematical programming.

The most popular robust measure for robust optimization problems is that of minimizing under the worst-case scenario. A common drawback in this case is that Problem (1.1) may yield solutions that are too conservative, since uncertainty sets may not be tight enough or decisions may be taken focusing solely on extreme cases that are unlikely to happen (Ben-Tal and Nemirovski, 2002). Two alternatives have been explored to overcome this over conservativeness: either to consider other robust measures in the objective function, or to modify the size and/or the shape of the uncertainty sets. Although other robust preferences (such as minimizing the maximum regret Perakis and Roels (2008)) have been treated in the literature, the most popular approach is to modify somehow the uncertainty sets.

Specifically, the over conservativeness derived by focusing on optimizing under the worst-case scenario can be reduced by excluding outlying cases. This can be attained by simply reducing the uncertainty set. Nevertheless, this must be done carefully since the user would become unprotected against any possible realization that is left out of the uncertainty set. The user, taking into account her own risk aversion, is the one to decide the proper trade off between performance and robustness. However, sometimes the nature of the problem provides an idea about how to determine an appropriate budget of uncertainty. As an example, consider the inventory approach proposed by Bertsimas and Thiele (2006), where the future demands for periods $t = T+1, ..., T+h$, X_t, are unknown, but nominal values for them, $\hat{X}_t$, are given. Then, interval-shaped uncertainty sets of length $2L_t$ are defined as follows:

$$\mathcal{U}_t = \left[\hat{X}_t - L_t, \hat{X}_t + L_t\right] \quad t = T+1, ..., T+h. \tag{1.2}$$

The scaled deviations for the demands, which take values in $[-1, 1]$, are defined as $z_t = (X_t - \hat{X}_t)/L_t$. The size of the uncertainty set $\mathcal{U}_{T+1} \times ... \times ...\mathcal{U}_{T+h}$ can be reduced by adding the constraint $\sum_{t=T+1}^{T+h} |z_t| \leq \Gamma$, where a $\Gamma = 0$ would imply there exists no uncertainty, and $\Gamma = h$ would allow the demands to always take the worst-case value. The implications of considering these uncertainty sets in a facility location problem will be discussed more carefully in Chapter 5.

Nevertheless, occasionally any reduction of the size of the uncertainty set may exclude important scenarios that should have been taken into account. That is the reason because it is also important to modify the shape of the uncertainty set as well, so as to construct a set that is tight over the cases to be considered. In this task, the nature and structure of the problem must be thoroughly taken into account, since, as mentioned before, some uncertainty sets are known to provide good performance when coupled with a specific type of problem Bertsimas et al. (2011). In addition, novel techniques that make use of statistical results to construct uncertainty sets is bringing a whole new perspective into the robust optimization field Bandi and Bertsimas (2012). For example, assume that $Z_1, ..., Z_s$ are independent and identically distributed random variables. Then, using the Central Limit Theorem, Bandi and Bertsimas (2012) defines an uncertainty set as:

$$\mathcal{U}\left(\Gamma_1^*\right) = \left\{(Z_1, \ldots, Z_s) \quad \text{s.t} \quad \left|\frac{1}{s}\sum_{i=1}^{s} Z_i\right| \leq \frac{\Gamma_1^* \sigma_Z \sqrt{s}}{s} + \mu_Z\right\}, \tag{1.3}$$

where μ_Z and σ_Z stand for the mean and the standard deviation of the random variables $Z_1, ..., Z_s$, respectively. The value Γ_1^* in (1.3) is a small constant that influences the accuracy of the fit, usually chosen as the $(1 - \tau/2)$th quantile of the standard normal distribution. Although this kind of approaches provides useful techniques to model the uncertainty sets and to determine a meaningful budget of uncertainty, they are not too spread in the statistical community yet. Nevertheless, recent improvements in computational settings are making these methods become a tractable tool that should be strongly considered.

In this Ph.D. dissertation, bridging Statistics and Optimization, the uncertainty set (1.3) will be used in Chapters 4 and 5 to model the random part of the demand in inventory and location problems. In contrast, a portfolio selection model is considered in Chapter 6, where the prior distributional assumptions of an investor will be related with her modelling of the uncertainty on the mean assets returns.

1.3 Problems to be treated

In this section the family of problems addressed in this PhD dissertation are briefly outlined.

1.3.1 Sparsity in Statistical problems

Linear Regression

One of the most common problems in real life is trying to predict a variable by making use of attributes that are deterministic or are easier to obtain. This idea can be gathered by expressing the stochastic variable $\mathbf{y}$ as a function of the attributes $\mathbf{x}_1, ..., \mathbf{x}_N$

$$\mathbf{y} = f(\mathbf{x}_1, ..., \mathbf{x}_N). \tag{1.4}$$

Since this relationship is hardly ever strictly verified when fitted to real observations, a random error term a is added. This shock accounts for the part of $\mathbf{y}$ that remains unexplained. The simplest model is to assume that f is a linear function, and thus aim to adjust a hyperplane to the observations in the following way:

$$\mathbf{Y} = \beta_0 + \boldsymbol{\beta}\mathbf{X} + \mathbf{a} \tag{1.5}$$

where $\mathbf{Y}$ is a vector containing the K realizations $Y_1, ..., Y_K$ of the stochastic variable $\mathbf{y}$ to be predicted, $\mathbf{X}$ is a matrix containing the observations $X_1^j, ..., X_K^j$ of the attributes $\mathbf{x}_j$, $j = 1, ..., N$, that influence on $\mathbf{y}$, and $\mathbf{a} = (a_1, ..., a_K)'$ is the vector of error terms. Once data containing the realizations of both the dependent variable and the predictors is provided, it remains to estimate the coefficients $\beta_0, \beta_1, ..., \beta_N$.

A natural estimation procedure may arise from minimizing somehow the unexplained part of $\mathbf{y}$, i.e., the error terms $a_1, ..., a_K$. For instance, the OLS method aims to minimize the sum of the squared errors and yields valid estimates for coefficients $\beta_0, \beta_1, ..., \beta_N$.

However, estimators obtained through this procedure may lack some desirable properties. First, the linear coefficients may be strongly perturbed by the presence of outliers, since the hyperplane may deviate from the main cloud of observations to better fit an extreme value whose error term may become aberrant when squared. Second, the estimation of such coefficients may be highly unstable when there exists strong collinearity on the data (Hesterberg et al., 2008). This is usually overcome either by manually removing variables that are highly correlated or by means of harvesting more observations to discard collinearity (Dormann et al., 2013; Montgomery et al., 2015). Finally, there are no reasons to expect the solution of the OLS to be sparse; i.e., a high number of non-zero coefficients is likely to be obtained, making the results difficult to interpret.

To address outliers and collinearity issues, the standard algorithms for linear regression estimation include a step in which unimportant variables are manually removed (Chatterjee and Hadi, 2015; Montgomery et al., 2015). To determine if a variable can be discarded, significance tests may be carried out. However, the results of these tests may not be faithful to the true importance of a variable in the presence of strong collinearity (Watson and Teelucksingh, 2002; Dormann et al., 2013).

Since this procedure can be tedious as well as misleading, in Chapter 3 we will propose tools based on Mathematical Optimization which improve sparsity in linear regression, avoiding these manual steps. This will be addressed by modelling so-called sparsity, significance and collinearity constraints and integrating them into the estimation procedure. The so-obtained tightened sparse models will become MINLPs, solvable by existing solvers.

Vector Autoregressive processes

Because of their simplicity and versatility, linear regression models as in (1.5) has been widely used to study the relationship between different features. Nevertheless, they do not allow to take into account the temporal dependence that may affect the considered variables, thus different tools are needed to model time series in different contexts where the temporal dependence is significant. A popular model in these cases is the autoregressive process. A time series $\{X_t\}_{t>0}$ follows an autoregressive process of order $n \geq 1$ (noted $AR(n)$) if it can be expressed in the form

$$X_t = \alpha + \sum_{k=1}^{n} \theta_k X_{t-k} + a_t \qquad t > 0, \tag{1.6}$$

where $\alpha, \theta_1, ..., \theta_n$ are coefficients and $\{a_t\}_{t>0}$ is the sequence of i.i.d model's error terms with expected value μ_a and variance σ_a^2, for all $t > 0$. For a more detailed description of the autoregressive process and interpretation of the coefficients in the model, see for example Box et al. (2008).

If the parameters in (1.6) are given (either they are known or estimated from sample data, available up to time T), then (1.6) can be used to forecast the process. In particular, if the errors are assumed to follow a normal distribution, then an $(1-\tau)\%$ prediction interval for a forecasted value at time $T+1$ is given by

$$\hat{X}_{T+1} \pm z_{1-\tau/2}\sigma_a \tag{1.7}$$

where $\hat{X}_{T+1}$ is the estimated forecast, and $z_{1-\tau/2}$ is the $(1-\tau/2)$th quantile of the standard normal distribution.

However, a plethora of real world data, such as e.g. brain functional connections or viruses activity, involve multivariate time series, i.e., different and inter-related features, evolving in time, are simultaneously measured and are to be forecasted. Since the components of such multivariate time series are not independent, inaccurate predictions are expected if the series are analyzed separately by repeatedly using one-dimensional time series forecasting tools. In order to properly address such dependencies, vector autoregressive models (VAR) are frequently applied (Arnold et al., 2007; Valdés-Sosa et al., 2005).

Let $\{\mathbf{X}_t\}_{t>0}$ be an N-dimensional vector autoregressive process of order n, VAR(n), i.e., each series i, $i = 1, ..., N$, can be expressed as

$$X_t^i = \alpha^i + \sum_{j=1}^{N}\sum_{k=1}^{n} \theta_{jk}^i X_{j,t-k} + a_t^i \qquad t > 0 \tag{1.8}$$

where $\{a_t^i\}_{t>0}$ denotes the series of contemporaneous shocks that affect feature i, and α^i and θ_{jk}^i are real numbers. A Vector Autoregressive process as in (1.8) can be written in a more compact way as follows:

$$\mathbf{X}_t = \boldsymbol{\alpha} + \sum_{k=1}^{n} A_k \mathbf{X}_{t-k} + \mathbf{a}_t \tag{1.9}$$

where $A_1, ..., A_n \in \mathbb{R}^{N\times N}$ are the parameters of the VAR that relates features k periods behind (i.e., $A_k = (\theta_{jk}^i)_{ji}$), and $\boldsymbol{\alpha} \in \mathbb{R}^N$ is the vector of intercepts.

VAR models have been widely used to study causalities in real processes (Eichler, 2012; Shojaie and Michailidis, 2010; Lozano et al., 2009; Arnold et al., 2007; Valdés-Sosa et al., 2005). The reason VAR models are an appropriate tool to learn about causalities is given by the definition of *Granger causality* (Granger, 1969), introduced by the Nobel prize laureate, Clive Granger. This notion of causality is strongly related to autoregressive processes, as a feature Y is said to *Granger-cause* a feature X if the autoregressive model considering past values not only of X but also of Y is statistically significantly more accurate than considering such a model based solely on X.

In Chapter 2 we will study a sparse variant of the VAR, in which the exact level of sparsity of the output can be controlled. This problem will be formulated in terms of a MINLP, whose objective function will express the desire to attain a good prediction power and the constraints will model different aspects of the sparsity, including constraints that do not allow spurious variables to be associated to non-zero coefficients.

1.3.2 Robustness in Operations Research problems

The newsvendor problem

The single-period problem (SPP), also known as the newsvendor problem, is a simple yet rich inventory model which has been widely studied in the Operations Research field due to its versatility and applicability to many business decision problems, in fields such as managing booking and capacity in airlines companies Weatherford and Pfeifer (1994), health insurances Rosenfield (1986); Eeckhoudt et al. (1991), scheduling Baker and Scudder (1990), retailers and managers order quantity decision in sports and fashion industries Gallego and Moon (1993), etc.

The basic version of the problem consists in making a one-step decision on the quantity Q to be bought of one single perishable product under the assumption that the demand is a random variable with known distribution F. If the decision maker buys each unit at cost c and sells it at price v, then the expected revenue is known to be maximized by buying exactly this quantity of product:

$$Q^* = F^{-1}\left(1 - \frac{c}{v}\right). \tag{1.10}$$

Numerous variants of the classical SPP have been proposed in the literature (see, for instance, Khouja (1999); Petruzzi and Dada (1999); Qin et al. (2011)). Of particular interest are the problems that assume uncertainty over the distribution function of the demand since the traditional assumption that the demand probability distribution is known may be unrealistic in many cases. In addition, if the demand is inferred from sample data, then the resulting estimate may lack of desirable statistical properties (consistency, asymptotic normality...), for example, for small sample sizes. To overcome these and other related problems, some distribution-free approaches have been considered in the literature. The most popular of these approaches may be the so called Scarf's rule Scarf (1958) , which is the solution to the robust optimization problem:

$$\min_{Q\geq 0} \max_{f\in H(\mu,\sigma)} G_f(Q)$$

where $H(\mu,\sigma)$ denotes the set of distributions with mean μ and variance σ, and $G_f(Q)$

is the expected total cost. The solution to this problem is given by:

$$Q_S^\star = \begin{cases} 0 & \text{if } \frac{c}{v}\left(1+\frac{\sigma^2}{\mu^2}\right) > 1 \\ \mu + \frac{\sigma}{2}\left(\frac{1-2\frac{c}{v}}{\sqrt{\frac{c}{v}\left(1-\frac{c}{v}\right)}}\right) & \text{if } \frac{c}{v}\left(1+\frac{\sigma^2}{\mu^2}\right) < 1; \end{cases} \tag{1.11}$$

see Scarf (1958) for more details. This solution is robust because it assumes uncertainty over the distribution function of the demand. Two more remarkable distribution-free works are also Gallego and Moon (1993), which provides an extension to Scarf's solution, and Yue et al. (2006), in which the demand density function is assumed to belong to a specific family of density functions. Other articles which cope with demand uncertainty are Ding et al. (2002); Dana and Petruzzi (2001); Godfrey and Powell (2001).

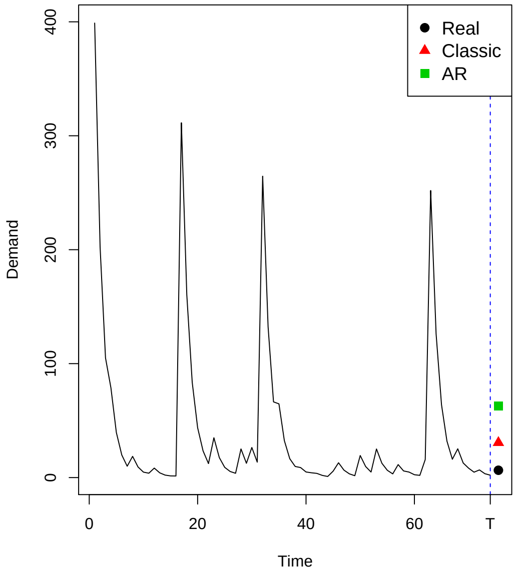


Figure 1.4: Demand time series generated from an autoregressive process with heavy-tailed errors, and real and predicted revenues under two classic approaches (classic newsvendor and AR).

However, as pointed out in See and Sim (2010); Bandi and Bertsimas (2012), not only the assumption of known distribution of the demand may be too strong, but also to estimate the mean and variance from the sample data and accommodate such estimates to an assumed distribution function may generate drastic errors in the inventory policy. Moreover, demand is in fact usually correlated along time, so assuming demands for each period are independent and identically distributed is in practice unrealistic (Lee et al., 2000; Graves, 1999; Kahn, 1987). In contrast, in Chapter 4 we will explore the newsvendor problem in which the distribution of the demand is unknown but it is assumed to follow an autoregressive process as in (1.6).

Different inventory policies yield rather different revenues. For instance, consider Figure 1.4, which depicts a time series for the demand, assumed to follow an autoregressive process with lognormal errors. This represents a realistic realization of the demand since, as mentioned by (Bimpikis and Markakis, 2015), a heavy-tailed distributed demand is common in practice. This is the reason why normally distributed errors may be too restrictive, as they are not able to capture the extreme behavior of the demand. Furthermore, when the errors are not normally distributed, the use of (1.7) may lead to inaccurate results, as will be illustrated in detail in Chapter 4.

From Figure 1.4 it can be observed that conventional methods such as the classic newsvendor and autoregressive approaches can lead to losses. In particular, for the data of Figure 1.4, the classic model yields a negative revenue of -1.21, while when assuming the basic AR model revenues decrease to -9.21. In contrast, robust approaches are usually too conservative and avoid ordering any quantity of product (Bertsimas and Thiele, 2006; Lin and Ng, 2011).

The approach proposed in Chapter 4 successfully copes with heavy-tailed distributed demands (as the lognormal) and usually outperforms both classic and autoregressive approaches in terms of average revenue and also probability of losses, while avoiding over conservativeness of previous robust approaches in the literature. Our method will be robust, since uncertainty over the demand is assumed, and distribution free, as no distributional assumptions will be made over the error terms in the AR.

The p-median location problem

The classic p-median problem (Mirchandani and Francis, 1990) consitsts on opening p facilities amongst a set of candidate locations to attend some customers. The problem of locating the plants must be done in such a way that the company maximizes its benefit. Therefore, the transportation and production costs must be taken into account when formulating the optimization problem.

For the sake of comprehension, we next present the notation used throughout this thesis.

p	number of facilities to open
$\{1,...,L\}$	set of candidate locations for the facilities
L_0	set of opened facilities
$\{1,...,N\}$	set of clients
$C^S=(c^S_{li})$	shipping costs matrix: c^S_{li} represents the cost of sending one unit of product from facility l to client i
$C^P=(c^P_l)$	vector of production costs: c^P_l represents the cost of producing one unit of product at plant l
$C=(c_{li})$	total cost matrix: $c_{li}=c^S_{li}+c^P$

The classic location-allocation p-median problem is formulated as:

$$\begin{aligned}
\min \quad & \sum_{l=1}^{L}\sum_{i=1}^{N} X^i c_{li} x_{li} \\
\text{s.t} \quad & \begin{cases}
\displaystyle\sum_{l=1}^{L} x_{li} = 1 & \forall i = 1,...,N \\
x_{li} \le y_l & \forall i = 1,...,N, l = 1,...,L \\
\displaystyle\sum_{l=1}^{L} y_l = p & \\
x_{li} \in \{0,1\} & \forall i = 1,...,N, l = 1,...,L \\
y_l \in \{0,1\} & \forall l = 1,...,L,
\end{cases}
\end{aligned} \tag{p-median}$$

where X^i is the demand of client i, c_{li} is the sum of the transportation and production costs (c^S_{li} and c^P_l), and the variables represent:

$$y_l = \begin{cases} 1 & \text{if facility } l \text{ is opened} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{li} = \begin{cases} 1 & \text{if client } i \text{ is supplied by facility } l \\ 0 & \text{otherwise.} \end{cases}$$

The Uncapacitated Facility Location Problem (UFLP) (Cornuéjols et al., 1983) is formulated similarly, but the number of facilities to open is not specified. On the other hand, an opening cost o_l is assumed for each facility l:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & \sum_{l=1}^{L}\sum_{i=1}^{N} X^i c_{li} x_{li} + \sum_{l\in L} o_l y_l \\ \text{s.t} \quad & \begin{cases} \sum_{l=1}^{L} x_{li} = 1 & \forall i = 1,...,N \\ x_{li} \leq y_l & \forall i = 1,...,N, l = 1,...,L \\ x_{li} \in \{0,1\} & \forall i = 1,...,N, l = 1,...,L \\ y_l \in \{0,1\} & \forall l = 1,...,L \end{cases} \end{aligned} \tag{UFLP}$$

In practice, the demand of client i to be satisfied, X^i, is unknown and thus needs to be estimated, which has yielded a wide range of robust location problems. Most of them are based either on distributional assumptions over the clients' demands or on scenario analysis. In contrast, in recent research nominal values for the future demands are assumed to be given and uncertainty structures are built taking into account those values (see, for instance, Baron et al. (2011); Bertsimas and Thiele (2006)). The drawback of this methodology is that, although an expert must have already harvested information about the behaviour of the demand in order to select a proper prediction method, unfortunately the user can incorporate such knowledge about the demand directly into the optimization problem only by making use of the estimated nominal values. For example, consider the approach in Bertsimas and Thiele (2006), whose uncertainty sets are modelled as in (1.2). If the demands are known to follow a VAR process, the only information that can be given about this is through the estimates $\hat{X}_t^i$. In constrast, in the theoretical work developed in Chapter 5 we will use the VAR (1.8) to model uncertainty sets for the future demands.

In Chapter 5 we also develop a numerical study that helps us assessing how different demand forecasting techniques affect the location-allocation decisions. To this aim we focus on a more simple problem: the p-median problem with independent autoregressive demands amongst clients.

Portfolio selection and transaction costs

A portfolio selection problem consists of determining how to distribute a unit of wealth among different assets. This decision will obviously depend on the investor's preferences, expressed as utility function, which is usually required to be concave so as to express risk aversion.

The modern portfolio theory was developed by Markowitz (1952), who summarized the behaviour of investors in two simple assumptions. First, it was assumed that an investor always chooses a portfolio with a higher rate of portfolio return. Second, a risk-averse investor will prefer a portfolio with a smaller standard deviation. This led Markowitz (1952) to express the investor preferences as a trade-off between risk and expected return as follows:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w}^T\Sigma\mathbf{w} \\ \text{s.t.} \quad & \mu^T\mathbf{w} = \mu_0 \\ & \mathbf{w}^T\mathbf{1}_N = 1, \end{aligned} \tag{1.12}$$

where $\mathbf{w} \in \mathbb{R}^N$ is the portfolio weight vector, $\Sigma \in \mathbb{R}^{N\times N}$ is the covariance matrix of asset returns, $\mu \in \mathbb{R}^N$ is the expected mean of asset returns, $\mu_0 \in \mathbb{R}$ is the expected rate of return of the portfolio, $\mathbf{1}_N \in \mathbb{R}^N$ is the vector of ones, and the constraint $\mathbf{w}^T\mathbf{1}_N = 1$ requires that the portfolio weights sum up to one.

The solution to problem (1.12) is called frontier portfolio, i.e., it is the one that possesses the minimum variance amongst all the portfolios with the same expected return. The set of all frontier portfolios, called the portfolio frontier, is an hyperbola in the standard deviation-expected return space. A portfolio frontier has been depicted in Figure 1.5, where *mvp* denotes the minimum variance portfolio; i.e., the portfolio that attains the minimum variance amongst all possible portfolios. Portfolios on the frontier with a higher expected return than the *mvp* are called efficient portfolios, and they have a correspondent inefficient portfolio in the frontier with the same variance but lower expected return.

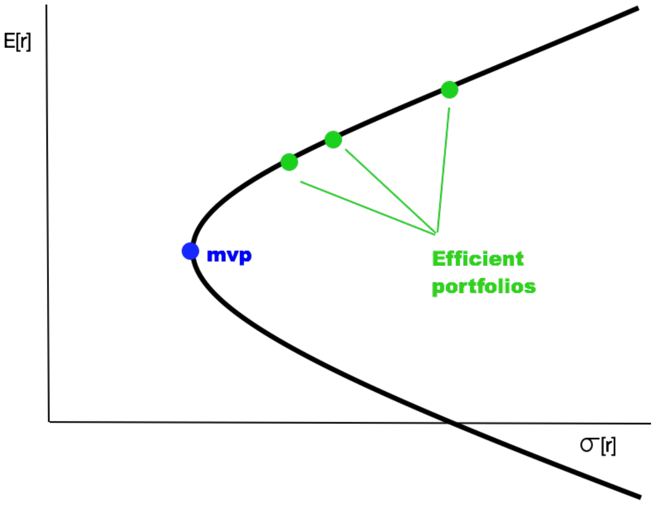


Figure 1.5: Portfolio frontier

Although the frontier portfolios are endowed with nice theoretical properties, in practice they may be subject to strong estimation error since Σ and μ are unknown and thus need to be estimated. Specifically, the estimation of the mean of the assets returns leads to such unstable portfolios that the solutions of the mean-variance problem (1.12) are frequently outperformed by the *mvp* (Michaud, 1989; Chopra and Ziemba, 1993; Broadie, 1993). To alleviate the impact of estimation error, many robust portfolios have been proposed in the literature (Goldfarb and Iyengar, 2003; Garlappi et al., 2007; Lu, 2011a,b).

Another crucial aspect in the successful implementation of the mean-variance portfolios are the transaction costs. Transaction costs are important because they can easily erode the gains from a trading strategy. For instance, DeMiguel et al. (2014) shows that the gains from a trading strategy that exploits serial dependence in stock returns does not survive even modest proportional transaction costs of ten basis points. Transaction costs can be generally modelled with the p-norm of the portfolio trade vector.

Specifically, in Chapter 6 we will consider the following mean-variance problem with p-norm transaction costs:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{\gamma}{2}\mathbf{w}^T\Sigma\mathbf{w} - \mu^T\mathbf{w} + \kappa\|\Lambda(\mathbf{w}-\mathbf{w}_0)\|_p^p \\ \text{s.t.} \quad & \mathbf{w}^T\mathbf{1}_N = 1, \end{aligned} \tag{1.13}$$

where $\gamma \in \mathbb{R}$ is the risk-aversion parameter, $\mathbf{w} \in \mathbb{R}^N$ is the portfolio weight vector, $\Sigma \in \mathbb{R}^{N\times N}$ is the estimated covariance matrix of asset returns, $\mu \in \mathbb{R}^N$ is the estimated mean of asset returns, $\kappa \in \mathbb{R}$ is the transaction cost parameter, $\Lambda \in \mathbb{R}^{N\times N}$ is the transaction cost matrix, which we assume to be symmetric and positive definite, $\mathbf{w_0} \in \mathbb{R}^N$ is the starting portfolio, and $\|s\|_p$ is the p-norm of vector s, $\|s\|_p^p = \sum_{i=1}^N |s_i|^p$. The first two terms in the objective function capture the risk-return trade-off: the first term is the portfolio return variance scaled by the risk-aversion parameter ($\frac{\gamma}{2}\mathbf{w}^T\Sigma\mathbf{w}$), and the second term is the portfolio return mean ($\mu^T\mathbf{w}$). More importantly for our purposes, the third term in the objective function is the p-norm transaction cost term, $\kappa\|\Lambda(\mathbf{w}-\mathbf{w}_0)\|_p^p$. Note that we allow for the portfolio trade vector $(\mathbf{w}-\mathbf{w}_0)$ to be transformed via a symmetric positive definite transaction cost matrix Λ before computing the p-norm. This formulation provides a good approximation for several different types of transaction costs. For small trades, which do not impact the market price, the transaction cost is generally assumed to be proportional to the amount traded on each asset, and thus it can be approximated by the p-norm transaction cost term with $p = 1$ and $\Lambda = I$, where $I \in \mathbb{R}^{N\times N}$ is the identity matrix. For larger trades, the literature has traditionally assumed that the transaction cost is quadratic, and thus it can be captured with $p = 2$. Moreover, Gârleanu and Pedersen (2013) argue that, for the case with large trades, a realistic choice of the transaction cost matrix is $\Lambda = \Sigma^{1/2}$. Finally, several authors have recently argued that the market impact costs associated with large trades grow as the square root of the amount traded (Torre, 1997; Farmer et al., 2013; Almgren et al., 2005; Frazzini et al., 2015), and thus they are better captured by the p-norm with $p = 1.5$. Our transaction cost term is general enough to capture all of these cases.

In Chapter 6 we will show that estimation error and transaction costs are intimately related, and use this relation to propose a data-driven approach to compute portfolios that are both efficient with respect to transaction costs and resilient with respect to estimation error.

Chapter 2

A Sparsity-Controlled Vector Autoregressive Model

Vector autoregressive (VAR) models constitute a powerful and well studied tool to analyze multivariate time series. Since sparseness, crucial to identify and visualize joint dependencies and relevant causalities, is not expected to happen in the standard VAR model, several sparse variants have been introduced in the literature. However, in some cases it might be of interest to control some dimensions of the sparsity, as e.g. the number of causal features allowed in the prediction. To authors extent none of the existent methods endows the user with full control over the different aspects of the sparsity of the solution. In this chapter we propose a sparsity-controlled VAR model which allows to control different dimensions of the sparsity, enabling a proper visualization of potential causalities and dependencies. The model coefficients are found as the solution to a mathematical optimization problem, solvable by standard numerical optimization routines. The tests performed on both simulated and real-life multivariate time series show that our approach may outperform a greedy algorithm and different Lasso approaches in terms of prediction errors specially when highly sparse graphs are sought, while avoiding the VAR's overfitting for more dense graphs.

2.1 Introduction

As mentioned in Chapter 1, a plethora of real world data involve multivariate time series, i.e., different and inter-related features, evolving in time, are simultaneously measured and are to be forecasted. Since the components of such multivariate time series are not independent, inaccurate predictions may be obtained if the series are analyzed separately by repeatedly using one-dimensional time series forecasting tools. In order to properly address such dependencies, vector autoregressive models (VAR) are frequently applied. However, although capturing features dependencies, there is no reason to expect the so-obtained VAR to be sparse. In other words, the output may be too complex when, on top of obtaining sharp forecasts, visualization of relevant causalities is sought, (Eichler, 2012; Shojaie and Michailidis, 2010; Lozano et al., 2009; Arnold et al., 2007; Valdés-Sosa et al., 2005). .

Due to its wide range of applications, several attempts have been proposed in the literature to obtain VAR models in which sparsity, as a potential for easy visualization of complex causal relations, is pursued. This is the case, for instance, of Stochastic Search Variable Selection (George, 2000; George and McCulloch, 1997), Bayesian approaches (Doan et al., 1984) and the Lasso (Tibshirani, 1996). Minimizing the forecasting errors plus an ℓ_1-penalty regularization term has not led to a unique Lasso method, but it has evolved into a full class of Lasso approaches, such as the so-called Adaptative Lasso, (Zou, 2006), the Group Lasso, (Song and Bickel, 2011; Haufe et al., 2010; Zhao et al., 2009; Yuan and Lin, 2006), the Maximum Likelihood Estimated Lasso (Davis et al., 2012; Hsu et al., 2008), or Lasso Granger methods, (Arnold et al., 2007). Such VAR models attempt to gain overall sparsity, without an explicit analysis of its different aspects.

Nevertheless, it might be of interest in certain real life situations to restrict not only the overall number of depicted dependencies but also other levels of sparsity. For example, the number of causal features might be wanted to be limited when acceding to the historical records of the features incurs into a cost. Consider a patient who is under surveillance and several tests need to be undertaken to control her health periodically. Those tests may not only be costly but also invasive for the patient end therefore it may be desirable to reduce their application without affecting much the quality of the diagnose. As an example, Griffin et al. (2005) suggest heart rate to predict neonatal sepsis, instead of obtaining blood from the infants for laboratory tests. This is also useful for the companies shares prices, whose time series are to be paid but, once done, all the historical records are available to use. For instance, generally in economy a three factor model is assumed to have good properties to predict asset pricing anomalies (Chen and Zhang, 2010). For this reason, many papers in this field have been devoted to determine the three most significant features that help predicting those anomalies (Kogan and Tian, 2015). In all these cases the number of causal features and the number of dependencies from each causal feature are valued differently: as we might seek to limit the first one, we might not be that strict with the second.

This subject was previously noted by Lozano et al. (2009), who stated that *"as a method of Granger graphical modeling, the relevant variable selection question is not whether an individual lagged variable is to be included in regression, but whether the lagged variables for a given time series as a group (i.e. the feature), are to be included"*. To address this issue they proposed to use Group Lasso, in which all the lagged variables of a feature were assigned to the same group. Although this approach reinforces to choose all the past values of a feature once one of them has already been selected for the prediction, it still does not grant control over the exact level of sparsity of the outcome, like none of the other Lasso methods. Moreover, the Lasso approaches perform a shrinkage over the coefficients which, as will be seen in Section 2.4, might not be advisable when highly sparse graphs are sought.

In contrast we propose a novel sparse approach, formulated as a mathematical optimization problem, which endows the user with the power to control different aspects of the sparsity of the solution. Sparsity is meant here in some of its many different dimensions, as the total number of nonzero coefficients, the total number of features used for the forecast, or the number of past observations of each feature used by the model to make predictions. The performance of this sparsity-controlled VAR method (SC-VAR henceforth) will be compared with three benchmark approaches, namely, the VAR and both the standard and Group Lasso, on simulated and real-life multivariate time series. The results show that the proposed approach outperforms the benchmark

Lasso methods in terms of prediction errors when highly sparse graphs are sought, while avoiding the VAR's overfitting for more dense graphs.

The chapter is structured as follows. Next section introduces mathematically the three benchmark methods, the VAR, the classic Lasso and the Group Lasso, and motivates our approach. In Section 2.3 the SC-VAR model is introduced and expressed as a mixed integer non-linear optimization program, solvable by standard optimization software. Also, a discussion about the choice of the parameters of the model is included. Competing approaches are compared against the proposed method in Section 2.4 in both simulated and real datasets. Finally, conclusions and future extensions are collected in Section 2.5.

2.2 Preliminaries

Let $\{\mathbf{X}_t\}_{t\geq 0}$ be an N-dimensional vector autoregressive process of order n, VAR(n), i.e., each series i, $i = 1, ..., N$, can be expressed as in (1.8). The usual estimation procedures for the coefficients α^i and θ^i_{jk} are Maximum Likelihood (which implies making distributional assumptions on the errors e^i_t) or, without imposing any statistical assumption, the Ordinary Least Squares method:

$$\min_{\mathbf{c},\mathbf{A}} \sum_{i=1}^{N}\sum_{t=n}^{T}\left(X^i_{t+1} - \alpha^i - \sum_{j=1}^{N}\sum_{k=1}^{n}\theta^i_{jk}X^j_{t+1-k}\right)^2 \tag{2.1}$$

where $\mathbf{c} = (\alpha^i)_i \in \mathbb{R}^{1\times N}$ and $\mathbf{A} = (A^1|...|A^N) \in \mathbb{R}^{N\times N\cdot n}$ stand for all unknown coefficients of the process to be estimated. Here $A^i = (\theta^i_{jk})_{jk}$ represents the $N \times n$ matrix of coefficients used to model series i.

There is no reason to expect sparsity in the estimates obtained by maximum likelihood estimation or by solving the nonlinear program (2.1), and therefore, it may be difficult to visualize causalities while leading to overfitting (Kalli and Griffin, 2014; Li, 2012; Kojima et al., 2009). Among the different procedures proposed in the literature with the aim of obtaining more sparse solutions, a prominent role is given to the Lasso-VAR (Lasso thereafter), in which an ℓ_1 regularization term is added to the objective function (2.1), and thus estimates are obtained by solving the following nonlinear nonsmooth optimization program:

$$\min_{\mathbf{c},\mathbf{A}} \sum_{i=1}^{N}\sum_{t=n}^{T}\left(X^i_{t+1} - \alpha^i - \sum_{j=1}^{N}\sum_{k=1}^{n}\theta^i_{jk}X^j_{t+1-k}\right)^2 + \sum_{i=1}^{N}\lambda^i\left(\sum_{j=1}^{N}\sum_{k=1}^{n}|\theta^i_{jk}|\right). \tag{2.2}$$

As previously mentioned, when features are presented as time series, the use of Group Lasso is encouraged in the literature if more than one-lagged values are considered; see, for instance, Lozano et al. (2009). Such a method groups lagged variables of the same feature, giving more importance to the number of causal features rather than to the overall number of dependencies depicted. This is done by adding an ℓ_2-penalty separately for each group (Yuan and Lin, 2006):

$$\min_{\mathbf{c},\mathbf{A}} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{t+1}^{i} - \alpha^{i} - \sum_{j=1}^{N} \sum_{k=1}^{n} \theta_{jk}^{i} X_{t+1-k}^{j} \right)^{2} + \sum_{i=1}^{N} \lambda^{i} \left(\sum_{r=1}^{R} \|\boldsymbol{\theta}_{\boldsymbol{j(r)}}^{\boldsymbol{i}}\|_2 \right) \tag{2.3}$$

where $\boldsymbol{\theta}_{\boldsymbol{j(r)}}^{\boldsymbol{i}} = \left(\theta_{j1}^{i}, \theta_{j2}^{i}, ..., \theta_{jp}^{i}\right)$. On the other hand, the Adaptative Lasso considers a weighted ℓ_1 penalization in which shrinkage is tried to be softened for significant features.

Infinitely many solutions may be obtained for the Lasso approaches when varying the parameter $\lambda^{i} \in \mathbb{R}^{+}$, where more sparse solutions are attained when increasing the value of this penalty (in fact, note that all $\theta_{jk}^{i} = 0$ when $\lambda^{i} \to +\infty$). However, although one may think that these popular Lasso approaches already cope with sparsity-controlled graphs, there is no way to discern the exact level of sparsity of the outcome when values for the penalties λ^{i} are given a priori, calling for a (multidimensional) parameter tuning which, at the end, show solutions with different levels of sparsity, but there is no guarantee that sparsity is fulfilled. Moreover, these methods perform a shrinkage over the coefficients, which may deteriorate their prediction power when highly sparse graphs are sought. Even although these methods often provide reasonable sparse solutions, according to some authors (Bertsimas and Copenhaver, 2014; Caramanis et al., 2012) adding an ℓ_1 or ℓ_2 penalty is a robust approach rather than a sparse method. In contrast, Bertsimas and Copenhaver (2014) present Mixed Integer Programming as a useful tool to attain sparsity. This recommendation is supported by the recent improvement on integer optimization solvers, which can attain considerably *good* solutions at a reasonable computational cost. In this chapter a sparsity controlled VAR is formulated as a MINLP, in which we manage to control different aspects of the sparsity of the outcome.

When control over sparsity is sought, a natural approach is as follows: start with a standard VAR and then, in a naive way, select the largest estimated coefficients (in absolute value) and set to zero the remaining (smaller) coefficients. This naive approach is easy to implement, quick to execute and sounds reasonable. However, its performance may be poor, as we show next. Consider the top left panel of Figure 2.1, where a simulated VAR of order $n = 3$ is graphically represented. The multivariate time series is represented as a directed graph; nodes correspond to the different features, i.e., the different one-dimensional time series composing the multivariate time series; edges in the graph visualize causality: an arrow from node j to node i means that the model uses feature j in the forecast of feature i; the thickness of the edges is proportional to the magnitude (in absolute value) of the coefficient relating the features, and therefore it measures the causality's strength when the series are normalized. The color of the edges is related to the lag: the arrow is plotted in black if the present (t) value of a feature is related with a data one period behind $(t-1)$, red for two periods $(t-2)$ and green for three $(t-3)$. Here node 1 receives arrows from nodes 2 and 3, meaning that, in order to forecast feature 1, past values of features 2 and 3 are used. Note also that node 1 receives here three arrows from node 3, so the present value t of feature 1 is caused by the previous three values $(t-1)$, $(t-2)$ and $(t-3)$ of series 3.

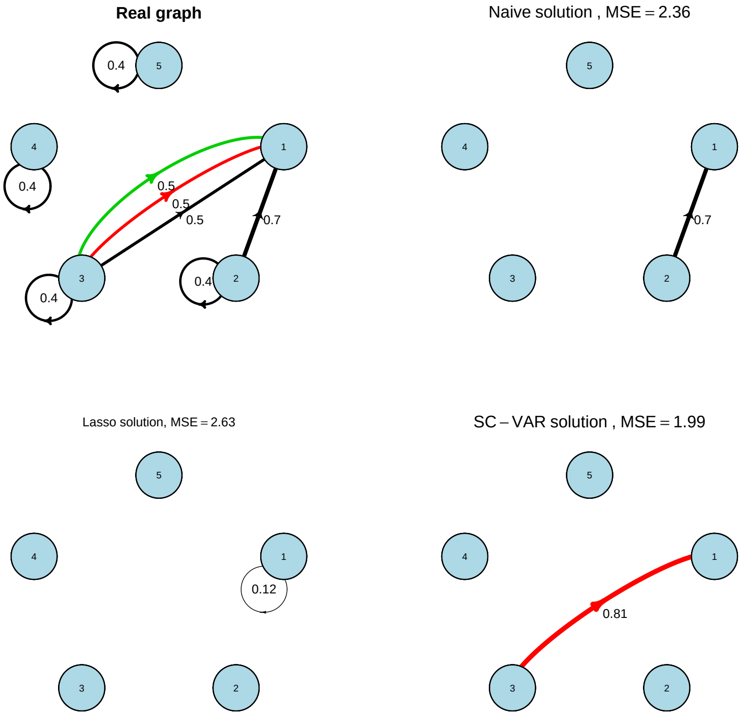

Real graph
Naive solution , MSE = 2.36
Lasso solution, MSE = 2.63
SC – VAR solution , MSE = 1.99
0.4
0.5
0.7
0.12
0.81
1
2
3
4
5

Figure 2.1: Graphical representation of a simulated VAR and its sparse counterparts. Naive algorithm (center) versus the proposed SC-VAR method (right).

Assume that only one observation of one single feature is allowed to be used to explain a feature. Then the naive approach, illustrated in the top right panel of Figure 2.1, underestimates strong persistence in favor of just one significant coefficient. This undesired phenomenon is a consequence of the nature of the naive approach: arrows are considered to be kept or removed one by one, and thus the overall picture may be lost. For this reason, instead of using the above-described naive approach, we suggest to express the problem of arrows selection as a mathematical optimization model, solvable by current standard numerical optimization routines. In particular, when applied to the time series of the example, the output of our procedure is visualized through the directed graph in Figure 2.1 (bottom right panel). It can be observed that the MSE obtained by the SC-VAR solution is considerably smaller than that of the naive approach. In order to illustrate the effect of the shrinkage of Lasso methods over the forecasting power we depict the solution under the standard Lasso in the bottom left panel. This effect is clear since the arrow is thinner, which could be detrimental when highly sparse solutions are sought. This phenomenon, also observed in the results of the experiments carried out in Section 2.4, will be discussed in more detail. Also it is interesting to note that the chosen causal feature differs for each method: as variable 1 is considered to cause itself for the Group Lasso, feature 3 is considered for our approach instead. This is not surprising since it is known that the VAR might not be identifiable (Lütkepohl, 2005).

2.3 The SC-VAR.

In this section we will first discuss the sparsity parameters in the SC-VAR model (Section 2.3.1), which will be later written as a mathematical optimization problem (Section 2.3.2). In Section 2.3.3 we will briefly discuss the choice of parameters of our model.

2.3.1 Different aspects of the sparsity

To the best of our knowledge the existent sparse VAR models attempt to gain overall sparsity, without an explicit analysis of the different aspects of sparsity. The Lasso approaches do not grant the user with the ability to manage the sparsity of the solution in some desirable dimensions either. Indeed, the number of features allowed to be used by the model to explain feature i (V_S^i) or the overall number of nonzero coefficients (V_A), are different measures usually masked under the generic term of sparsity. Other aspects of the sparsity, that will be also under the user's control in our proposed sparse VAR, are the number of non-zeroes used to explain feature i (V_T^i) or the number of observations per each causal feature of variable i (V_{Sa}^i). Moreover, when series are normalized then the strength of a potential causality might be related with the magnitude of its associated coefficient. Hence, in order to avoid spurious dependencies, clear cut-offs ϵ_j^i will be introduced, so that only coefficients with an absolute value greater than or equal to ϵ_j^i are allowed when relating feature i with feature j. All these parameters, included in our SC-VAR, allow the user to obtain an output with the desired level and structure of sparsity.

2.3.2 Mathematical Programming formulation

The objective of the SC-VAR approach if twofold: control the sparsity of the solution while not damaging much the forecasting capacity. Therefore, in our SC-VAR model, the VAR estimates are obtained by solving the optimization problem (2.1) imposing on the coefficients of $\mathbf{A}$ the bounds represented by the sparsity parameters $V_A, V_T^i, V_S^i, V_{Sa}^i$ and ϵ_j^i above. They can be expressed as linear constraints by adding logical (binary) variables. Indeed, define the variables δ_{jk}^i to indicate whether a coefficient θ_{jk}^i is zero or not, and variables γ_j^i to indicate if feature j is meant to cause variable i (i.e., if $\theta_{jk}^i \neq 0$ for some k).

Now the SC-VAR model is formulated as the optimization problem (2.4)-(2.11), whose outputs are the estimates of the sparse coefficients α^i and θ_{jk}^i, as well as the solution for the indicator variables δ_{jk}^i and γ_j^i.

$$\min_{\mathbf{c},\mathbf{A},\Delta,\Gamma,} \quad \sum_{i=1}^{N}\sum_{t=p}^{T}\left(X_{t+1}^i - \alpha^i - \sum_{j=1}^{N}\sum_{k=1}^{n}\theta_{jk}^i X_{t+1-k}^j\right)^2 \tag{2.4}$$

s.t

$$\delta_{jk}^i \leq \gamma_j^i \qquad \forall k \in K, j, i \in I \tag{2.5}$$

$$\sum_{j=1}^{N}\gamma_j^i \leq V_S^i \qquad \forall i \in I \tag{2.6}$$

$$\sum_{k=1}^{n}\gamma_j^i\delta_{jk}^i \leq V_{Sa}^i \qquad \forall j, i \in I \tag{2.7}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{n}\delta_{jk}^i \leq V_A \tag{2.8}$$

$$\sum_{j=1}^{N}\sum_{k=1}^{n}\delta_{jk}^i \leq V_T^i \qquad \forall i \in I \tag{2.9}$$

$$M\delta_{jk}^i \geq |\theta_{jk}^i| \geq \epsilon_j^i\delta_{jk}^i \qquad \forall k \in K, j, i \in I \tag{2.10}$$

$$\delta_{jk}^i, \gamma_j^i \in \{0,1\} \qquad \forall k \in K, j, i \in I \tag{2.11}$$

where $K = \{1,...,p\}$, $I = \{1,...,N\}$, $\Delta = (\delta_{jk}^i)_{i,j,k}$, $\Gamma = (\gamma_j^i)_{i,j}$ and M is a *large* constant.

Let us briefly discuss the correctness of the formulation above. The objective function (2.4) minimizes the sum of squared errors. Constraint (2.5) forces the variable γ_j^i to take the value 1 when some δ_{jk}^i takes the value 1, i.e., as soon as some θ_{jk}^i is non-zero. The remaining constraints model different aspects of the sparsity of the process. Indeed, constraints (2.6) and (2.7) bound the number of features that are said to cause variable i

and the number of non-zero coefficients per each of the chosen causal features of variable i, respectively, for $i = 1, .., N$. Constraints (2.8) and (2.9) bound respectively the total number of non-zero entries in matrix $\mathbf{A}$ and the total number of non-zero coefficients for each variable i. The shrinking parameter ϵ^i_j is included in the model via constraint (2.10), which assigns zero to any coefficient that is not allowed to appear on the model, but otherwise it requires $|\theta^i_{jk}|$ to belong to the interval $[\epsilon^i_j, M]$. Here M is assumed to be a *large* fixed number, and thus this constraint does not exclude reasonable values of the parameters θ^i_{jk}.

Problem (2.4)-(2.11) is a MINLP (Burer and Letchford, 2012; Lee and Leyffer, 2012), with convex quadratic objective function. See e.g. Bertsimas and Copenhaver (2014); Bertsimas and Mazumder (2014) for other statistical problems recently addressed via optimization in integer numbers. All constraints are linear, except for (2.10). However, this can be rewritten by introducing new auxiliary variables ν^{i+}_{jk}, ν^{i-}_{jk}, allowing to reformulate Problem (2.4)-(2.11) as:

$$\begin{aligned}
&\min_{\mathbf{c},\mathbf{A},\Delta,\Gamma,} \sum_{i=1}^{N}\sum_{t=p}^{T}\left(X^i_{t+1} - \alpha^i - \sum_{j=1}^{N}\sum_{k=1}^{n}\theta^i_{jk}X^j_{t+1-k}\right)^2 \\
&\quad \text{s.t} \\
&\qquad (2.5)\text{-}(2.9) \\
&\qquad \alpha^i_{jk} \geq \epsilon^i_j \nu^{i+}_{jk} - \nu^{i-}_{jk} M && \forall k \in K, j, i \in I \\
&\qquad \alpha^i_{jk} \leq -\epsilon^i_j \nu^{i-}_{jk} + \nu^{i+}_{jk} M && \forall k \in K, j, i \in I \\
&\qquad \gamma^i_j, \nu^{i+}_{jk}, \nu^{i-}_{jk} \in \{0,1\} && \forall k \in K, j, i \in I
\end{aligned} \tag{P}$$

where δ^i_{jk} has been replaced by $\nu^{i+}_{jk} + \nu^{i-}_{jk}$ in constraints (2.5)-(2.9). Note that the new constraints require that either $-M \leq \theta^i_{jk} \leq -\epsilon^i_j$ or $\epsilon^i_j \leq \theta^i_{jk} \leq M$ if coefficient θ^i_{jk} is chosen to appear in the model (i.e., if $\delta^i_{jk} = 1$). Now, Problem (P) is a MINLP with quadratic convex objective function and linear constraints. Hence, it can be solved using standard solvers, such as `CPLEX` or `Gurobi`, and it can be easily written using a simple algebraic language such as `AMPL` (Fourer et al., 2002). The lines of the `AMPL` code are included in Appendix at the end of this chapter.

Note that (2.8) is the only constraint of Problem (P) linking the count of the non-zeroes of all features. Hence, if constraint (2.8) were redundant, the separability of the objective function (2.4) would allow to solve Problem (P) by solving separately the problem for each feature i :

$$\min_{\alpha^i, A^i} \sum_{t=p}^{T} \left(X_{t+1}^i - \alpha^i - \sum_{j=1}^{N} \sum_{k=1}^{n} \theta_{jk}^i X_{t+1-k}^j \right)^2 .$$

Since such problems are of much smaller dimension, we see that removing constraint (2.8) would allow one to cope with databases with a larger number of time series.

2.3.3 Choice of parameters

Compared with other methods seeking sparsity, our approach has many more parameters, which may be, at first glance, discouraging. Some comments are in order to treat this issue. First note that, as opposed to the Lasso approaches, the parameters which are to be determined in Problem (P) have a precise meaning in terms of the sparsity, and therefore can be fixed by the user a priori without carrying out any tuning if a determined level of sparsity is sought. In other words, they are not parameters to tune but decisions to make with respect to the sparsity desired. Second, although apparently the proposed methodology consists of four parameters per series (V_T^i, V_S^i, V_{Sa}^i and ϵ_j^i) and one global parameter (V_A), not all of them need to be fixed to address the problem. Indeed, it suffices to fix the parameters that are significant for the user and choose the rest of them so that the remaining constraints are redundant. For instance, if the user only cares about the number of causal features, it suffices to fix only V_S^i. As an illustration consider Figure 2.1, discussed in Section 2.2, where the SC-VAR approach was solved fixing nothing but $V_T^1 = 1$. Nevertheless, if the modeler does not possess any information about the data and/or lack preferences regarding the sparsity of the outcome, the parameters of concern may be tuned taking into account the preferred trade off between the predictive quality and the sparsity of the output, as detailed in Section 2.4.4.

Third, Problem (P) can be solved by fixing solely V_A, which represents the overall sparsity of the whole graph. This means that the SC-VAR can be solved for all features simultaneously, while the Lasso approaches are always solved separately for each feature. This provides our model with more flexibility than Lasso methods to obtain a graph with a specified level of overall sparsity, placing pools of non-zeroes where necessary, as will be seen in Figure 2.3. Summarizing, the more thorough the user wants to be with the structure of the solution, the more parameters she has to fix. This is the price to pay to control sparsity, if sought, in many of its dimensions.

2.4 Numerical illustrations

In this section the SC-VAR is compared with five benchmark approaches, the VAR, the Lasso, the Group Lasso, the Adaptative Lasso and a greedy approach on simulated as well as real datasets. Further descriptions on such methods and the choice of the parameters used are provided below.

2.4.1 Comparison methodology

In this chapter five different methodologies were chosen to compare against out SC-VAR approach. They were: the VAR, the Lasso, the Group Lasso, the Adaptative Lasso, and a greedy approach.

1. The **VAR**, whose solutions were obtained by OLS.

2. The **Lasso**, whose coefficients were obtained via the Least Angle Regression algorithm, LAR. This algorithm solves the Lasso for each variable i separately. Since at each step the LAR incorporates a new predictor for variable i, it provides a set of solutions with different levels of sparsity (i.e. different number of non-zeroes per node, V_T^i). The Lasso set of solutions was obtained by using the `lars()` function of `R-cran` package `lars` (Hastie and Efron, 2013).

3. The **Adaptative Lasso**, whose coefficients were also obtained by using the same `lars()` function. Its weights were fixed to $1/|\hat{\alpha}^i_{jk,VAR}|^\gamma$, by applying the algorithm proposed in Zou (2006).

4. The **Group Lasso**, whose solutions were obtained by using the functions of the `R-cran` package `gglasso` (Yang and Zou, 2015). In particular, the function `cv.gglasso` was used to perform a 5-fold cross-validation over the fits obtained for each value of the penalties λ^i and calculate their mean cross-validated errors. Recall that the Group Lasso understands sparsity in terms of the number of causal features, rather than the overall number of non-zero coefficients per node. As the sparsity of the output cannot be discerned in advance, we had to tune its parameters: we solved the Group Lasso for a grid of λ^i large enough, so solutions for all possible values of V_S^i were obtained.

5. The **greedy** method, which consists on applying a refitting to the feasible solution obtained by the naive approach described on Section 2.2; i.e., the non-zero coefficients are chosen to be the ones with a larger coefficient in absolute value for the VAR, and then those coefficients are optimized so as to minimize the OLS (2.1).

The SC-VAR solution was obtained by solving Problem (P) using `Gurobi` version `6.5.0`. As previously commented, the Lasso approaches are solved independently for each node. Hence, in order to fairly compare against these methods, the parameter V_A was fixed so as the constraint (2.8) was redundant. In this way, Problem (P) could be solved separately for each feature.

With the purpose of testing the influence of the shrinking parameter of constraint (2.10), the SC-VAR was solved for $\epsilon = 0$ and $\epsilon = 0.2$, where $\epsilon^i_j = \epsilon$ for all i, j. However, we observed the obtained forecasting qualities were often similar but more sparse solutions were obtained for $\epsilon = 0.2$, thus generally we only report here the results for $\epsilon = 0.2$ to conserve space. In contrast to the VAR, some constraints of the SC-VAR bounds the sparse coefficients by a constant M. Observe that, as customary in the literature, the choice of such M is problematic, since a very small value may exclude reasonable values of the coefficients θ^i_{jk}, whilst a very large value of M is to cause severe numerical troubles (Camm et al. (1990)). In our experiments M was fixed to 2, although other values were tested, with similar performance.

Concerning the computational costs, a time limit of 30 seconds per node was imposed for the SC-VAR. However, this upper bound was only reached for the large simulated data sets, that will be discussed in Section 2.4.2 of the manuscript. For the rest of databases sparse graphs were generally obtained in less than 2 seconds. A more detailed analysis of the computational times will also be carried out in Section 2.4.2.

In order to test the performance of all the approaches under different perspectives, two ways of comparison were considered. In the first one, the modeler knows a priori the level of the sparsity to be attained and thus the parameters are fixed beforehand, and in the second one she does not possess such an information and therefore the parameters are tuned. Hence, first the problems were solved for all possible combinations of parameters V^i_T or V^i_S so as to carry out a deeper analysis on the performance of the approaches under different requirements of sparsity. For VAR(1) processes, the greedy and the SC-VAR were compared against the Lasso, since no grouping effects amongst lagged variables were necessary for these cases. For VAR(n) processes with $p > 1$ the greedy and the SC-VAR were also compared against the Group Lasso when it was sought to test the influence of the number of causal features over the prediction quality. In this case each time series was divided in train (50% of the data) and test (50%) sets. The solutions for each method were obtained using only the data of the train set, and the MSE for such solutions was calculated in the test set. These results are reported in Sections 2.4.2 and 2.4.3.

In the second comparison methodology, the parameters $V^i_T \in \{1, ..., N\}$ were tuned, as proposed at the end of Section 2.3.3. In this case, the Adaptative Lasso was considered so as to compare with a more competitive lasso approach. The parameter γ of this method was also tuned using the same values proposed in Zou (2006), i.e., $\gamma \in \{0.5, 1, 2\}$.

The case $\gamma = 0$ was also considered so as to encompass the classic Lasso in our study. Here each time series was divided in train, test and validation sets ($50\%, 25\%$ and 25%, respectively). The solutions for each method were obtained using only the data of the train set, from which the output that minimized the MSE in the test set was chosen, and the reported MSE for that solution was calculated in the validation set. The results are reported in Section 2.4.4.

2.4.2 Simulation study

Data generation

In order to test the performance of all the approaches we generate synthetic data following VAR(1) and VAR(2) processes of 10 nodes with i.i.d. errors drawn from a standard Normal distribution. To generate the coefficients of such multivariate time series we roughly follow the experiments conducted in Arnold et al. (2007); Lozano et al. (2009), in which an *affinity* parameter is fixed. Such a parameter is the probability that an edge is included in the graph. For example, if we want a graph to have a 20% density we fix the *affinity* parameter to 0.2. Then, we randomly generate all the coefficients of the process matrix $\mathbf{A}$ and decide whether each off-diagonal element is included by simulating a Binomial distribution with success probability 0.2.

So as to test the performance of the sparse approaches under graphs with different levels of sparsity, VAR processes with densities 0%, 10%, 20%,...,100% have been generated. A 0% density means that each feature follows an independent autoregressive process (diagonal matrix) and a 100% density implies that there exists correlation amongst all nodes. For each level of density 100 instances of VAR processes were generated. Each time series has 1000 observations, whose first 500 were assigned to the train set and the remaining to the test set. Although the VAR, Lasso, Group Lasso, SC-VAR and greedy were solved for all levels of density, the results were mainly equivalent and thus only the results for simulations with a 10%, 50% and 90% density are included here. Analogously, in order to test the performance of the SC-VAR for large databases, 5 instances of VAR(1) processes with 200 and 500 nodes were generated. In these cases the processes were generated with a 5% density and parameter V_T was fixed to the number of expected non-zeroes.

Table 2.1 reports the median of the normalized MSEs and the standard deviations (in brackets) for such densities. Specifically, Panel A of Table 2.1 reports the results for Lasso, SC-VAR and greedy for VAR(1) processes, while Panel B of Table 2.1 reports them for Group Lasso, SC-VAR and greedy for VAR(2) processes. The best method in terms of MSE has been highlighted in bold for each case.

	Density 10%			Density 50%			Density 90%		
	Lasso	SC-VAR	Greedy	Lasso	SC-VAR	Greedy	Lasso	SC-VAR	Greedy
Panel A. Simulated datasets with $n = 1$									
$V_T = 1$	1.09	**1.04**	**1.04**	1.23	1.15	**1.13**	1.23	**1.17**	1.18
	(0.36)	(0.14)	(8.87)	(0.33)	(0.23)	(0.32)	(0.35)	(0.26)	(0.86)
$V_T = 2$	1.03	**1.00**	**1.00**	1.15	1.07	**1.06**	1.20	**1.11**	**1.11**
	(0.38)	(0.17)	(0.33)	(0.34)	(0.25)	(0.17)	(0.36)	(0.26)	(0.45)
$V_T = 3$	1.01	**0.99**	**0.99**	1.09	1.03	**1.02**	1.15	1.08	**1.07**
	(0.41)	(0.19)	(0.08)	(0.36)	(0.26)	(0.07)	(0.38)	(0.27)	(0.25)
$V_T = 4$	**0.99**	**0.99**	**0.99**	1.05	1.02	**1.01**	1.12	**1.05**	**1.05**
	(0.42)	(0.19)	(0.07)	(0.39)	(0.27)	(0.03)	(0.40)	(0.28)	(0.16)
$V_T = 5$	**0.99**	**0.99**	1.00	1.03	1.01	**1.00**	1.09	1.04	**1.03**
	(0.43)	(0.19)	(0.01)	(0.41)	(0.28)	(0.01)	(0.44)	(0.29)	(0.08)
$V_T = 6$	1.00	**0.99**	1.00	1.01	1.01	**1.00**	1.06	1.04	**1.01**
	(0.43)	(0.19)	(0.00)	(0.42)	(0.29)	(0.01)	(0.48)	(0.29)	(0.04)
$V_T = 7$	1.00	**0.99**	1.00	**1.00**	1.01	**1.00**	1.03	1.03	**1.00**
	(0.43)	(0.19)	(0.00)	(0.45)	(0.29)	(0.00)	(0.53)	(0.30)	(0.02)
$V_T = 8$	1.00	**0.99**	1.00	**1.00**	1.01	**1.00**	1.02	1.03	**1.00**
	(0.43)	(0.19)	(0.00)	(0.48)	(0.29)	(0.00)	(0.57)	(0.30)	(0.01)
$V_T = 9$	1.00	**0.99**	1.00	**1.00**	1.01	**1.00**	**1.00**	1.03	**1.00**
	(0.44)	(0.19)	(0.00)	(0.50)	(0.29)	(0.00)	(0.65)	(0.30)	(0.00)
$V_T = 10$	1.00	**0.99**	1.00	**1.00**	1.01	**1.00**	**1.00**	1.03	**1.00**
	(0.45)	(0.19)	(0.00)	(0.52)	(0.29)	(0.00)	(0.74)	(0.30)	(0.00)
Panel B. Simulated datasets with $n = 2$									
$V_S = 1$	1.13	**0.95**	1.06	1.11	**1.01**	1.19	1.12	**1.02**	1.29
	(0.36)	(0.14)	(0.88)	(0.33)	(0.23)	(0.97)	(0.35)	(0.26)	(6.56)
$V_S = 2$	1.15	**0.90**	0.99	1.11	**0.96**	1.12	1.13	**0.99**	1.19
	(0.38)	(0.17)	(0.17)	(0.34)	(0.25)	(0.49)	(0.36)	(0.26)	(0.69)
$V_S = 3$	1.16	**0.90**	0.99	1.14	**0.90**	1.07	1.14	**0.94**	1.14
	(0.41)	(0.19)	(0.03)	(0.36)	(0.26)	(0.24)	(0.38)	(0.27)	(0.42)
$V_S = 4$	1.17	**0.90**	0.99	1.17	**0.89**	1.04	1.15	**0.91**	1.10
	(0.42)	(0.19)	(0.01)	(0.39)	(0.27)	(0.16)	(0.40)	(0.28)	(0.29)
$V_S = 5$	1.18	**0.90**	0.99	1.20	**0.87**	1.01	1.18	**0.89**	1.06
	(0.43)	(0.19)	(0.00)	(0.41)	(0.28)	(0.13)	(0.44)	(0.29)	(0.19)
$V_S = 6$	1.19	**0.90**	0.99	1.24	**0.87**	1.00	1.21	**0.88**	1.04
	(0.43)	(0.19)	(0.00)	(0.42)	(0.29)	(0.11)	(0.48)	(0.29)	(0.14)
$V_S = 7$	1.20	**0.90**	1.00	1.28	**0.87**	1.00	1.25	**0.87**	1.02
	(0.43)	(0.19)	(0.00)	(0.45)	(0.29)	(0.10)	(0.53)	(0.30)	(0.11)
$V_S = 8$	1.20	**0.90**	1.00	1.30	**0.87**	1.00	1.31	**0.87**	1.01
	(0.43)	(0.19)	(0.00)	(0.48)	(0.29)	(0.10)	(0.57)	(0.30)	(0.10)
$V_S = 9$	1.21	**0.90**	1.00	1.33	**0.87**	1.00	1.40	**0.87**	1.00
	(0.44)	(0.19)	(0.00)	(0.50)	(0.29)	(0.10)	(0.65)	(0.30)	(0.10)
$V_S = 10$	1.24	**0.90**	1.00	1.36	**0.87**	1.00	1.48	**0.87**	1.00
	(0.45)	(0.19)	(0.00)	(0.52)	(0.29)	(0.10)	(0.74)	(0.30)	(0.10)

Table 2.1: Normalized median MSE (and standard deviations) under the SC-VAR, greedy and the Lasso for VAR(1) processes, and the SC-VAR, greedy and the Group Lasso for VAR(2) processes, all generated with a 10, 50 and 90% density, for different requirements of sparsity in terms of number of causal features (V_S=1,...,10)

Analysis of the results

From Panel A of Table 2.1 it can be observed that the SC-VAR outperforms the Lasso in terms of MSE when highly sparse graphs are sought (that is, for small values of V_T). This outperformance over the Lasso increases as the density of the process does. Indeed, the difference between the Lasso and the SC-VAR is at least a 5% deterioration over the VAR MSE when the maximum number of non-zeroes required is less than 2, 4 or 6 for processes with densities 10%, 50% and 90%, respectively.

While the sparsity-controlled approach can sometimes attain an MSE 9% better than the Lasso ($V_T = 2$, 90% density), the Lasso outperforms it for highly dense graphs ($V_T \geq 8$, 90% density). Furthermore, the standard deviations suggests that the SC-VAR is a more stable approach in all cases. In comparison with the greedy approach we could conclude that in terms of MSE both methods perform similarly for $n = 1$. However, the greedy seems to become quite unstable when highly sparse graphs are sought, although its standard deviations decrease very quickly when the number of zeroes of the output does.

From Panel B of Table 2.1 it can be concluded that the SC-VAR usually helps avoiding VAR overfitting. Note that the SC-VAR outperformance over the VAR is clearer for $n = 2$ than for $n = 1$, which is reasonable since the number of parameters of the VAR increases with n. As solving the SC-VAR with fixing nothing but $V_S = 10$ is equivalent to the standard VAR, the improvement over its MSE is thanks to the choice $\epsilon = 0.2$. This poor performance in comparison with the SC-VAR also affects the greedy, which always attains a worse MSE than the proposed approach. Specifically, this behaviour is even more evident when highly sparse graphs are sought. For instance, the MSE of the greedy presents a 27% more of deterioration over the VAR than the SC-VAR for VAR(2) processes with 90% of density and $V_S = 1$.

In Panel B of Table 2.1 it can be observed that the SC-VAR always outperforms the Group Lasso in terms of MSE, the outperformance being clearer as the sparsity of the output decreases. Moreover, it seems that the SC-VAR performs equivalently with respect to the VAR no matter the density of the original graph, but the Group Lasso performs worse for large V_S in truly dense processes. For instance, while the SC-VAR improves the VAR's forecasting error in a 13% for $V_S = 10$ and processes with 90% density, the Group Lasso attains a 48% of worsening.

In order to test the ability of the different approaches to recover the original graph, the median F_1 score is depicted in Figure 2.2 for $n = 1$ and $n = 2$. In this case, the results for the SC-VAR with $\epsilon = 0$ were reported so as to show that its performance, although different from that of $\epsilon = 0.2$, is mainly equivalent to that of the greedy. From Figure 2.2 some conclusions arise. First, considering an $\epsilon > 0$ may help avoiding the overfitting. This is observed clearly for VAR processes with 10% density, where the capability of the approaches to recover the original graph deteriorates as more non-zeroes are allowed in the model. In this case, the F_1 score for the SC-VAR with $\epsilon = 0.2$ holds constant, since constraint (2.10) avoids adding all the variables to the model by requiring to include only relevant features. Second, although a *large* ϵ may avoid overfitting and may focus solely on the most significant features, it might not be advisable if a dense graph wants to be recovered completely. To see this, pay attention to the performance of the SC-VAR with $\epsilon = 0.2$ for VAR processes with 90% density. However, note that although the rest of approaches attain a better F_1 in these cases, the SC-VAR with $\epsilon = 0.2$ yields a better MSE. Therefore, it is an useful tool to discover relevant causalities without damaging much the prediction power.

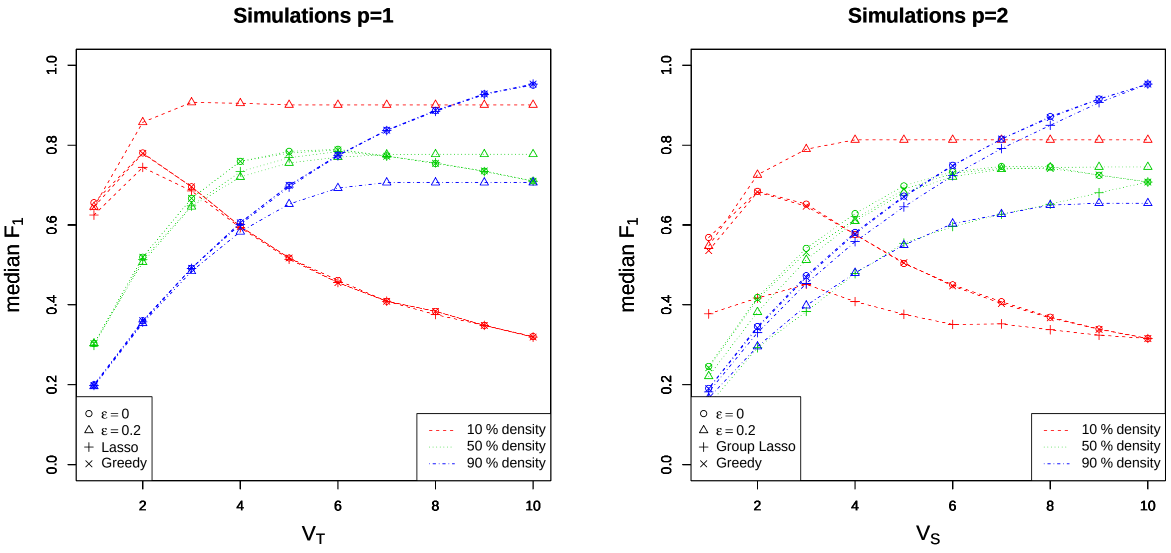


Figure 2.2: Median F_1 measure for Lasso/Group Lasso, greedy and SC-VAR with $\epsilon = 0, 0.2$ for simulated databases with $n = 1$ and $n = 2$ and different levels of densities.

In order to illustrate more in depth the results, randomly selected instances are depicted in Figure 2.3, together with their VAR solutions. For the sake of abbreviation, only the most extreme solutions of the SC-VAR and the Lasso ($V_T = 1$ and $V_T = 10$) have been depicted. From Figure 2.3 some observations become apparent. First, the VAR leads to overfitting. This is clearer as more sparse the real graph is. Second, the Lasso is equivalent to the VAR for $V_T = 10$. However, the SC-VAR attains a much more sparse solution while providing a similar MSE. It seems that requiring the absolute value of the non-zero coefficients to be larger than a threshold ($\epsilon = 0.2$ in constraint (2.10)) helps avoiding overfitting in these particular cases, leading to graphs that are more similar to the original ones. Observe that this idea of defining a clear cut between non-zero and zero coefficients is rather different to the behaviour of the Lasso, which shrinks coefficients towards zero.

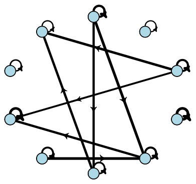
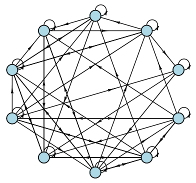
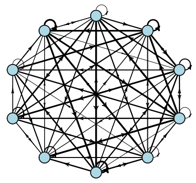
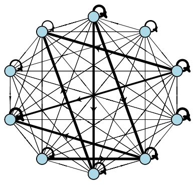
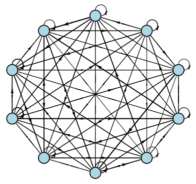
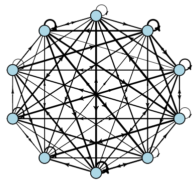
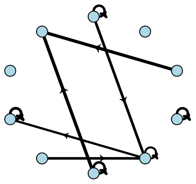
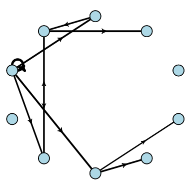
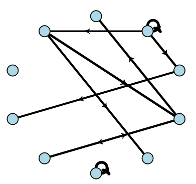
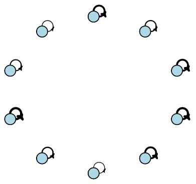
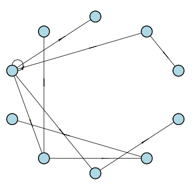
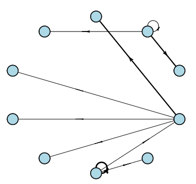
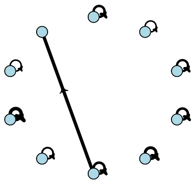
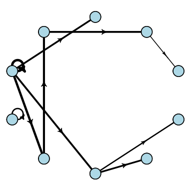
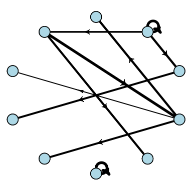
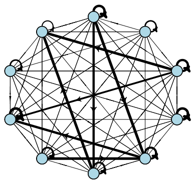
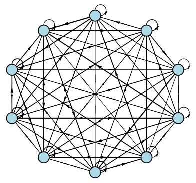
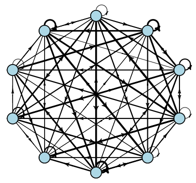
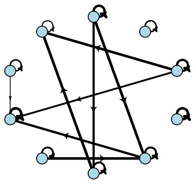
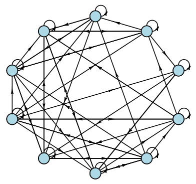
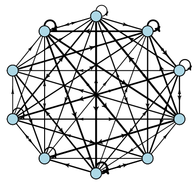
Density		10%	50%	90%
Real				
VAR		MSE = 1 	MSE = 1 	MSE = 1 
$V_A = 10$	SC-VAR	MSE = 1.19 	MSE = 1.32 	MSE = 1.78 
$V_T = 1$	Lasso	MSE = 1.54 	MSE = 1.9 	MSE = 2.4 
	SC-VAR	MSE = 1.34 	MSE = 1.35 	MSE = 1.8 
$V_T = 10$	Lasso	MSE = 1 	MSE = 1 	MSE = 1 
	SC-VAR	MSE = 0.99 	MSE = 1 	MSE = 1.03 

Figure 2.3: Real graphs of randomly selected instances of processes with different densities, represented along with their VAR solutions. SC-VAR and Lasso solutions are represented for a couple of levels of sparsity, $V_T = 1$ and $V_T = 10$, and the SC-VAR is also depicted fixing solely $V A_= 10$.

Third, although the two sparse approaches under comparison attain equally sparse graphs for $V_T = 1$, the SC-VAR yields better MSEs. Note that the chosen causal features are different for the two approaches. Moreover, as noted in Section 2.2, larger penalties for the Lasso imply more sparsity but stronger shrinkage, entailing a loss in its prediction power. Finally, as an illustration results for the SC-VAR fixing nothing but parameter V_A have been also depicted. We required a total of 10 non-zero coefficients for the whole graph, obtaining the same number of arrows as when requiring a maximum of 1 non-zero per feature.

Observe that the obtained graphs for processes with 50 and 90% density are similar to the outputs of SC-VAR solved fixing $V_T = 1$ for each node. However, for the process with 10% density the MSE is improved by encouraging extra-diagonal elements. Note that some features receive more than one arrow, although the level of sparsity of the whole graph is the same. In conclusion, the forecasting power can be improved while maintaining the same level of sparsity by allowing a pooling effect on the non-zeroes (i.e., by solving the SC-VAR with binding (2.8) constraint).

In order to test the performance of the SC-VAR for large databases, VAR(1) processes with 200 and 500 nodes were generated. Despite a time limit of 30 seconds per node was imposed for the SC-VAR for all the results presented in this chapter, this upper bound was only reached for these large simulated data sets. For the rest of databases sparse graphs were generally obtained in less than 2 seconds. More details about the computational times are given in next section. For 200 nodes, the Lasso and the greedy approach obtain a graph with exactly 2200 non-zeroes, while the output of the SC-VAR has 1671 non-zero coefficients. Although the SC-VAR solution is more sparse, its interpretability is gained by minor deterioration over the MSE of the greedy approach and an improvement over the Lasso's (the MSE equals 0.662 for the SC-VAR, 0.657 for the greedy and 0.849 for the Lasso). The behaviour is mainly analogous for VAR processes with 500 features: the SC-VAR yields an extremely more sparse solution than its competitors (3091 non-zeroes instead of 130000) with an improvement over the forecasting power (the MSE equals 0.003 for the SC-VAR, 0.004 for the greedy and 0.156 for the Lasso). From these experiments we can conclude that the choice of an $\epsilon > 0$ avoids reaching the upper bound over the non-zeroes, leading the SC-VAR to yield solutions that are much more sparse than its competitors' but with similar or better predictive quality.

Computational times.

Although it would be interesting to study how data parameters (N, n and T) and chosen parameters (such as V_T, V_S, V_{Sa}...etc.) affect computational times when solving the MINLP (P), carrying out such costly experiments are out of the scope of this work. However, some intuitions arise from our numerical experience. As the VAR consists

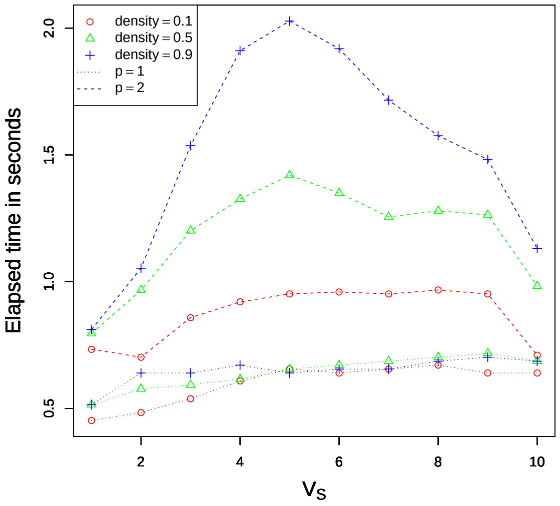


Figure 2.4: Median elapsed times in seconds taken by the SC-VAR to be solved for different requirements of the sparsity in VAR(1) and VAR(2) processes with various levels of density.

of $N^2n + N$ coefficients, the computational times are expected to increase specially with N and n, and we have experienced so in our experiments. On the other hand, we found that the SC-VAR is not that sensitive to the length of the time series T. In order to illustrate the effect of the sparsity requirements of the SC-VAR over the computational cost, consider Figure 2.4, where the median elapsed time in seconds has been depicted against the sparsity of the output, measured in terms of V_S. For the sake of clarity, only the times obtained for VAR(1) and VAR(2) processes with 10, 50 and 90% densities are depicted. Some comments are in order here. First, Figure 2.4 supports the intuition about computational times increasing with n. Second, although there is no much difference in the computational times for VAR(1) processes, it seems that for VAR(2) processes the computational times increase with the density of the true graph. Third, for $n = 2$ it is clear that the SC-VAR computational burden is consistently lower when either highly sparse or dense graphs are sought. To conclude, for $n = 1$ the SC-VAR takes to solve around 0.6 seconds on a PC Intel$^{®}$ Core^{TM} Quad CPU 4GB RAM, while for $n = 2$ the behavior is less consistent, but it usually obtains highly sparse graphs ($V_S \leq 2$) in less than 1 second.

2.4.3 Real data sets

In this section the performance of the three competing approaches is compared in two real databases, whose main features are summarized in Table 2.2. The series have not only been normalized but they have also undergone the Augmented Dickey-Fuller test for stationarity, implemented in the function `adf.test()` of the `R-cran`'s package `tseries`. The order n in the autoregressive model was unknown and hence chosen from $n = \{1, 2, ..., 7\}$ by the Schwartz criterion, implemented on the function `VARselect()`, available in the `R-cran`'s package `vars`.

Abbreviation	Name	N	T_{train}	T_{test}	n	Reference
Google flu	Google Flu Trends	48	221	220	1	Davis et al. (2012)
Air pollution	Concentration levels of air pollutants	5	4185	4185	4	Davis et al. (2012)

Table 2.2: Summary of real databases used for numerical illustrations

The main results for both databases are depicted in Figure 2.5. Panel A shows the results obtained for the considered Lasso approach and the SC-VAR in terms of MSE, which is represented against the upper-bound on the number of causal features per node V_S or the number of non-zero entries per node V_T. Panel B depicts some solutions for the methods under comparison. In the case of the Google Flu database, those solutions are depicted as heatmaps for the VAR, Lasso and SC-VAR. The color represents the sign of the coefficients θ^i_{jk} (blue for negative, red for positive) and the intensity is related to the magnitude of such coefficients. For the airpollution database Panel B of Figure 2.5 depicts, through directed graphs, the solutions of the VAR, and a couple of solutions for the Group Lasso and the SC-VAR. In such graphs color blue is associated with a 4-lag dependency. For the sake of abbreviation, only one or two of the most sparse solutions of the sparse approaches are included.

Google flu database.

This database is derived by the 45 Google user search terms that are considered to be indicative of influenza activity in the U.S. The sample is measured weekly from the beginning of the year 2006 until the week of June 6, 2014. According to the Centers for Disease Control and Surveillance (CDC) the probability that a patient query is related to influenza-like-illnes is closely related to the data in the Google flu database. From the 51 considered regions (50 states and the District of Columbia), North and South Dakota as well as Wyoming have been removed due to missing data.

From Panel A of Figure 2.5 some conclusions become apparent. First, the SC-VAR can considerably reduce the density of the matrix $\mathbf{A}$, containing the coefficients of the VAR, with an improvement of the MSE. Although one may think that fewer non-zero entries in $\mathbf{A}$ would lead to better MSE, increasing the freedom of the model may lead to overfitting. For instance, see that the SC-VAR solutions for $V_T \leq 15$ or Lasso solutions for $V_T \geq 3$ report a MSE smaller than 1; i.e., the prediction power of those sparse graphs is better than that of the solution provided by the VAR. Finally, it can be observed that although Lasso outperforms the SC-VAR for $V_T \geq 6$, it can attain extremely worse

MSE when highly sparse graphs are sought.

In the heatmaps of Panel B of Figure 2.5, the names of the states have been replaced by their abbreviations. Although Lasso provides a much more sparse solution than the VAR and enhance the visualization of potential causalities of the flu for the different regions of the US, the price of gaining such a level of sparsity is a 192% deterioration over its MSE. Nevertheless, the SC-VAR solution for $V_T = 1$ considerably improves the VAR's forecasting capacity (it attains a MSE with a 38% improvement over the VAR) while also allowing for an easier interpretation thanks to its sparsity.

Also note that, since the absolute value of the coefficients must be larger than the threshold 0.2, the obtained heat map is sharper than that of the Lasso. It is also interesting to note that SC-VAR approach tends to strengthen diagonal elements; i.e., with the SC-VAR in most cases the chosen causal feature for a node is itself when only one non-zero is allowed. This behavior is not as evident with the Lasso.

Air pollution database.

The data consists of hourly records of the solar radiation intensity (R) and the levels of four air pollutants, namely CO, NO, NO_2 and O_3, measured in Azusa, California during the year 2006.

In Panel A of Figure 2.5 we observe that there exists a clear difference in the impact of the parameters V_S and V_{Sa} over the MSE: increasing the parameter V_{Sa} seems more efficient to reduce the prediction error than increasing V_S. Therefore it is advisable to treat both parameters separately, since they clearly represent different aspects of the sparsity. This is done in a natural way with our approach, but not with the Lasso. Furthermore, note that our approach seems to stabilize for $V_T \geq 6$, whatever the values of V_S and V_{Sa} are. The constant MSE seems to denote that no further coefficients are being added to the model. Also, some of these constant lines explain that the constraint involving V_T is redundant, as the $V_S \cdot V_{Sa} \leq V_T$.

The behaviour observed from both plots of Panel A of Figure 2.5 is analogous to that of the previous results: when highly sparse graphs are sought, the SC-VAR seems to be more appropriate as it attains a better MSE. We point out that when only one non-zero is allowed ($V_T = 1$), the SC-VAR yields a 16% worsening over the VAR, while the Group Lasso reports a 49% deterioration. Moreover, when the number of causal features wants to be limited, the differences between the prediction errors of the SC-VAR and Group Lasso approaches are roughly a 200% and 50% for one and two causal features, respectively. Note also that the MSE deterioration of the SC-VAR over the VAR is less than a 2% for $V_S \geq 2$.

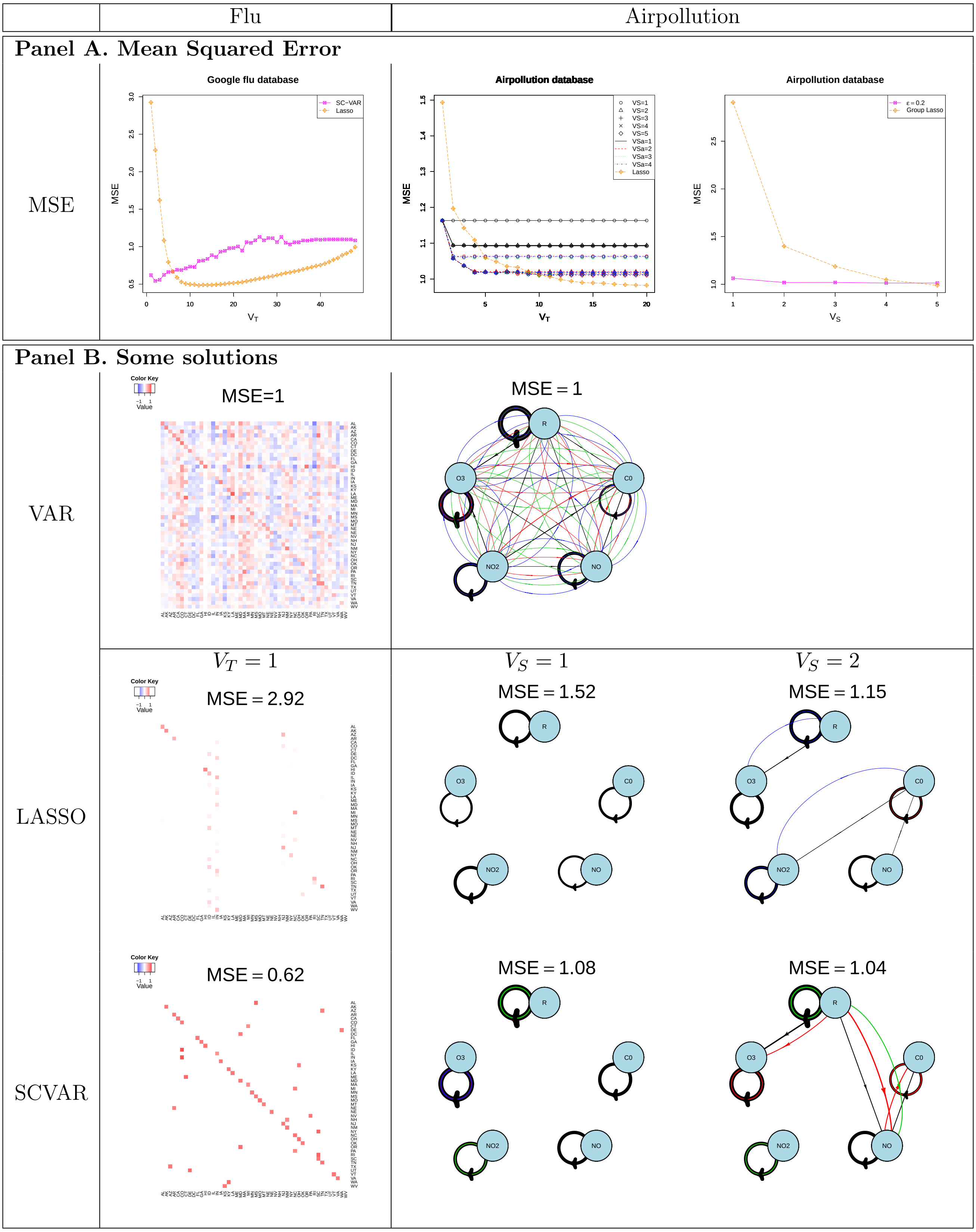


Figure 2.5: MSE (Panel A) and some heatmaps and graphs (Panel B) representing some solutions of the VAR, SC-VAR, and standard or Group Lasso for Flu and Airpollution databases. For the Flu database the SC-VAR and the standard Lasso are depicted for $V_T = 1$, while for the Airpollution dataset the SC-VAR and the Group Lasso outputs are allowing one ($V_S = 1$) or two ($V_S = 2$) causal features.

From Panel B of Figure 2.5 some comments arise. First, note that although the chosen relevant causal features are the same for the Group Lasso and SC-VAR when $V_S = 1$, the Group Lasso implies a 52% worsening over the VAR's prediction error, while the SC-VAR deteriorates it in a 8%. Second, the unraveled potential causal features obtained for each method are different when $V_S = 2$. The choice of the SC-VAR provides a much more sparse graph than the VAR while obtaining 11% less deterioration than its sparse counterpart.

In order to assess the stability of the SC-VAR we have solved the SC-VAR in the airpollution database for a grid of ϵ with V_T fixed, and a grid of V_T for two different values of ϵ fixed. The performance is illustrated in Figure 2.6, where Panel A depicts the MSE of the outputs and a heatmap representing the path of solutions for a grid of ϵ with $V_T = 10$ fixed, and Panel B represents, also using heatmaps, the path of solutions for a grid of V_T with $\epsilon = 0$ and 0.2 fixed. Note that the stability of the MSE for different values of V_T in the airpollution database was already analyzed when commenting Figure 2.5.

From Panel A of Figure 2.6 various conclusions emerge. First, the MSE tends to deteriorate when ϵ increases. However, note that an improvement of the predictive quality is attained for $\epsilon = 0.07$. From what it has been commented in the simulation study and real databases, it seems that constraint (2.10) is an useful tool to avoid overfitting when the database is truly dense, and/or to attain sparsity. Second, from the right plot of Panel A it can be deduced that the sparsity of the output increases as the value of ϵ does. Although the path of solutions is quite stable, note that as mentioned previously (see comments for Figure 2.5) sometimes it is not optimal to gradually drop the variables with the smallest coefficients to attain sparsity, but often it is better to let a new feature entering the model. For instance, look at the bottom rows for $\epsilon = 0.12$.

From Panel B of Figure 2.6 it can be deduced that the SC-VAR is highly stable when varying the parameter V_T. However, sometimes variables can leave the model ($V_T = 3, 4$ for $\epsilon = 0.2$). Also, it is interesting to observe how choosing an $\epsilon > 0$ stabilizes the process by avoiding adding features to the model.

2.4.4 Tuning of the parameters

In this section we comment the performance of the methods after applying the parameter tuning proposed at the end of Section 2.4.1. The results in terms of MSE and number of non-zeroes of the graph (denoted by NZ) are shown in Table 2.3. Specifically, Panel A displays the results for the simulated data sets, while Panel B does for the real databases.

Panel A of Table 2.3 shows that for simulated VAR(1) all methods are equivalent in

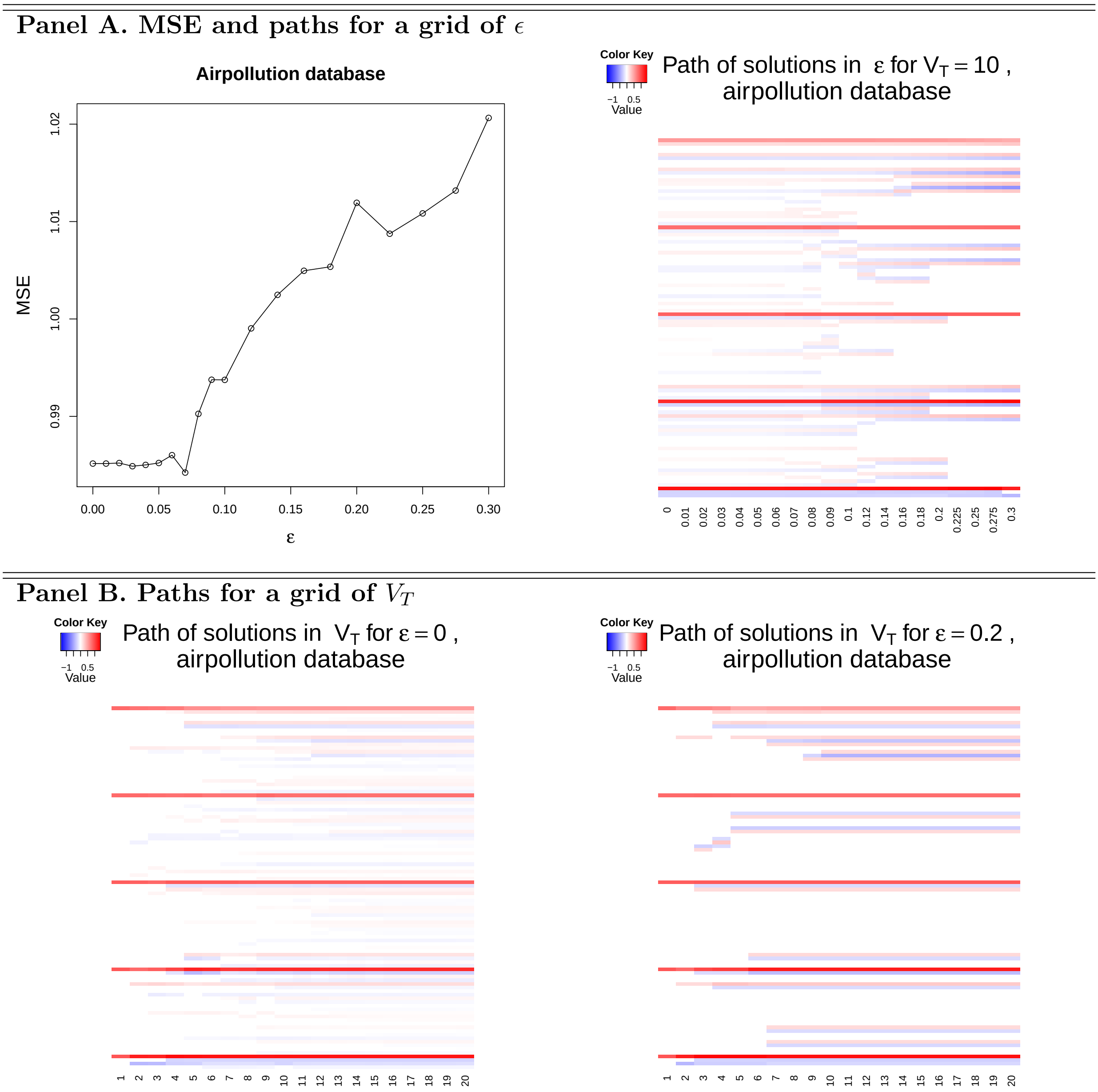


Figure 2.6: Heatmap representing the paths of solution for a fixed $V_T = 10$ and different values of ϵ together with its MSEs (right and left plots of Panel A, respectively), and heatmaps for different V_T with fixed $\epsilon = 0, 0.2$ (left and right plots of Panel B, respectively).

terms of MSE, but the most sparse solutions are obtained for the SC-VAR. Nevertheless, for VAR(2) processes the Adaptative Lasso is outperformed by its competitors, which also attain more sparse solutions. Nevertheless, the sparsity of the greedy becomes more similar to that of the Lasso as the density of the graph increases, while the SC-VAR is always a lot more sparse. In particular, although for simulated VAR(2) data with

90% of density the SC-VAR prediction power is slightly worse to that of the greedy, it manages to reduce the number of non-zero coefficients in nearly 100.

From Panel B of Table 2.3 it can be observed that for Flu database the greedy performs poorly in terms of forecasting power. Although the Adaptative Lasso improves its MSE, it attains a solution that is a lot more dense. Nevertheless, the SC-VAR yields similar MSE to the Lasso while obtaining a more sparse output. For the airpollution database the Adaptative Lasso performs poorly, although its output is the most sparse. Even though the greedy improves its predictive quality, the SC-VAR performs similarly with a more sparse solution.

Summarizing, from Table 2.3 we can conclude that the SC-VAR attains either a similar or better MSE than the rest of methods but with a more sparse output.

		Adapt. Lasso		SC-VAR		Greedy	
		MSE	NZ	MSE	NZ	MSE	NZ
Panel A. Simulated datasets							
	10% density	1.00	48	1.00	17	1.00	29
$n = 1$	50% density	1.01	70	1.02	35	1.01	53
	90% density	1.02	84	1.04	50	1.01	75
	10% density	1.00	90	0.76	26	0.76	62
$n = 2$	50% density	1.02	147	0.77	58	0.75	132
	90% density	1.02	176	0.74	76	0.70	172
Panel B. Real datasets							
	Flu	0.51	652	0.54	190	1.47	317
	Airpollution	1.21	49	1.02	56	1.01	84

Table 2.3: Results after tuning the parameters for the adaptative lasso, SC-VAR with $\epsilon = 0.2$ and greedy approaches in real and simulated databases. Normalized median MSE and median number of non-zeroes are displayed for the simulated databases.

2.5 Concluding remarks and extensions

In this chapter a sparse vector autoregressive model, the SC-VAR, that allows the user to control the sparsity of the output from various perspectives has been introduced. The model's sparsity is expressed in terms of different parameters, such as the number of total non-zero entries per series, the number of features involved or the number of periods chosen per feature. The method is expressed as an optimization problem, solvable by standard numerical optimization software.

The ability of the proposed approach to unravel potential causalities and, in many cases, to improve the fit, has been tested in simulated multivariate time series as well as in two real data bases, referred in the existent literature. It is concluded from the experiments that (i) the proposed approach is able to yield very sparse solutions either improving or without significantly increasing the VAR's forecasting error, (ii) the SC-VAR usually yields better MSEs than the greedy method and Lasso approaches when highly sparse graphs are sought, leading to a much better visualization of the process dependencies, as e.g. depicted in Figure 2.3, and (iii) the parameters considered to measure the sparsity play different roles, thus it might not be advisable to aggregate them into one single measure, as done by other sparse methods. In particular, it seems that the parameter ϵ plays an important role, as it helps avoiding overfitting and attaining sparsity with a minor deterioration of the prediction power.

In the future we plan to address grouping effects by considering jointly different features in constraints of type (2.10). The idea is to incorporate the modeller knowledge about groups of variables that may have little individual effect but whose impact may be significant when considered altogether. Hence, the user may want to group the features in different sets $J_1, ..., J_S$ and consider constraints of the type $\sum_{j \in J_s} |\theta^i_{jk}| \leq \epsilon^i_s$ for all $s = 1, .., S$. It is our aim to further model other grouping constraints and to asses the influence over the performance of the approach.

Appendix: AMPL code

```
#PARAMETERS OF THE PROCESS
param T;                     #Series' length
param N;                     #Number of series
param p;                     #Order of the process

#SETS OF INDEXES
set Nseries:=1..N;
set time:=1..T;
set tallowed:=(p+1)..T;
set periods:=1..p;

#THE MULTIVARIATE TIME SERIES
param X {j in Nseries,t in time};       #Matrix of time series
#PARAMETERS OF THE SC-VAR
param VA;                               #Upper-bound on the total number of non-zeroes
param VT {i in Nseries};                #Upper-bounds on total non-zeroes for feature i
param Vs {i in Nseries};                #Upper-bounds on number of causal features for i
param Vsa {i in Nseries};               #Upper-bounds on number of non-zeroes
                                        #of each causal feature
param eps {i in Nseries,j in Nseries};  #Thresholds
param M;                                #Upper bound for alpha

#VARIABLES
var c {i in Nseries};
var alpha {i in Nseries,j in Nseries,k in periods};
var gamma {i in Nseries,j in Nseries}, binary;
var nupos {i in Nseries,j in Nseries,k in periodos}, binary;
var nuneg {i in Nseries,j in Nseries,k in periodos}, binary;
#OBJECTIVE FUNCTION

minimize fun:  (1/T)*sum{t in tallowed,i in Nseries} (X [i,t]
               -c[i]-sum{j in Nseries, k in periods} (alpha[i,j,k]*X[j,(t-k)]))^2;
```

```
#CONSTRAINTS
subject to totalNZ:
             sum{i in Nseries,j in Nseries,k in periods} (nupos[i,j,k]+nuneg[i,j,k])<=VA;
subject to totalNZi{i in Nseries}:
             sum{j in Nseries,k in periods} (nupos[i,j,k]+nuneg[i,j,k])<=VT[i];
subject to nocausalnoalpha {i in Nseries,j in Nseries,k in periods}:
             nupos[i,j,k]+nuneg[i,j,k]<=gamma[i,j];
subject to seriesNZ {i in Nseries}:
             sum{j in Nseries} gamma[i,j]<=Vs[i];
subject to elemserieNZ {i in Nseries,j in Nseries}:
             sum{ k in periodos} (nupos[i,j,k]+nuneg[i,j,k])<=Vsa[i]*gamma[i,j];
subject to epspos {i in Nseries,j in Nseries,k in periods}:
             alpha[i,j,k]>=eps[i,j]*nupos[i,j,k]-M*nuneg[i,j,k];
subject to epsneg {i in Nseries,j in Nseries,k in periods}:
             alpha[i,j,k]<=-eps[i,j]*nuneg[i,j,k]+M*nupos[i,j,k];
```

Chapter 3

Enhancing Sparsity by Tightening Linear Regression Models

As a preprocessing step, prior to estimating parameters in linear regression models (1.5), a common practice consists of removing highly correlated variables or unimportant features. In this chapter we propose tools based on Mathematical Optimization which improve sparsity in linear regression, avoiding these manual steps. This will be addressed by modelling sparsity, significance and collinearity constraints, and integrating them into the preferred estimation procedure. In particular, the so-obtained tightened sparse models will become MINLPs. The numerical experiments carried out on real and simulated datasets support this statement and also show that tightening the search space of some standard linear regression models may enhance the interpretability of the outcomes with competitive predictive quality.

3.1 Introduction

In this chapter we propose to add constraints to the existing linear regression approaches with the aim of avoiding the manual steps to remove unimportant or highly correlated variables, as described in Chapter 1, while enhancing sparsity. This will be addressed by modelling the so-called sparsity, significance and collinearity constrains and integrating them into the preferred estimation procedure. Formally, assume the estimation method, selected by the user to estimate the parameters β_0, $\boldsymbol{\beta}$ of the linear regression model (1.5), is derived from solving an optimization problem of the form:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & f(\boldsymbol{\beta}) \\ \text{s.t} \quad & \boldsymbol{\beta} \in \mathcal{B} \end{aligned} \tag{3.1}$$

where $\mathcal{B}$ denotes the feasible region. In this chapter we propose to tighten the feasible region for the $\boldsymbol{\beta}$ in order to enhance the performance and interpretability of the solutions of Problem (3.1). For instance, the OLS and penalized regression may be good representatives of estimation methods as in (3.1). For these cases, $\mathcal{B} = \mathbb{R}^N$; in fact, the search spaces of the most commonly used estimation methods are usually unrestricted. Hence, tightening approaches have been considered in the literature in order to improve the performance of linear regression problems. For example, in the recent paper Bertsimas and King (2015) some sparsity constraints were added to the Lasso objective. In this chapter, it is our aim to generalize this kind of approaches by replacing any estimation procedure (3.1) by:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & f(\boldsymbol{\beta}) \\ \text{s.t} \quad & \boldsymbol{\beta} \in \mathcal{B} \cap \mathcal{S} \end{aligned} \tag{3.2}$$

where $\mathcal{S}$ will gather the proposed constraints. In particular, the so-obtained sparse models (3.2) will be MINLPs. Although not frequently used amongst the statistical community, MINLPs have recently been used to tackle sparse models in a linear regression (Bertsimas and King, 2015) and a time series (Carrizosa et al., 2016b) context. Both cases were supported by the recent improvements in computational times that MINLPs have enjoyed, becoming a useful tool to address statistical problems in a tractable and versatile manner. We will show by means of different tests the advantages of tightening (3.1) via specific constraints.

The chapter is structured as follows. Next section is devoted to motivate and model different types of constraints which define the set $\mathcal{S}$ in (3.2); they will be called sparsity, significance and collinearity constraints. The numerical experiments are carried out in Section 3.3, where the methods under comparison and the design of experiments are also discussed. The last section is devoted to concluding remarks and extensions.

3.2 New constrains for linear regression approaches

In this section we define the constraints to be added to a classic (possibly sparse) linear regression model (3.1), in order to enhance the sparsity of the outcome and to allow only significant features to be considered in the model. We also model some constraints aiming to avoid some misleading interpretations that are common in the presence of data multicollinearity. The idea is that the user should feel free to add any of these constraints, when compatible, to her selected regression approach.

3.2.1 Sparsity

One way to limit the number of non-zero coefficients in a regression model is by using the so called ℓ_0-*norm*, which is no other thing than a cardinality function $||\boldsymbol{\beta}||_0 = \#(j : \beta_j \neq 0)$. This function provides a valid tool to model a constraint which can be added to some regression approaches (for example, the basic OLS objective) to grant control over the sparsity of the outcome:

$$||\boldsymbol{\beta}||_0 \leq V_T, \tag{3.3}$$

where V_T denotes an upper bound on the number of non-zeroes of the solution. In Mathematical Optimization the formulation of this constraint is tackled defining the integer variables:

$$\gamma_j \;=\; \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{if } \beta_j = 0. \end{cases} \tag{3.4}$$

Once these indicator variables are defined, it is straightforward to formulate a constraint that bounds the density of the solution as follows:

$$\sum_{j=1}^{N} \gamma_j \leq V_T. \tag{3.5}$$

This constraint is equivalent to the ℓ_0-norm constraint (3.3). The tractability of approaches which model sparsity constraints as in (3.5) is supported by the recent development of MINLP solvers (Bertsimas and King, 2015; Carrizosa et al., 2016b). Nevertheless, an ℓ_1-norm constitutes an even more tractable surrogate to attain sparsity in regression models. Indeed, adding the constraint

$$\sum_{j} |\beta_j| \leq t \tag{3.6}$$

performs a shrinkage over the coefficients, which reduces the variance of the estimates but eventually attains sparsity as a side effect. Note that this constraint can be passed to the objective function as a penalization, leading to the classic expression of the Lasso.

The tractability and the theoretical results that prove that the Lasso recovers the real underlying sparse structure under some conditions, has led the Lasso to become the benchmark sparse regression method. As it will be illustrated in Section 3.3, the Lasso improves the interpretability of the OLS, but a better trade off between sparsity and predictive power may be attained by tightening its feasible region with constraints such as (3.5).

3.2.2 Significance

In this section we model constraints that allow only *significant* variables to be represented in the model. Intuitively, *large* coefficients are identified with the importance of a feature when the data is normalized. For instance, best subset selection sets the smallest coefficients to zero and keeps the largest ones (LeBlanc and Tibshirani, 1998). This is also the idea underlying the Adaptative Lasso (Zou, 2006), which avoids highly shrinkage over the largest coefficients so as to allow them to pervive in the model. This is done by introducing weights that are related to the size of the estimated coefficients. For instance, Zou (2006) proposes weights that are inversely proportional to the magnitude of the OLS estimates. In this way, the largest coefficients on the OLS solution are subject to a smaller penalization and hence are more likely to be included in the model.

Following this reasoning we propose to establish a threshold of *importance* that a feature must be able to exceed to be allowed in the model. This approach is mainly different from the Lasso methods, which perform an shrinkage over the coefficients. We model such constraint as follows:

$$|\beta_j| \in \{0\} \cup [\epsilon, +\infty) \qquad j = 1, ..., N \tag{3.7}$$

where ϵ is called here the significance threshold, to be fixed by the user or to be tuned. Note that we define an unfeasibility region in $(0, \epsilon)$, trying to avoid shrinkage and forbid espureous coefficients in the solution. Constraint (3.7) can be linearly rewritten using the following binary variables:

$$\nu_j^+ = \begin{cases} 1 & \text{if } \beta_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\nu_j^- = \begin{cases} 1 & \text{if } \beta_j < 0 \\ 0 & \text{otherwise} \end{cases}$$

Indeed, constraint (3.7) is expressed via two linear constrains as follows:

$$\begin{aligned} \beta_j &\geq \epsilon\nu_j^+ - \nu_j^- M \quad && \forall j = 1, ..., N \\ \beta_j &\leq -\epsilon\nu_j^- + \nu_j^+ M && \forall j = 1, ..., N \end{aligned} \tag{3.8}$$

where M is a *large* constant. This big M, often appearing when modelling problems with integer variables, is large enough so it does not exclude reasonable values of the parameters β_j (see, e.g. Camm et al. (1990)). To the authors extent, this constraints were first modelled by Carrizosa et al. (2016b) in a sparse context, where they were used to discover potential causalities in multivariate time series.

Variables ν_j^+, ν_j^- are linked with γ_j, defined in Equation (3.4), as follows:

$$\gamma_j = \nu_j^+ + \nu_j^-,$$

and thus (3.5) can also be written as:

$$\sum_{j=1}^{N} \gamma_j \leq V_T.$$

Despite the usefulness of the above mentioned sparse approaches, the association between the importance and the magnitude of the estimated coefficient may sometimes be misleading, since coefficients obtained by OLS or Lasso approaches may be inflated in the presence of collinearity (Montgomery et al., 2015). This issue will be treated in Section 3.2.3, and therefore we propose to combine constraints (3.8) with the collinearity constraints that will be described in the next section.

3.2.3 Collinearity

The presence of correlated variables in the data is demonstrated to lead to the appearance of undesired consequences, such as the unstability of the solutions and the inflation of the estimated coefficients, which may lead to misleading interpretations, or misrepresentative results of significance tests (Watson and Teelucksingh, 2002; Hesterberg et al., 2008; Dormann et al., 2013; Montgomery et al., 2015).

As an illustration, consider the example given by Hesterberg et al. (2008) with the diabetes database. This database consists of the measures of 10 variables (namely age, sex, body mass index, average blood pressure and six different blood serums) on 442 patients. Figure 3.1 depicts the path of solutions of the Lasso for this database; that is to say, the estimates of the coefficients $\boldsymbol{\beta}$ obtained for different values of the penalty t in constraint (3.6) are depicted against their level of sparsity.

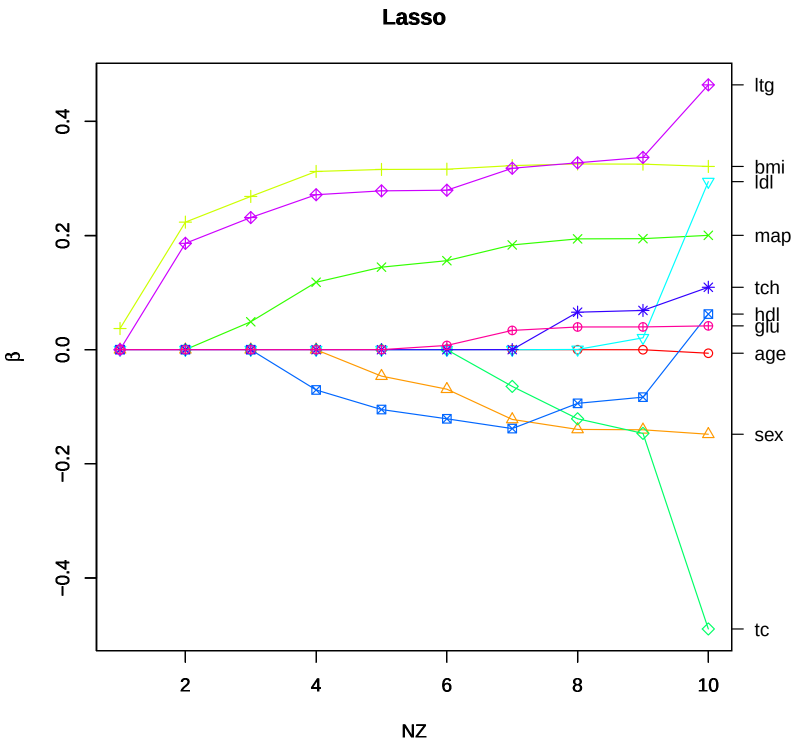


Figure 3.1: Path of solutions of the Lasso for the diabetes database. The size of the coefficients $\boldsymbol{\beta}$ are depicted against the number of non zeroes of the solution (NZ).

As noted by Hesterberg et al. (2008), features *tc* and *ldl*, which have a correlation of 0.89, enter the model relatively late and their coefficients grow quickly in opposite directions. Another bizarre behaviour is observed for variables *hdl* and *tch*, which show a correlation of −0.73. In this case, the coefficient of *hdl* (who had entered the model far before *tch*), changes its tendency in the presence of *tch*. This unstable behaviour leads the coefficient associated with *hdl* to change its sign at the end. In particular, this ultimate change of sign may cause a misinterpretation of the process: at some point *hdl* is assumed to have a positive impact over the evolution of the diabetes, while for another level of sparsity it is assumed to have the opposite effect.

The standard approach to avoid such negative consequences often consists of manually removing some correlated variables after carrying out some tests (Chatterjee and Hadi, 2015). Further remedies consist of applying methods that decorrelate the data (see, for example, Cao et al. (2010); Massy (1965)). Nevertheless, the later approaches imply the transformation of the variables and thus complicate the interpretation of the final models with respect to the original features.

A novel approach was undertaken by Bertsimas and King (2015), who proposed to integrate the manual step of removing correlated variables into their optimization problem. They did so by adding the constraints:

$$\gamma_i + \gamma_j \leq 1 \quad \forall (i,j) \in \Omega_\eta, \tag{3.9}$$

where $\Omega_\eta = \{(i,j) : \rho_{ij} \geq \eta\}$ is the set of pairs of features considered to be highly correlated, and the variables γ_i, γ_j are defined in (3.4). That is to say, a pair of features whose correlation exceed a threshold η are considered not to give different information about the response variable $\mathbf{y}$. These constraints may be too restrictive since if $(i,j) \in \Omega_\eta$, either feature i or j is excluded from the model. Hence, for a large η collinearity issues might not be avoided, while for smaller η a large number of variables might be excluded.

In contrast, we would like to propose less restrictive constraints in order to allow two correlated variables in the model at the same time but forbid misleading interpretations and misrepresentative coefficients. In other words, our aim is to avoid unstable behaviours while allowing the model to include two correlated variables only if they both contribute separately to the explanation of $\mathbf{y}$. In this way a clearer interpretation of the true impact of the features over the variable under study may be obtained. Therefore, we propose to model constraints that avoid the *compensation* of coefficients for correlated variables. Under the light of the previous example, these are the requirements we want to gather when modelling the pairwise collinearity constraints:

1. The weight of two correlated variables should not be spread out between them. If two features do not give information important enough by themselves, at least one should be removed.

2. The coefficients of two features that are highly positively correlated must have the same sign.

3. The coefficients of two features that are highly negatively correlated must have

opposite signs.

Since any coefficient β_j must be larger than ϵ when imposing constraints (3.8), intuitively the first point will be alleviated by considering the constraints (3.8) for ϵ sufficiently large. The rest of requirements can be easily formulated as constraints using the indicator variables defined in Section 3.2.2:

$$\nu_i^+ + \nu_j^- \leq 1 \quad \forall (i,j) \in \Omega_\alpha^+ \tag{3.10}$$

$$\nu_i^- + \nu_j^+ \leq 1 \quad \forall (i,j) \in \Omega_\alpha^+ \tag{3.11}$$

$$\nu_i^+ + \nu_j^+ \leq 1 \quad \forall (i,j) \in \Omega_\alpha^- \tag{3.12}$$

$$\nu_i^- + \nu_j^- \leq 1 \quad \forall (i,j) \in \Omega_\alpha^- \tag{3.13}$$

where Ω_α^+ and Ω_α^- are the sets of pairs of features that are moderately or highly correlated, expressed as $\Omega_\alpha^+ = \{(i,j) : \rho_{ij} \geq \alpha\}$ and $\Omega_\alpha^- = \{(i,j) : \rho_{ij} \leq -\alpha\}$. That is to say, constraints (3.10)-(3.11) mean that, if two variables i, j are highly positively correlated (i.e. $(i,j) \in \Omega_\alpha^+$), then we do not allow one of the coefficients to be positive and the other negative. Similarly, constraints (3.10)-(3.11) imply that, if two variables i, j are highly negative correlated (i.e. $(i,j) \in \Omega_\alpha^-$), we forbid their coefficients to be both positive or both strictly negative.

In order to illustrate the impact of these constraints over the Lasso approach, Figure 3.2 depicts the path of solutions of the tightened Lasso for the diabetes database, as was analogously done in Figure 3.1. As it can be observed, the use of constraints (3.10)-(3.13) to tighten the feasible region of the Lasso avoids the unstable behaviours noted above, easing the interpretation of the impact of the predictors over the development of the illness. In particular, the use of the proposed constraints yields a more sparse solution, where the variables *age, hdl, tch* and *ldl* are omitted.

To the authors extent, this kind of correlation constraints has never been modelled before, except for the above mentioned more strict constraint (3.9). In comparison, the inclusion of the new collinearity constraints (3.10)-(3.13) would imply a larger number of constraints and variables, yielding a more complex model. However, it will be seen in the next section that a good performance is attained without increasing the time limit.

3.3 Numerical illustrations

In this section we describe and undertake the numerical experiments performed to compare three benchmark estimation methods in linear regression against their tightened versions derived from reducing the search of the coefficients $\boldsymbol{\beta} \in \mathcal{B}$ to the set $\mathcal{B} \cap \mathcal{S}$, where $\mathcal{S}$ is defined through the constraints proposed in the previous section. Specifi-

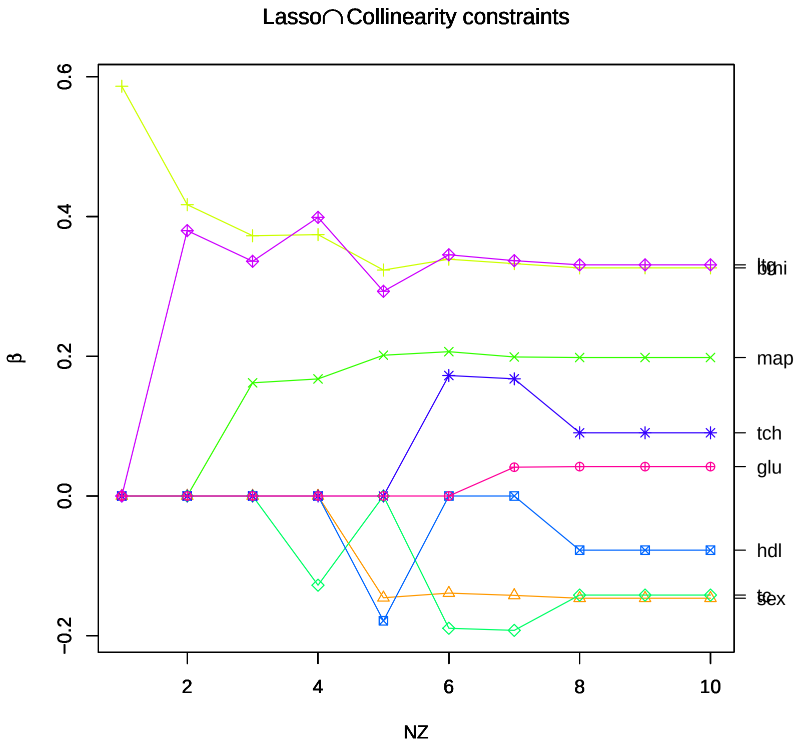


Figure 3.2: Path of solutions of the tightened Lasso for the diabetes database.

cally, in the next section we introduce the linear regression methods used in the tests, while Section 3.3.2 outlines the design of the experiments. Sections 3.3.3 and 3.3.4 show the results for the real and simulated databases, respectively.

3.3.1 Methods under comparison

Now we will introduce the linear regression models to which we will add the constraints proposed in the previous section, defining the so-called tightened region $\mathcal{B} \cap \mathcal{S}$. The aim is to assess their impact over the solutions of the original problems; i.e. to analyze the differences in the performance between Problems (3.1) and (3.2). In particular, we will consider three different sets of constraints $\mathcal{S}$. In order to test separately the effect of the significance and collinearity constraints, the tightening sets $\mathcal{S}_1 = \{(3.5), (3.8)\}$ and $\mathcal{S}_2 = \{(3.5), (3.10) - (3.13)\}$ are considered. Finally, the set $\mathcal{S}_3 = \{(3.5), (3.8), (3.10) - (3.13)\}$ is also used to further tighten the set of constraints $\mathcal{B}$ of the original problem. Note that the sparsity constraint (3.5) is always included in the tightened versions since it is the most simple and interpretable of all the constraints and has already proven to enhance the sparsity of the outcome in linear regression (Bertsimas and King, 2015). Also, the integrality constraints $\nu_j^+, \nu_j^- \in \{0, 1\}$ need to be included in the set $\mathcal{S}$ when such a kind of variables does not appear in the original problem.

Below, we explicitly formulate the benchmark estimation methods (Problem (3.1)) and their tightened versions (Problem (3.2)) used for comparison in our numerical results. In order to conserve space, only the problems with the tightening set $\mathcal{S}_3$ are shown. To use $\mathcal{S}_1$ or $\mathcal{S}_2$ as tightening sets instead some of the constraints should be removed.

Ordinary Least Squares

Following the notation used to define Problem (3.1), the Ordinary Least Squares estimation method can be written as:

$$\begin{aligned} f(\boldsymbol{\beta}) &= \sum_{k=1}^{K}\left(Y_k - \beta_0 - \sum_{j=1}^{N}\beta_j X_k^j\right)^2 \\ \mathcal{B} &= \mathbb{R}^N \end{aligned} \tag{3.14}$$

OLS is the most straightforward procedure to derive solutions for $\beta_0, \boldsymbol{\beta}$ in linear regression problems. It is extremely tractable, since Problem (3.14) is a unconstrained quadratic program. The restricted approach with tightening set $\mathcal{S}_3$ can explicitly be written as:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \sum_{k=1}^{K}\left(Y_k - \beta_0 - \sum_{j=1}^{N}\beta_j X_k^j\right)^2 \\ \text{s.t} \quad & \begin{cases} \sum_{j=1}^{N}(\nu_j^+ + \nu_j^-) \le V_T & \\ \beta_j \ge \epsilon\nu_j^+ - \nu_j^- M & \forall j = 1, ..., N \\ \beta_j \le -\epsilon\nu_j^- + \nu_j^+ M & \forall j = 1, ..., N \\ \nu_i^+ + \nu_j^- \le 1 & \forall (i,j) \in \Omega_\alpha^+ \\ \nu_i^- + \nu_j^+ \le 1 & \forall (i,j) \in \Omega_\alpha^+ \\ \nu_i^+ + \nu_j^+ \le 1 & \forall (i,j) \in \Omega_\alpha^- \\ \nu_i^- + \nu_j^- \le 1 & \forall (i,j) \in \Omega_\alpha^- \\ \nu_j^+, \nu_j^- \in \{0,1\} & \forall j = 1, ..., N \end{cases} \end{aligned} \tag{3.15}$$

Lasso

The Lasso can be formulated as an optimization problem as in (3.1) as follows:

$$
\begin{aligned}
&f(\boldsymbol{\beta}) = \sum_{k=1}^{K}\left(Y_k - \beta_0 - \sum_{j=1}^{N}\beta_j X_k^j\right)^2 + \lambda\|\boldsymbol{\beta}\|_1 \\
&\mathcal{B} = \mathbb{R}^N
\end{aligned}
\tag{3.16}
$$

The sparsity of the solution attained by the Lasso increases as its regularization parameter $\lambda \in \mathbb{R}^+$ does. The critical values of λ are easily computed using any implementation of the LAR algorithm in standard statistical softwares. In particular, in this chapter the Lasso set of solutions was obtained by using the `lars()` function of `R-cran` package `lars` (Hastie and Efron, 2013). Specifically, the expression of the most tightened version of the Lasso that will be solved in our numerical experiments is:

$$
\begin{aligned}
\min_{\boldsymbol{\beta}} \quad & \sum_{k=1}^{K}\left(Y_k - \beta_0 - \sum_{j=1}^{N}\beta_j X_k^j\right)^2 + \lambda\|\boldsymbol{\beta}\|_1 \\
\text{s.t} \quad & \left\{
\begin{array}{ll}
\sum_{j=1}^{N}(\nu_j^+ + \nu_j^-) \leq V_T & \\
\beta_j \geq \epsilon\nu_j^+ - \nu_j^- M & \forall j = 1, ..., N \\
\beta_j \leq -\epsilon\nu_j^- + \nu_j^+ M & \forall j = 1, ..., N \\
\nu_i^+ + \nu_j^- \leq 1 & \forall (i,j) \in \Omega_\alpha^+ \\
\nu_i^- + \nu_j^+ \leq 1 & \forall (i,j) \in \Omega_\alpha^+ \\
\nu_i^+ + \nu_j^+ \leq 1 & \forall (i,j) \in \Omega_\alpha^- \\
\nu_i^- + \nu_j^- \leq 1 & \forall (i,j) \in \Omega_\alpha^- \\
\nu_j^+, \nu_j^- \in \{0,1\} & \forall j = 1, ..., N
\end{array}
\right.
\end{aligned}
\tag{3.17}
$$

The optimization problem of Bertsimas and King (2015)

Now we introduce the following optimization problem:

$$
\begin{aligned}
&f(\boldsymbol{\beta}) = \sum_{k=1}^{K}\left(Y_k - \beta_0 - \sum_{j=1}^{N}\beta_j X_k^j\right)^2 + \lambda\|\boldsymbol{\beta}\|_1 \\
&\mathcal{B} = \{(3.5), (3.9), \gamma_j \in \{0,1\} \;\; \forall j = 1, ..., N\}
\end{aligned}
\tag{3.18}
$$

Bertsimas and King (2015) proposed an algorithm to obtain interpretable solutions in linear regression. As part of such an algorithm, the resolution of a more complex version of Problem (3.18), which may include constraints that allow to incorporate the modellers prior knowledge or avoid undesirable results, can be undertaken more than once. In our numerical results we restrict ourselves to the more simple version Problem (3.18).

Although Problem (3.18) already includes the sparsity constraint (3.5), it lacks of the significance constraints (3.8). Moreover the correlation constraint (3.9) is different to the ones proposed here, as discussed in Section 3.2.3. Therefore a different performance of Problem (3.18) is expected by considering the tightened feasible region $\mathcal{B}\cap\mathcal{S}$ instead. In particular, the tightened version with search space $\mathcal{B}\cap\mathcal{S}_3$ can be explicitly formulated as:

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \quad \sum_{k=1}^{K}\left(Y_k-\beta_0-\sum_{j=1}^{N}\beta_j X_k^j\right)^2+\lambda\|\boldsymbol{\beta}\|_1\\
&\text{s.t}\quad \begin{cases}
\sum_{j=1}^{N}(\nu_j^+ + \nu_j^-)\le V_T & \\
\beta_j \ge \epsilon\nu_j^+ - \nu_j^- M & \forall j=1,...,N\\
\beta_j \le -\epsilon\nu_j^- + \nu_j^+ M & \forall j=1,...,N\\
\nu_i^+ + \nu_j^- \le 1 & \forall (i,j)\in\Omega_\alpha^+\\
\nu_i^- + \nu_j^+ \le 1 & \forall (i,j)\in\Omega_\alpha^+\\
\nu_i^+ + \nu_j^+ \le 1 & \forall (i,j)\in\Omega_\alpha^-\\
\nu_i^- + \nu_j^- \le 1 & \forall (i,j)\in\Omega_\alpha^-\\
\nu_i^+ + \nu_i^- + \nu_j^+ + \nu_j^- \le 1 & \forall (i,j)\in\Omega_\eta\\
\nu_j^+,\nu_j^- \in\{0,1\} & \forall j=1,...,N
\end{cases}
\end{aligned} \tag{3.19}$$

As in Bertsimas and King (2015) the maximum pairwise correlation allowed is $\eta = 0.8$; that is to say, the set Ω_η in (3.9) is defined here as $\Omega_\eta = \{(i,j) : \rho_{ij} \ge 0.8\}$. Further details about the choice of λ and V_T are given in the following section.

3.3.2 Design of experiments

In order to make a fair comparison against existing procedures, the experiments developed here closely follow those in Bertsimas and King (2015). First, the datasets are normalized and divided in train, test and validation (50%, 25% and 25% of the data, respectively). All the problems are solved in the train set, and the solution that minimizes the Mean Squared Error (MSE thereafter) in the test set is chosen. Two criteria are used to compare the methods, namely, the MSE and the sparsity. All the MSEs reported in this chapter correspond to the values obtained in the validation sets and are normalized by dividing by the OLS solution; that is to say, when any method attains a MSE greater than 1 their prediction capacity is estimated to be worse than that of the OLS, while for smaller values the predictive power has improved.

For the tightened MINLPs (3.2), the pairwise correlation considered to generate the sets Ω_{α}^{+} and Ω_{α}^{-} in constraints (3.10)-(3.13) is fixed to $\alpha = 0.6$. The significance parameter ϵ in constraints (3.7) is tuned by chosing amongst the ten values $\{0.05, 0.06, 0.08, 0.1, 0.125, 0.15, 0.175, 0.2, 0.25, 0.3\}$ so as to minimize the MSE in the test set. As done in Bertsimas and King (2015), when comparing against any method that has a Lasso objective, the grid of values of the parameter λ to be tuned for the tightened MINLP is logarithmically generated in the interval $(0, \lambda_{max}]$, where λ_{max} is the penalty provided by the LAR for which only one coefficient is non-zero. The parameter V_T on the sparsity constraint is chosen in $\{1, ..., N\}$. However, in order restrict the search only to likely values of V_T, a stopping criterion is imposed: when no more features are added to the model (i.e., when the constraint (3.5) becomes inactive), no larger values of V_T are considered.

Problems (3.1) and (3.2) are, at their worst, MINLPs with quadratic convex objective function and linear constraints. Unless otherwise specified, the optimization problems were solved using the standard `CPLEX 12.6.3.0` and were easily coded in the algebraic language `AMPL` (Fourer et al., 2002). A time limit of 20 seconds was imposed, although this limit was reached only for the largest datasets and in most cases cases the optimal solution was attained in a few seconds.

3.3.3 Real datasets

In this section we show the results obtained for some real datasets, which are easily reachable on internet and well referenced in the literature (Bertsimas and King, 2015). Further details about the specifications of the data sets and their sources are displayed in Table 3.1. The columns provide information about the name, number of observations (K), the number of predictors (N), and the data source.

	K	N	Source
cpu	105	6	Lichman (2016)
yacht	154	6	Lichman (2016)
whitewine	2499	11	Lichman (2016)
redwine	800	11	Lichman (2016)
golf2008	78	6	Winner (2016)
golf2009	73	11	Winner (2016)
compact	4096	21	Torgo (2016)

Table 3.1: Real data sets specifications and sources

The databases were randomly divided ten times in train, test and validation sets. The median MSE and number of non-zeroes are displayed in Table 3.2, where the first column of results corresponds to the normalized MSE and numer of non-zero coefficients (NZ) for the original Problem (3.1) (*Baseline*), and the second, third and fourth columns contain the results of Problem (3.2) with sparsity and significance constraints ($\mathcal{S}_1$), sparsity and collinearity ($\mathcal{S}_2$), and sparsity, significance and collinearity constraints ($\mathcal{S}_3$), respectively. Each panel of Table 3.2 shows the results of tightening the feasible regions of the considered methods, described in Section 3.3.1.

From Table 3.2 it can be concluded that reducing the search of the coefficients $\boldsymbol{\beta}$ by adding any of the constraints proposed in Section 3.2 always improves or maintains the sparsity of the solution. In particular, adding solely the significance constraints usually improves the sparsity while attaining a better or equivalent MSE. In contrast, the effect of the collinearity constraints yields an even more sparse solution, although sometimes performing slightly worse than the significance constraints in terms of predictive power. The combination of these constraints may yield even more sparse solutions than its counterparts, while improving the MSE of the solutions obtained searching in $\mathcal{B} \cap \mathcal{S}_2$.

More specifically, as it can be observed in Panel A, the density of the outputs is always substantially reduced for the OLS. The predictive power improves in 9 out of the 21 cases, up to a 35% in the cpu dataset. For the rest of the databases the predictive quality is similar to that of the OLS. The maximum deterioration of the MSE (4%) is found for the *compact* database with the tightening set $\mathcal{S}_3$; it is the price paid for gaining 12 zeroes in the output.

In Panel B it can be observed that the tightened Lasso always reduces considerably the density of the outputs except for *golf2008* database, where the number of non-zeroes is maintained. For instance, the number of non-zeroes is reduced in 11 and 4 for *compact* and *whitewine* databases. The MSE is slightly improved (around a 2%) in 8 out of the 21 cases, and slightly deteriorated or similar for the rest of the cases. An exception is the *cpu* database with collinearity constraints, where 1 zero is gained by paying a 16% deterioration over the MSE. Nevertheless, this worsening is considerably alleviated by also using the significance constraints.

In Panel C the performance of the new constraints is not as extreme, since Problem (3.18) already considered sparsity and an alternative type of collinearity constraints. However, adding the proposed new constraints has a positive effect over the outputs: the density is reduced in 12 out of the 21 cases with a similar or better MSE. For *cpu* database a much better predictive quality is obtained for the same level of sparsity when including solely the significance constraints.

	Baseline		$\mathcal{S}_1$		$\mathcal{S}_2$		$\mathcal{S}_3$	
	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ
Panel A: OLS								
cpu	1.00	6.0	0.65	4.5	0.72	4.0	0.67	4.0
yacht	1.00	6.0	1.00	1.0	1.00	1.0	1.00	1.0
whitewine	1.00	11.0	1.02	8.0	1.03	6.5	1.03	6.0
redwine	1.00	11.0	1.03	7.0	1.03	6.0	1.03	6.0
golf2008	1.00	6.0	0.69	3.0	0.69	3.0	0.69	3.0
golf2009	1.00	11.0	0.90	7.0	0.91	7.0	0.89	6.0
compact	1.00	21.0	1.02	13.0	1.04	11.0	1.04	9.0
Panel B: Lasso								
cpu	0.62	5.0	0.64	4.5	0.72	4.0	0.66	4.0
yacht	1.02	2.5	1.00	1.0	1.00	1.0	1.00	1.0
whitewine	1.02	10.0	1.02	8.0	1.03	6.5	1.03	6.0
redwine	1.05	9.5	1.03	7.0	1.03	6.0	1.03	6.0
golf2008	0.68	3.0	0.68	3.0	0.68	3.0	0.68	3.0
golf2009	0.88	9.0	0.86	6.0	0.89	7.5	0.86	5.5
compact	1.02	20.0	1.02	13.0	1.04	11.5	1.04	9.0
Panel C: Problem (3.18)								
cpu	0.81	4.0	0.76	4.0	0.81	4.0	0.80	4.0
yacht	1.00	1.0	1.00	1.0	1.00	1.0	1.00	1.0
whitewine	1.03	8.0	1.03	6.0	1.03	8.0	1.03	6.0
redwine	1.03	7.0	1.03	7.0	1.03	6.0	1.03	6.0
golf2008	0.68	3.5	0.68	3.0	0.68	3.0	0.68	3.0
golf2009	0.95	6.5	0.92	4.5	0.94	6.0	0.92	4.5
compact	1.04	14.0	1.04	10.0	1.05	11.5	1.05	9.5

Table 3.2: Median MSE and number of non-zeroes (NZ) of the ten random shuffles for the real data sets.

3.3.4 Simulated data

Data generation

Analogously to the overall design of experiments, the synthetic generation of the data closely follows that of Bertsimas and King (2015). The k-th vector of observations $\left(X_k^1, ..., X_k^N\right)$ was generated following a multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = \rho^{|i-j|}$. In particular, we chose $\rho = -0.9$ and $\rho = -0.5$ so as to test the performance of the constraints under highly and moderate correlations. The regression model is taken in small dimension, but with quite a number of irrelevant covariates. More precisely, the number of features was set to 500, although the number of non-zero $\boldsymbol{\beta}$ coefficients was 10. The β_i were uniformly generated in the interval $(-2, 2)$ for i such that $i \bmod p/k = 0$. The response was generated following (1.5), with $\beta_0 = 0$ and the error terms i.i.d. following a $N \sim (0, \sigma^2)$. The variance σ^2

was chosen so as the signal-to-noise ratio (SNR) was 6.32, as in Bertsimas and King (2015)

Results

Analogously to the real datasets results, Table 3.2 shows the median MSE and number of non-zeroes (NZ). The first column corresponds again to the results for the original Problem (3.1) (*Baseline*), and the remaining columns contain the results of the tightened versions. Each panel of Table 3.2 corresponds to one of the methods described in Section 3.3.1.

From Table 3.2 it can be observed that tightening the feasible regions of the baselines methods by adding the significance or the collinearity constraints has a different impact over the solution, usually attaining a better trade-off between sparsity and prediction power when combined. More specifically, reducing the search of the coefficients $\boldsymbol{\beta}$ by adding all the constraints proposed in Section 3.2 improves the MSE yielded with the tightening sets $\mathcal{S}_1$ and $\mathcal{S}_2$ while maintaining or increasing slightly the density of the output in the presence of highly correlated features.

From Panel A some conclusions arise. First, the density of the OLS outputs is always reduced by adding any of the constraints proposed in Section 3.2. In particular, the number of non-zeroes is usually reduced from 500 (the total number of features) to 6 or 7 with around a 25% improvement of the MSE. An exception is found for highly correlated features and collinearity constraints, where an extremely sparse solution is obtained paying with an 85% deterioration over the predictive power. Finally, the tightening set $\mathcal{S}_3$ is observed to balance the effect of the significance and collinearity constraints, either improving the predictive quality of the most sparse solution ($\rho = -0.5$), or attaining a considerably sparse solution with a better MSE ($\rho = -0.9$).

From Panel B it can be concluded that the tightened versions of the Lasso substantially reduce the sparsity but implying a small deterioration over the MSE. For instance, for $\rho = -0.5$ the price of gaining 72.5 zeroes is a 5.6% worsening of the predictive power. Despite the use of the collinearity constraints alone yields the highest deterioration over the MSE for highly correlated features, it seems that the ℓ_1-norm regularization alleviates the extreme behaviour observed for the OLS. Nevertheless, the collinearity constraints help improving the MSE produced by the significance constraints, although they may increase the density of the output.

From Panel C it is observed that tightening Problem (3.18) always leads to a considerable improvement over the predictive power but slightly more dense solutions. In particular, significance constraints help improving the MSE in more than a 30%, although producing around 2 more non-zeroes. As observed previously, the most sparse solutions are obtained for the tightening sets $\mathcal{S}_2$, although in this case the MSE of the baseline method is improved.

	Baseline		$\mathcal{S}_1$		$\mathcal{S}_2$		$\mathcal{S}_3$	
	MSE	NZ	MSE	NZ	MSE	NZ	MSE	NZ
Panel A: OLS								
$\rho = -0.5$	1.00	500.0	0.81	6.0	0.60	7.5	0.75	6.0
$\rho = -0.9$	1.00	500.0	0.77	7.0	1.85	3.0	0.73	6.0
Panel B: Lasso								
$\rho = -0.5$	0.53	80.5	0.63	7.0	0.60	7.5	0.56	8.0
$\rho = -0.9$	0.53	183.0	0.70	6.5	1.67	3.5	0.68	6.5
Panel C: Problem (3.18)								
$\rho = -0.5$	0.76	6.0	0.53	8.0	0.60	7.5	0.55	8.5
$\rho = -0.9$	1.10	5.0	0.69	7.5	0.84	5.5	0.65	8.0

Table 3.3: Median MSE and number of non-zeroes (NZ) of the ten simulated instances.

3.4 Concluding remarks

In this chapter we propose to tighten the parameter region of linear regression models in order to enhance the sparsity of the outputs while attaining a good predictive power. To this aim the underlying optimization problems are modified by adding new constraints to those defining the search space. These constraints detect the most important features for the prediction and avoid misleading estimators that may be obtained in the presence of highly correlated variables.

In order to assess the impact of the new constraints over various linear regression models, three approaches are considered. First, the standard OLS estimation is taken into account. Second, the classic Lasso is considered. Last, a tightening procedure introduced in Bertsimas and King (2015) is included in our numerical illustrations. The results show that the tightened versions are computationally tractable and that sparsity may be enhanced while often improving or maintaining the predictive quality by tightening the feasible regions of the three above-mentioned regression approaches.

Chapter 4

Robust Newsvendor Problem with Autoregressive Demand

In this chapter we explore the single-item newsvendor problem under a novel setting which combines temporal dependence and tractable robust optimization. First, the demand is modeled as a time series which follows an autoregressive process $AR(n)$, $n \geq 1$. Second, a robust approach to maximize the worst-case revenue is proposed: a robust distribution-free autoregressive method for the newsvendor problem, which copes with non-stationary time series, is formulated. A closed-form expression for the optimal solution is found for $n = 1$; for the remaining values of n, the problem is expressed as a nonlinear convex optimization program, to be solved numerically. The optimal solution under the robust method is compared with those obtained under three versions of the classic approach, in which either the demand distribution is unknown, and autocorrelation is neglected, or it is assumed to follow an $AR(n)$ process with normal error terms. Numerical experiments show that our proposal usually outperforms the previous benchmarks, not only with regard to robustness, but also in terms of the average revenue. Extensions to multiperiod and multiproduct models are also discussed.

4.1 Introduction

As mentioned in Chapter 1, the classic approach to the well-known newsvendor problem (1.10) and its famous robust counterpart (1.11) assume the demand is independent and identically distributed along time, which is in practice unrealistic. For this reason, some authors have studied inventory models with time-correlated demand, including AR models (Aviv, 2002; Reyman, 1989; Johnson and Thompson, 1975), compound Poisson processes (Shang and Song, 2003), martingale models of forecast evolution (Dong and Lee, 2003; Lu et al., 2006; Wang et al., 2012), factor models (See and Sim, 2010) or estimation via Kalman filter (Aviv, 2003). Most of these papers either assume perfect knowledge of the distribution function (Levi et al., 2008; Aviv, 2003, 2002; Shang and Song, 2003; Wang et al., 2012; Reyman, 1989) or are focused in calculating bounds of the objective function, which are distribution-free in See and Sim (2010); Lu et al. (2006); Dong and Lee (2003). In contrast, in the work developed here no distributional assumptions are made and the optimal solution is obtained with a closed expression for a particular case, and the problem to be solved in the remaining cases is extremely tractable due to its structural properties: it is a low-dimensional convex problem. Moreover, we do not only cope with temporal demand but also take into account robustness in terms of uncertainty and risk aversion, which provides novelty to this work.

In recent years, risk-averse models have received increasing attention in the inventory literature (see Choi et al. (2011) and the references therein). There, instead of maximizing the expected revenues, other utilities are optimized. Depending on the decision maker's preferences, it may be reasonable, for example, to optimize the probability of achieving a target profit (see for example Kabak and Schiff (1978) or Lau (1980)), the

Return of Investment (Thakkar et al. (1983)), the Cost-Volume-Profit, the CVaR (Chen et al. (2009)) or other risk-averse policies (Eeckhoudt et al. (1995); Choi et al. (2011)). The above-mentioned paper of Yue et al. (2006), as well as Perakis and Roels (2008); Zhu et al. (2013); Jiang et al. (2011) consider the minimax regret decision criterion instead. The robust approach in the newsvendor problem deals with uncertainty in the demand while minimizing the impact over the optimal solution of the worst-case scenario. For example, the landmark Scarf's rule adopts such a criterion, although it enforces independence of the demand along time. Bertsimas and Thiele (2006) also propose a robust inventory approach, where uncertainty intervals for the demand are supposed to be already given. On the contrary, our approach would address the worst-case analysis while coping with time-correlated demands and including information of the historical observations of the demand into the model.

In this chapter we address the newsvendor problem from a new perspective, integrating a distribution-free design with temporal dependence in the demand, into a robust optimization approach. Throughout this chapter we perform a worst-case analysis, seeking the policy maximizing the worst-case revenue. Specifically, our main contributions are:

1. We consider the demand as a time series with non-negligible autocorrelation coefficients. For simplicity, the basic yet versatile autoregressive process of some order n, $AR(n)$, is used as time series model. We follow a distribution-free approach, in the sense that no distributional assumption is imposed over the error terms of the autoregressive model.

2. We implement a robust optimization method based on the uncertainty sets of Bandi and Bertsimas (2012), where the goal is to minimize the losses in the worst-case realization of the parameters. For the particular case when the uncertainty set is modeled with the l_2-norm, a closed-form expression for the optimal solution is obtained in the case $n = 1$. For $n \geq 2$ the problem turns into a tractable nonlinear convex optimization program, solved numerically.

3. We show that our approach outperforms three different classic approaches. In the first one, the demand distribution is assumed to be unknown and is estimated from the sample observations, which are assumed to be independent; in the second one, the demand distribution is assumed to follow an $AR(n)$ process with normal error terms; the third method is the robust distribution-free solution of Scarf (1958) (1.11).

4. We briefly discuss the robust multi-product newsvendor problem with demands correlated over time and between products, and the multi-period case.

The chapter is organized as follows. In next section we briefly introduce autoregressive processes and model various robust newsvendor problems with autoregressive demands. Specifically, we formulate the single-item single-period case in terms of an optimization problem in Section 4.2.1. There, we also discuss the choice of parameters and we show that, for a particular case, the problem is a convex optimization problem, and we obtain a closed-form solution for $n = 1$. A brief extension to the multi-period case is outlined in Section 4.2.2, where the robust modelling of the autoregressive processes is integrated into the inventory model of Bertsimas and Thiele (2006). An extension to the multiple item newsvendor problem is carried out in Section 4.2.3, where the demands of the products are assumed to follow a Vector Autoregressive Process. In Section 4.3 we design and present some numerical examples, where the robust autoregressive model is tested against three different but classic methods, outlined in Section 4.3.1. Data generation and presentation of results are addressed in Sections 4.3.2 and 4.3.3, respectively. Last section is devoted to concluding remarks and extensions.

4.2 The model

Now we describe how to obtain a robust solution for the single period newsvendor problem with $AR(n)$ demand, under the assumption that a realization of the process up to time T is available and no probability distribution is imposed on the errors. To do this we use the uncertainty sets of Bandi and Bertsimas (2012).

4.2.1 The single-product case

In this section we focus on the single-item newsvendor with autoregressive demand, obtaining a closed-form solution when the demand follows and $AR(1)$ process. The solution we obtain is robust in two senses. First, we optimize revenue under the worst-case scenario of the demand. Second, the forecast of the demand is distribution-free, although it takes into account autocorrelation. We undertake this task via a robust optimization problem with uncertainty sets (Ben-Tal and Nemirovski, 1998, 1999): the data is not deterministic but required to belong to an uncertainty set, and all the constraints of the optimization problem must hold for every valid value of the uncertainty set.

Ben-Tal and Nemirovski (2000) show that nominal solutions of optimization problems may become drastically unfeasible under small perturbations in the data. Although they recognize the need of robustifying the nominal problem, they also admit that this is done under a cost over the optimal value. Therefore, risk aversion against uncertainty in the data has often been addressed by alleviating the impact of the worst case scenario over the solution (Bertsimas et al., 2011; Bertsimas and Copenhaver, 2014; Ben-Tal et al., 2009). As mentioned in Hanasusanto et al. (2014), the worst case approach is strongly justified from the theory of choice in economics (Ellsberg, 1961; Gilboa and Schmeidler, 1989). Furthermore, this risk averse approach has been frequently used in the newsvendor context (Scarf, 1958; Gallego and Moon, 1993; Vairaktarakis, 2000; Hanasusanto et al., 2014; Han et al., 2014). Although this type of robust solutions might be too conservative, Bertsimas and Thiele (2006) show that the choice of the uncertainty set can play an important role to help avoiding this over conservativeness. In this section we formulate our robust newsvendor problem with autoregressive demand by means of a minimax optimization problem. We give some guidelines on how to choose the parameters of our model so as to take into account risk preferences and avoid over conservativeness.

Assume we possess the historical data of the demand up to time T $\{X_t\}_{t=1}^T$, but it is unknown for next period. Moreover assume that such demand roughly follows an autoregressive process as in (1.6). Let Q denote the quantity of product to order, assumed to belong to an interval $\mathbb{Q} = [\underline{Q}, \overline{Q}]$. If the demand takes the value X_{T+1}, then a revenue $R(Q, X_{T+1})$ is obtained. Typically, when no shortage costs nor salvage values are taken into account, R is calculated using the following expression:

$$R(Q, X_{T+1}) = \min\{Q, X_{T+1}\} - \left(\frac{c}{v}\right) Q. \tag{4.1}$$

where c and v are the unit cost and selling price, respectively. The goal is to find Q maximizing the revenue $R(Q, X_{T+1})$ under the worst-case scenario for the demand X_{T+1}, assumed to vary in a given uncertainty set. We formulate the robust single-item newsvendor problem as follows:

$$\max_{Q\in\mathbb{Q}} \quad \min_{X_{T+1}} \quad R(Q, X_{T+1}) \tag{4.2}$$

$$\text{s.t}$$

$$\left|\frac{1}{T-n}\sum_{t=n+1}^{T} a_t\right| \leq \Gamma_1, \tag{4.3}$$

$$\|(a_{n+1}, ..., a_T)\|_q \leq \Gamma_2, \tag{4.4}$$

$$X_t = \alpha + \sum_{k=1}^{n} \theta_k X_{t-k} + a_t \qquad t = n+1, ..., T+1, \tag{4.5}$$

$$|a_{T+1}| \leq \Delta \tag{4.6}$$

$$a \leq X_{T+1} \leq b \tag{4.7}$$

The rationale behind the constraints is the following. The vector $(a_{n+1}, ..., a_T)$ is the vector of residuals in the $AR(n)$ model in constraint (4.5). Constraint (4.4) forces the error vector to be small, by bounding its l_q norm. Typical choices are $q = 1$, and thus we make small the absolute value of the residuals, and for $q = 2$ we bound their variance. On the other hand constraint (4.3) requires the mean of the observed errors to be bounded. Note that, as no constraints over the θ_k $k = 1, ..., n$ are made, nonstationary processes can be addressed by problem (4.2)-(4.7). Finally, (4.6) implies that the absolute value of the random value a_{T+1} (which represents the prediction error) is bounded above by some constant Δ. In some real-world situations, the nature of the time series requires to obtain a value of X_{T+1} within a specific interval. This is for example the case of rainfall data and exchange rates, which take non-negative values, or unemployment rates and diseases prevalence, which must lie in $[0, 1]$ (see Carrizosa et al. (2013)). Constraint (4.7) expresses this requirement.

Denote the set of constraints (4.3)-(4.7) by $\mathbb{X}$, and fix $Q_0 \in \mathbb{Q}$. As the function $R(Q_0, \cdot)$ decreases in X_{T+1}, then $\min\limits_{X_{T+1} \in \mathbb{X}} R(Q_0, X_{T+1}) = R(Q_0, \underline{X}_{T+1})$, where $\underline{X}_{T+1}$ is either a, b, or the solution to problem $ARUS(n)$, defined as:

$$\begin{aligned} \min_{\substack{\alpha,\theta_1,\ldots,\theta_n,\\ a_1,\ldots,a_{T+1}}} \quad & X_{T+1} \\ \text{s.t} \quad & \\ & (4.3), (4.4), (4.5), (4.6) \end{aligned} \qquad (ARUS(n))$$

We will refer to this problem as an AutoRegressive process based on Uncertainty Sets, in short $ARUS(n)$. Then, the solution Q^* to problem (4.2)-(4.7) is also the solution to

$$\max_{Q \in \mathbb{Q}} \left(\min\{Q, \underline{X}_{T+1}\} - \left(\frac{c}{v}\right) Q \right).$$

As $R(Q, \underline{X}_{T+1}) = \min\{Q, \underline{X}_{T+1}\} - \left(\frac{c}{v}\right) Q$ is piecewise linear concave function in Q, with a global maximum at $\underline{X}_{T+1}$, then the solution to problem (4.2)-(4.7) is finally given by:

$$Q^* = \begin{cases} \underline{Q} & \text{if} \quad \underline{Q} \geq \underline{X}_{T+1} \\ \underline{X}_{T+1} & \text{if} \quad \underline{Q} \leq \underline{X}_{T+1} \leq \overline{Q} \\ \overline{Q} & \text{if} \quad \underline{X}_{T+1} \geq \overline{Q} \end{cases} \tag{4.8}$$

Choice of parameters

The values of Γ_1, Γ_2 and Δ in the formulation of the $ARUS(n)$ problem are chosen according to the practitioner's criterion. Next, we describe a procedure to select such parameters using the concept of uncertainty sets along the lines of Bandi and Bertsimas (2012), who define the uncertainty set (1.3). Since the errors terms $a_{n+1}, ..., a_T$ are assumed i.i.d., we can define an uncertainty set of the form (1.3) for them. As there is no loss of generality in assuming $\mu_a = 0$, in combination with (4.5) we would have the following uncertainty set:

$$\mathcal{U}\left(\Gamma_1^*\right) = \left\{ (\alpha, \theta_1, \ldots, \theta_n) \quad \text{s.t} \quad \frac{1}{T-n} \left| \sum_{t=n+1}^{T} X_t - \alpha - \sum_{k=1}^{n} \theta_k X_{t-k} \right| \leq \frac{\Gamma_1^* \sigma_a \sqrt{T-n}}{T-n} \right\}.$$

Since σ_a is unknown in practice, we suggest to set it to:

$$\sigma_a \approx (1 + \xi_1/100)\sigma_0,$$

where $(1 + \xi_1/100)$ indicates a perturbation (depending on the value of ξ_1) of σ_0, which denotes the optimal value to the problem of minimizing the variance of the errors $a_{n+1}, .., a_T$.

Consider now the constraint (4.4). For $q = 2$ this constraint bounds the variance of the errors. In this case we can proceed similarly as before and substitute Γ_2 by a certain perturbation of the minimum value attained by the errors' variance:

$$\Gamma_2 \approx (1 + \xi_2/100)\,\gamma_2,$$

where γ_2 is the obtained by minimizing the variance of the errors subject to constraints (4.3) and (4.5).

Finally, consider (4.6). The choice of Δ is crucial, since it bounds the value of the prediction error a_{T+1}. Since the sequence $\{a_t,\ t > 0\}$ is i.i.d., it seems reasonable to relate Δ with the values $a_1, a_2, \ldots, a_T$. Since we have already obtained the optimal $\alpha^\star, \theta_1^\star, \ldots, \theta_n^\star$ feasible for (4.3) and (4.5) and for which the errors' variance is minimum, then it is straightforward to obtain the values $a_1^*, \ldots, a_T^*$ (by substituting $\alpha^\star, \theta_1^\star, \ldots, \theta_n^\star$ into (1.6)). Therefore, one possible choice of Δ is the empirical $k-$th quantile of the sample $(a_1^*, \ldots, a_T^*)$, for some value of k, large enough. In the context of the newsvendor problem it seems reasonable to relate the value of Δ to the *profitability* of the product $\frac{c}{v}$, which should influence the decision maker's risk aversion. From Schweitzer and Cachon (2000), in the case of high profit products (that is, $\frac{c}{v} \leq 0.5$), risk aversion may be reduced by allowing the decision maker to buy more items. This implies that $\underline{X}_{T+1}$ is allowed to be higher, which may be obtained by reducing the value of Δ. Thus, in the context of the newsvendor problem, it makes sense to choose Δ as the empirical $U\left(\frac{c}{v}\right)$-th quantile of the sample $(a_1, \ldots, a_T)$, where U is an utility function that depends on the decision's maker risk aversion.

Properties and closed-form solution for $q = 2$

In this section we analyze the properties of the robust newsvendor problem with autoregressive demand (4.2)-(4.7) for $q = 2$ and we obtain a closed-form solution when the demand follows an $AR(1)$.

Consider problem

$$-\Delta + \min_{\boldsymbol{\theta}} \left(\frac{-\sum_{t=n+1}^{T} \varphi_t(\boldsymbol{\theta})}{T-n} - \min\left\{\Gamma_1, \sqrt{\Gamma_2 - H(\boldsymbol{\theta})}\right\} + \sum_{k=1}^{n} \theta_k X_{T+1-k} \right) \tag{4.9}$$

$$\text{s.t}$$

$$\Gamma_2 - H(\boldsymbol{\theta}) \geq 0, \tag{4.10}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, and $\varphi_t(\boldsymbol{\theta})$, $H(\boldsymbol{\theta})$ are defined as

$$\begin{aligned} \varphi_t(\boldsymbol{\theta}) &= \sum_{k=1}^{n} \theta_k X_{t-k} - X_t, \\ H(\boldsymbol{\theta}) &= \frac{1}{T-n} \sum_{t=n+1}^{T} \varphi_t(\boldsymbol{\theta})^2 - \frac{1}{(T-n)^2} \left(\sum_{t=n+1}^{T} \varphi_t(\boldsymbol{\theta}) \right)^2, \end{aligned}$$

and fix the parameters Γ_1, Γ_2 and Δ. Then the next results hold.

Proposition 1. *The solution to problem (4.2)-(4.7) is either* $\underline{Q}$, $\overline{Q}$, *a, b, or the solution to (4.9)-(4.10).*

Proof. First, we provide a lemma that will be necessary for the proof.

Lemma 1. *If*

$$\frac{1}{T-n} \sum_{t=n+1}^{T} \left(\alpha + \sum_{k=1}^{n} \theta_k X_{t-k} - X_t \right)^2 \leq \Gamma_2, \quad \textit{for all } \alpha, \theta_1, \ldots, \theta_n,$$

then $\Gamma_2 - H(\boldsymbol{\theta}) \geq 0$.

Proof. Note first that if $\Gamma_2 - H(\boldsymbol{\theta}) < 0$ then

$$\Gamma_2 - \frac{1}{T-n} \sum_{t=n+1}^{T} \varphi_t^2 < \frac{-1}{(T-n)^2} \left(\sum_{t=n+1}^{T} \varphi_t \right)^2 \tag{4.11}$$

and

$$\frac{1}{T-n} \sum_{t=n+1}^{T} (\alpha + \varphi_t)^2 = \alpha^2 + \frac{2\alpha}{T-n} \sum_{t=n+1}^{T} \varphi_t + \frac{1}{T-n} \sum_{t=n+1}^{T} \varphi_t^2$$

Assume there exists $(\alpha, \boldsymbol{\theta})$ such that $\Gamma_2 - H(\boldsymbol{\theta}) < 0$ and $\frac{1}{T-n}\sum(\alpha+\varphi_t)^2 \leq \Gamma_2$. Then,

$$\alpha^2 + \frac{2\alpha}{T-n}\sum_{t=n+1}^{T}\varphi_t \leq \Gamma_2 - \frac{1}{T-n}\sum_{t=n+1}^{T}\varphi_t^2$$

but from (4.11)

$$\alpha^2 + \frac{2\alpha}{T-n}\sum_{t=n+1}^{T}\varphi_t < -\frac{1}{(T-n)^2}\left(\sum_{t=n+1}^{T}\varphi_t\right)^2,$$

and therefore

$$\left(\alpha + \frac{1}{T-n}\sum_{t=n+1}^{T}\varphi_t\right)^2 = \alpha^2 + \frac{2\alpha}{T-n}\sum_{t=n+1}^{T}\varphi_t + \frac{1}{(T-n)^2}\left(\sum_{t=n+1}^{T}\varphi_t\right)^2 < 0$$

which is a contradiction. □

From (4.3),

$$\alpha \leq U_1(\boldsymbol{\theta}) = \Gamma_1 - \frac{1}{T-n}\sum_{t=n+1}^{T}\varphi_t(\boldsymbol{\theta}),$$

$$\alpha \geq L_1(\boldsymbol{\theta}) = -\Gamma_1 - \frac{1}{T-n}\sum_{t=n+1}^{T}\varphi_t(\theta).$$

and (4.4) implies

$$\alpha \leq U_2(\boldsymbol{\theta}) = \frac{-1}{T-n}\sum_{t=n+1}^{T}\varphi_t(\boldsymbol{\theta}) + \sqrt{\Gamma_2 - H(\boldsymbol{\theta})},$$

$$\alpha \geq L_2(\boldsymbol{\theta}) = \frac{-1}{T-n}\sum_{t=n+1}^{T}\varphi_t(\boldsymbol{\theta}) - \sqrt{\Gamma_2 - H(\boldsymbol{\theta})},$$

Then given a fixed $\boldsymbol{\theta}$, (4.3)-(4.5) is written as:

$$-\Delta + \min \quad \left(\alpha + \sum_{k=1}^{n}\theta_k X_{T+1-k}\right)$$

$$\text{s.t} \qquad \begin{cases} \alpha \geq \max\{L_1(\boldsymbol{\theta}), L_2(\boldsymbol{\theta})\} \\ \alpha \leq \min\{U_1(\boldsymbol{\theta}), U_2(\boldsymbol{\theta})\} \\ \sqrt{\Gamma_2 - H(\boldsymbol{\theta})} \geq 0 \end{cases},$$

or equivalently as

$$\begin{array}{ll} -\Delta + \min & \left(\max\left\{L_1(\boldsymbol{\theta}), L_2(\boldsymbol{\theta})\right\} + \sum_{k=1}^{n} \theta_k X_{T+1-k} \right) \\ \text{s.t} & \left\{ \sqrt{\Gamma_2 - H(\theta)} \geq 0 \right. \end{array}$$

which is equivalent to (4.9)-(4.10).

□

Proposition 2. *Equations (4.9)-(4.10) define a smooth convex optimization problem.*

Proof. We provide two lemmas needed for the proof of Proposition 2.

Lemma 2. *The function $H(\boldsymbol{\theta})$ is convex.*

Proof. Since $H(\boldsymbol{\theta})$ is an estimator of the variance of $\varphi(\boldsymbol{\theta})$, it can be rewritten as

$$H(\boldsymbol{\theta}) = \frac{1}{T-n} \sum_{t=n+1}^{T} \left(\varphi_t(\boldsymbol{\theta}) - \frac{1}{(T-n)} \sum_{t=n+1}^{T} \varphi_t(\boldsymbol{\theta}) \right)^2$$

which is a convex function on $\boldsymbol{\theta}$ since each term is the square of an affine function.

□

Lemma 3. *The feasible region of the problem defined by (4.9)-(4.10) is convex.*

Proof. The feasible region of the problem defined by (4.9)-(4.10) is the set $F = \{\boldsymbol{\theta} \text{ s.t. } \Gamma_2 - H(\boldsymbol{\theta}) \geq 0\}$. After some algebra, we have

$$\begin{aligned} \Gamma_2 - H(\boldsymbol{\theta}) &= \Gamma_2 - \frac{1}{T-n} \sum_{t=n+1}^{T} \left(\sum_{k=1}^{n} \theta_k X_{t-k} \right)^2 + \frac{2}{T-n} \sum_{t=n+1}^{T} \sum_{k=1}^{n} \theta_k X_{t-k} X_t \\ & -\frac{1}{T-n} \sum_{t=n+1}^{T} X_t^2 + \frac{1}{(T-n)^2} \left(\sum_{t=n+1}^{T} \sum_{k=1}^{n} \theta_k X_{t-k} \right)^2 \end{aligned}$$

From the definitions in (4.16) it can be seen that

$$\Gamma_2 - H(\boldsymbol{\theta}) = \Gamma_2 - \sum_{k=1}^{n} \theta_k^2 V_k - 2 \sum_{k=1}^{n} \theta_k \sum_{r=k+1}^{n} \theta_r C_{k,r} + 2 \sum_{k=1}^{n} \theta_k C_{k,0} - V_0 \tag{4.12}$$

Since by Lemma 2, H is convex, the region defined by $\Gamma_2 - H(\boldsymbol{\theta}) \geq 0$ is convex. □

Proof of Proposition 2: Since the feasible region is convex then, it suffices to prove that the objective function is convex too. From Lemma 2, $H(\boldsymbol{\theta})$ is convex and therefore

$$-\min\left\{\Gamma_1, \sqrt{\Gamma_2 - H(\boldsymbol{\theta})}\right\}$$

is also convex. On the other hand, $\sum_{t=n+1}^{T} \varphi_t(\boldsymbol{\theta})$ is linear on $\boldsymbol{\theta}$, thus

$$-\Delta + \min_{\boldsymbol{\theta}} \left(\frac{-\sum_{t=n+1}^{T} \varphi_t(\boldsymbol{\theta})}{T-n} - \min\left\{\Gamma_1, \sqrt{\Gamma_2 - H(\boldsymbol{\theta})}\right\} + \sum_{k=1}^{n} \theta_k X_{T+1-k} \right)$$

is convex.

□

Moreover, the global optimum in the case $p = 1$ can be obtained in closed form as the next result shows.

Theorem 1. *For $n = 1$, the optimal solution to the minimization problem defined by (4.9)-(4.10) is reached at one of these values for $\theta_1^\star$:*

$$\theta_1^{\star(1)} = \frac{-C_{1,0} \pm \sqrt{C_{1,0}^2 + V_1\left(\Gamma_2 - V_0 - \Gamma_1^2\right)}}{-V_1}, \tag{4.13}$$

$$\theta_1^{\star(2)} = \frac{-C_{1,0} \pm \sqrt{C_{1,0}^2 + V_1\left(\Gamma_2 - V_0\right)}}{-V_1}, \tag{4.14}$$

$$\theta_1^{\star(3)} = \frac{-C_{1,0}(V_1 + a^2) \pm \sqrt{C_{1,0}^2(V_1 + a^2)^2 + a^2 V_1 (V_1 + a^2)\left(\Gamma_2 - V_0 - C_{1,0}\right)}}{-V_1}, \tag{4.15}$$

where

$$V_k = var\left(X^k\right), \quad C_{k,h} = cov\left(X^k, X^h\right) \tag{4.16}$$

respectively denote the variance of X^k and covariance matrix between X^k and X^i where

$$X^k = \left(X_{n+1-k} \ldots, X_{T-k}\right), \tag{4.17}$$

and $a = \dfrac{-1}{T-1} S_1 + X_T$.

Proof. Two main cases may be distinguished

1. Either $\min\left\{\Gamma_1, \sqrt{\Gamma_2 - H(\theta_1)}\right\} = \Gamma_1$,

2. or $\min\left\{\Gamma_1, \sqrt{\Gamma_2 - H(\theta_1)}\right\} = \sqrt{\Gamma_2 - H(\theta_1)}$

Consider the first case. Then, the problem to be solved is written as

$$-\Delta - \Gamma_1 + \min \left(\frac{-\sum\limits_{t=n+1}^{T} \varphi_t(\theta_1)}{T-n} + \sum_{k=1}^{n} \theta_k X_{T+1-k} \right)$$

$$\text{s.t} \qquad \begin{cases} \Gamma_2 - H(\theta_1) \geq 0 \\ \sqrt{\Gamma_2 - H(\theta_1)} \geq \Gamma_1 \end{cases}$$

Note that the first constraint is redundant, as Γ_1 is supposed to be positive. Since the objective function is linear on θ_1, the optimum is reached at the frontier of the feasible region

$$F = \left\{ \theta_1 \text{ s.t. } \Gamma_1 = \sqrt{\Gamma_2 - H(\theta_1)} \right\}.$$

Consider (4.12) with $n = 1$. Then

$$\Gamma_2 - \Gamma_1^2 - H(\theta_1) = -\theta_1^2 var(X^1) + 2\theta_1 cov(X^1, X^0) + (\Gamma_2 - var(X^0) - \Gamma_1^2)$$

which is equal to zero if $\theta_1 = \theta_1^{\star(1)}$ as in (4.13).

Assume now that $\min\left\{\Gamma_1, \sqrt{\Gamma_2 - H(\theta_1)}\right\} = \sqrt{\Gamma_2 - H(\theta_1)}$. Then the objective function to be minimized can be written as

$$A_2(\theta_1) = -\Delta + \frac{-1}{T-1} \sum_{t=2}^{T} (\theta_1 X_{t-1} - X_t) - \sqrt{\Gamma_2 - H(\theta_1)} + \theta_1 X_T,$$

and let $F_2(\theta_1) = A_2(\theta_1) + \Delta$. This function is convex and therefore, the optimal solution for the unconstrained problem is reached when the derivatives are null, where

$$\frac{dF_2(\theta_1)}{d\theta_1} = a + \frac{\frac{1}{T-1} \sum_{t=2}^{T} (\theta_1 X_{t-1} - X_t)\, b_t}{\sqrt{\Gamma_2 - H(\theta_1)}},$$

and where $a = \dfrac{-1}{T-1} \sum\limits_{t=2}^{T} X_{t-1}$ and $b_t = X_{t-1} - \dfrac{1}{T-1} \sum\limits_{t_0=2}^{T} X_{t_0-1}$. Note that $\dfrac{dF_2(\theta_1)}{d\theta_1}$ exists if and only if $\Gamma_2 - H(\theta_1) > 0$. Thus, the optimal θ_1 is either the one which the partial derivative of F_2 in the feasible region of the considered problem is zero or it is found at the frontier of such feasible region. Three cases may therefore be considered

(2.1) The minimun is reached at $\Gamma_2 - H(\theta_1) = 0$,

(2.2) The minimun is reached at $\sqrt{\Gamma_2 - H(\theta_1)} = \Gamma_1$, (2.3) The optimal θ_1 is the one such that $\dfrac{\partial F_2(\theta_1)}{\partial \theta_1} = 0.$

Case (2.2) has been already solved. Consider the case (2.1). Then, after substituting $p = 1$ in (4.12), expression (4.14) is obtained. Finally, the problem to be solved for the case (2.3) turns into

$$-\Delta + \min F_2(\theta_1) \quad \text{s.t} \begin{cases} \Gamma_2 - H(\theta_1) > 0 \\ \sqrt{\Gamma_2 - H(\theta_1)} \leq \Gamma_1 \end{cases} \tag{4.18}$$

An optimal solution to (4.18) is obtained in the case (2.3) if there exists θ_1 such that $\dfrac{dF_2(\theta_1)}{d\theta_1} = 0$, which in addition satisfies the constraints of problem (4.18). Such value θ_1 is obtained by

$$\left(-a\sqrt{\Gamma_2 - H(\theta_1)}\right)^2 = \left(\frac{1}{T-1}\sum_{t=2}^{T} \varphi_t(\theta_1) b_t\right)^2 .$$

After some algebra, the right term is expressed as:

$$\frac{1}{(T-1)^2}\left(\sum_{t=2}^{T} \varphi_t(\theta_1) b_t\right)^2 = \left(cov\left(\theta_1 X^1, X^1\right) - cov\left(X^0, X^1\right)\right)^2 ,$$

from which the next quadratic function is obtained:

$$\theta_1^2(V_1^2 + a^2 V_1) - 2\theta_1 C_{1,0}(V_1 + a^2) - a^2(\Gamma_2 - V_0 - C_{1,0}), \tag{4.19}$$

which is equal to zero if and only if $\theta_1 = \theta_1^{\star(3)}$ as in the expression (4.15).

□

4.2.2 The multi-period case

It is reasonable to study the multi-period newsvendor problem whenever the product perishes at a rate that extends to more than one period. Bertsimas and Thiele (2006) propose a robust multi-period inventory approach in which prediction intervals for the demand are supposed to be given. Although this approach deals with holding and shortage costs, it does not provide any information about the demand or guidelines on how to obtain its forecasts. Here we propose to use our modeling of the robust autoregressive demand to obtain prediction intervals that can be embedded into the multi-period approach of Bertsimas and Thiele (2006). In this way we are able to: (i) take into account the temporal autocorrelation of the demand, (ii) provide robust forecasts for the demands, which fits in with the robust approach Bertsimas and Thiele (2006) propose, and (iii) intimately bound the demand estimation with the inventory problem, as the forecasts can be done by taking into account the risk aversion of the user, which depends on the profitability of the product $\frac{c}{v}$.

Once provided the extremes values $\underline{X}_{T+1}$ and $\overline{X}_{T+1}$, a multi-stage prediction approach, as e.g. in Cheng et al. (2006), is possible: we minimize and maximize the $ARUS(n)$ at each time step and use the obtained values for the next period. Note that if the demand follows and $AR(n)$ process then X_{T+l} depends on the n previous values, thus the prediction interval obtained for the unknown data in $X_{T+l-1}, ..., X_{T+l-n}$ may affect the forecast on instant $T + l$. The robust multi-stage approach we propose to estimate the demand consists on simply adding the constraints:

$$\underline{X}_{T+m} \leq X_{T+m} \leq \overline{X}_{T+m} \quad m = l-n, ..., l-1$$

to the $ARUS(n)$ optimization problem when X_{T+m} is unknown. The values of $\underline{X}_{T+m}$ and $\overline{X}_{T+m}$ have been previously obtained by minimizing (respectively maximizing) the $ARUS(n)$ problem for X_{T+m}. Note that the value of parameter Δ can be modified on each step to deal with variations in the selling price and the cost.

4.2.3 The multi-product case

In this section we analyze the extension to the multiple item newsvendor problem when the products are not replaceable. Many papers, such as Zhang (2012); Lau and Lau (1995); Abdel-Malek and Montanari (2005); Choi et al. (2011), have treated this problem considering demands possibly correlated amongst products. Without exception, they assume the demands' marginal density functions are given for each product and disregard temporal correlation. In contrast, we propose a distribution-free approach and we assume that the demands of the products are not only correlated over time but also correlated between items. Specifically, we extend the model in (1.6) by assuming that the demands follow a Vector Autoregressive Process of or order n, $VAR(n)$ thereafter. Such a process can be written as in (1.8). Denote the demand of the N products in time instant t by $\mathbf{X}_t = \left(X_t^1, ..., X_t^N\right)'$, the vector of constants by $\boldsymbol{\alpha} \in \mathbb{R}^N$, and $A_k = (\theta^i_{jk})_{i,j} \in \mathbb{R}^{N \times N}$ gathers the coefficients relating the values of the time series with the demand k periods behind. Finally $\mathbf{a}_t \in \mathbb{R}^N$ represents the contemporaneous error terms for all demands series at time instant t.

For the multi-product case we consider an uncertainty set gathering the constraints (4.3) and (4.4) for each product, while including the demands' correlation between time and items expressed by (1.8). In order to bound the prediction errors in such a way that a contemporaneous dependence between the shocks of the different products can be addressed, we add the constraint $\left\| \left(\frac{a^1_{T+1}}{\Delta^1}, ..., \frac{a^N_{T+1}}{\Delta^N} \right) \right\|_s \leq 1$ to our multiproduct problem. Note that for $s = \infty$, this is equivalent to requiring (4.6) to hold for each product $j = 1, ...N$.

In conclusion, we model the robust multi-item newsvendor with correlated demand as:

$$\max_{\mathbf{Q}\in\mathbb{Q}} \min_{\substack{\mathbf{X}_{T+1},\boldsymbol{\alpha},\\ A_k,\mathbf{a}_t}} \mathbf{R}(Q^1,...,Q^N N X^1_{T+1},...,X^N_{T+1}) \tag{4.20}$$

s.t

$$\left|\frac{1}{T-n}\sum_{t=n+1}^{T} a^j_t\right| \leq \Gamma^j_1, \qquad j=1,..,N \tag{4.21}$$

$$\|(a^j_{n+1},...,a^j_T)\|_q \leq \Gamma^j_2, \qquad j=1,..,N \tag{4.22}$$

$$\mathbf{X}_t = \boldsymbol{\alpha} + \sum_{k=1}^{n} A_k \mathbf{X}_{t-k} + \mathbf{a}_t \qquad t=n+1,...,T+1 \tag{4.23}$$

$$\left\|\left(\frac{a^1_{T+1}}{\Delta^1},...,\frac{a^N_{T+1}}{\Delta^N}\right)\right\|_s \leq 1 \tag{4.24}$$

Here $\mathbf{R}(Q^1,...,Q^N,X^1_{T+1},...,X^N_{T+1}) = \sum_{j=1}^{N}\mathbf{R_j}(Q^j,X^j_{T+1})$, $\mathbb{Q}$ is a set of constraints on the quantity of products to acquire. Note that problem (4.20)-(4.24) takes into account the correlation of the demand in two aspects (time and products) while providing a distribution free approach. Define the function

$$\phi(Q^1,...,Q^N) = \min_{\text{s.t.}\ (4.21)\text{-}(4.24)} \sum_{j=1}^{N}\mathbf{R_j}(Q^j,X^j_{T+1})$$

which gives, for quantities $Q^1,...,Q^N$ the worst-case revenue. Observe that, for each $\mathbf{R}_j$ as in (4.1), the function ϕ is concave and thus (4.20)-(4.24) is a concave maximization problem. It will be linearly constrained if, as customary in the literature, capacity, weight or budget constraints are written as linear functions (Hadley and Whitin, 1963; Abdel-Malek and Montanari, 2005)

4.3 Numerical illustrations

Now we are going to test the performance of our single-product approach against those described in Section 4.3.1. To make the results as complete as possible we have checked the obtained average revenue and small quantiles for a large number of simulated data sets with different properties, such as the correlation, distribution of errors or seasonality. To further explore the behavior of our approach, different values of n have been considered to generate possibly periodic $AR(n)$ processes, but only results for $n = 1$ are included here as the same performance was observed for the other cases tested.

The proposed approach will be compared with three classic approaches, which will be called *static*, *Scarf* and $AR(n)$. In the first approach, the demand distribution will be unknown and estimated from the sample observations (assumed to be independent), the second assumes the distribution function for the demand is uncertain, as in the third approach normality in the error terms will be assumed.

4.3.1 Benchmark methods

In this section we describe in detail the three benchmark approaches: *static*, *Scarf* and *AR*(n).

If the demand distribution F were known and the expected revenue were to be maximized, then the optimal quantity $Q^\star_s$ would be given by the specific quantile of the distribution function (1.10), which depends on the cost (c) and selling prices (v). Note that under this approach the temporal dependence in the data is ignored, so we call it *static* thereafter. In practice the distribution of the demand is unknown and therefore in the static approach, an estimation of the distribution function must be employed. Usually, the empirical distribution function $\hat{F}$, which converges to the true cumulative density function (cdf) F for a large enough sample, is considered.

The so called Scarf's rule is a robust solution, given by (1.11), in which the distribution of the demand is unknown except for its mean and variance. This solution is robust because it assumes uncertainty over the distribution function of the demand. However, as our numerical results show, this solution may suffer from an excessive conservativeness.

In the case of the classic $AR(n)$ approach the forecast was assumed to follow a normal distribution. Therefore, under normality, the optimal solution for the newsvendor problem would be

$$Q_{AR(n)} = \Phi^{-1}_{(\hat{X}_{T+1},\sigma_a)}\left(1-\frac{c}{v}\right) \tag{4.25}$$

where $\Phi^{-1}_{(\hat{X}_{T+1},\sigma_a)}$ is the inverse cdf of a normal distribution with mean $\hat{X}_{T+1}$ and standard deviation σ_a. Since $Q_{AR(n)}$ may be negative and the demand always takes non-negative values, the quantity $Q^\star_{AR(n)} = \max\left\{0, Q_{AR(n)}\right\}$ will be considered instead.

4.3.2 Synthetic data generation and experiments design

The performance of the different inventory policies is illustrated by different simulational experiments, for which samples of the $AR(n)$ process (representing the demand series) are artificially generated. In this section we describe how the synthetic $AR(n)$ data have been generated, and we specify the choice of parameters for our model.

We have generated the demand series $\{X_t\}_{t>0}$ following an $AR(n)$ process as in (1.6). Three different distributions for the error terms were tested in our experiments, all of them chosen so as to generate non-negative demand series. First, $a_t \sim N(4, 1)$; also, we found of interest to check the behavior of the methods when heavy-tails are incorporated in the generator model, as done in Huh et al. (2011), who choose Pareto and Lognormal distributions to check the performance of their inventory approach under samples of time-independent demand generated with heavy-tailed distributions. Therefore, we also set $a_t \sim LN(0, 3)$ and $a_t \sim Par(1, 1)$, where LN and Par denote the standard Lognormal and Pareto distributions.

A different aspect to be considered when simulating the data is the strength of the temporal dependence. In our experiments, two values of $\boldsymbol{\theta}$ were set. Note that for $p = 1$ the coefficient θ_1 represent the lag-1 autocorrelation coefficient thus, in order to test the methods on highly and moderately correlated time series, $\boldsymbol{\theta} = \theta_1 = 0.9$ and $\boldsymbol{\theta} = \theta_1 = 0.5$ were fixed. Figure 4.1 illustrates demand time series generated with the highest value of the correlation and both normal and lognormal errors distributions. As mentioned before, other values of p have been tested with different values of $\boldsymbol{\theta}$ but the conclusions obtained were analogous to those for $p = 1$.

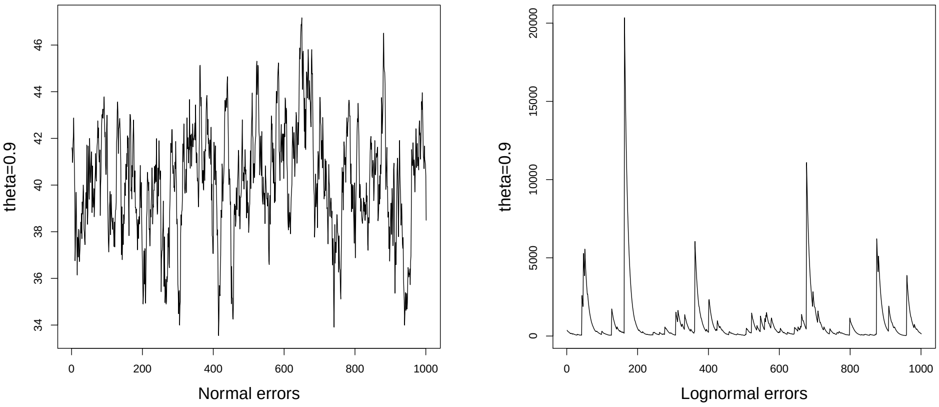


Figure 4.1: Examples of highly correlated autoregressive demands generated with errors following $N(4, 1)$ and $LN(0, 3)$.

Finally, after some testing, the risk seeker strategy where Δ is the $\left(\frac{c}{v}\right)^2$-th quantile of $(a_1, \ldots, a_T)$, has been adopted under three types of profit products: $\frac{c}{v} \in \{0.75,\ 0.5,\ 0.25\}$, representing low-, neutral- and high-profit products, respectively. In this way the proposed robust approach avoids being too conservative. The perturbation parameters β and ν of the $ARUS(n)$ were both set to 5, allowing therefore a 5% perturbation over the minimum variance and standard deviation, respectively. The parameter Γ_1^* was set as the 0.95 quantile of the standard normal distribution following the reasoning of Bandi and Bertsimas (2012).

A total of 1000 series of length $T + 1 = 1000$ were generated for each $\boldsymbol{\theta}$ and each error's probability distribution. The first $T = 999$ values have been used as train set in order to estimate the parameters of the inventory policies proposed in Section 4.3.1

and the $ARUS(n)$, and the remaining value $t = 1000$ has been used as validation set. Therefore, the next process has been followed in order to calculate the revenue of the different approaches:

1. Determine the optimal quantity Q^* of products to buy for instant $T+1$ having available the demand historical records for $t = 1, ..., T$
2. The demand in instant $T+1$ is realized, and then the revenue is calculated by using the expression (4.1).

4.3.3 Results

The results obtained are illustrated by Table 4.1 and Figure 4.2. Table 4.1 shows the average revenue and the frequency of losses for the different approaches (namely, static, Scarf, classic $AR(1)$ prediction method, and robust $ARUS(1)$ method) under three different statistical distributions for the error terms ($Par(1,1)$, $LN(0,3)$ and $N(4,1)$). In Table 4.1 two levels of dependence ($\boldsymbol{\theta} = 0.9$ versus $\boldsymbol{\theta} = 0.5$) are considered.

From Table 4.1 several conclusions can be obtained. We first analyze highly autocorrelated series. First, we point out that under normally distributed errors, the four competing approaches perform similarly in terms of both the average revenue and frequency of losses. However, significant differences are found when the errors follow a distribution with heavier tails. Second, in both the Pareto and Lognormal cases, the robust approach outperforms the other three in terms of both the average revenue and percentage of losses, being the $LN(0,3)$ under neutral-profit product ($c/v = 0.5$) an exception, as the $AR(1)$ attains a better average revenue. Third, the methods that do not take into account the correlation of the data present the poorest performance, Scarf being the worst approach, usually yielding a higher percentage of losses that the $ARUS(1)$ and a lower average revenue. Fourth, the frequency of losses is always zero under the proposed robust approach, while it may be moderately high for both the static approach and the classic $AR(1)$ prediction method when high-profit products are considered.

Consider now moderately correlated series. Although the overall picture is analogous to that of the highly correlated case, some differences arise. First, it is interesting to note how for the low-profit products case with normally distributed errors all approaches attain a fraction of runs with losses bigger than zero, Scarf being the most extreme one. Second, it is worth highlighting the poor performance of the $AR(1)$ forecasting method under heavy-tailed distributions for the errors. Note that for high-profit products negative average revenues are obtained, the frequency of losses being extraordinarily high (this last phenomenon is also observed under neutral-profit products). On the contrary, the Scarf's rule is too conservative, always attaining a zero average rev-

			$Par(1,1)$		$LN(0,3)$		$N(4,1)$	
$\boldsymbol{\theta}$	c/v	Method	Avg. rev.	% loss	Avg. rev.	%loss	Avg. rev.	%loss
0.9	0.25	Static	25.70	7.50	157.64	27.50	29.16	0.00
		Scarf	11.76	12.90	46.26	15.30	29.16	0.00
		$AR(1)$	145.75	22.50	583.51	36.10	29.58	0.00
		$ARUS(1)$	160.22	0.00	674.83	0.00	29.45	0.00
	0.50	Static	13.01	12.60	65.09	23.50	19.02	0.00
		Scarf	1.52	3.20	0.09	0.10	19.02	0.00
		$AR(1)$	118.60	4.50	490.27	10.90	19.52	0.00
		$ARUS(1)$	105.22	0.00	437.73	0.00	19.49	0.00
	0.75	Static	4.94	11.60	20.19	16.30	9.25	0.00
		Scarf	0.00	0.00	0.00	0.00	9.21	0.00
		$AR(1)$	50.51	0.00	187.70	0.00	9.64	0.00
		$ARUS(1)$	52.25	0.00	217.49	0.00	9.64	0.00
0.5	0.25	Static	2.72	28.40	9.29	47.70	5.56	0.00
		Scarf	0.00	0.20	0.00	0.00	5.56	0.00
		$AR(1)$	-27.63	82.00	-145.12	87.80	5.61	0.00
		$ARUS(1)$	7.23	0.10	33.08	0.00	5.46	0.00
	0.50	Static	1.16	26.00	2.80	35.70	3.48	0.00
		Scarf	0.00	0.00	0.00	0.00	3.48	0.00
		$AR(1)$	3.02	63.60	4.63	76.90	3.55	0.00
		$ARUS(1)$	4.16	0.00	19.19	0.00	3.51	0.00
	0.75	Static	0.35	17.70	0.61	21.90	1.60	1.90
		Scarf	0.00	0.00	0.00	0.00	1.60	2.60
		$AR(1)$	2.29	0.00	9.31	0.00	1.66	0.90
		$ARUS(1)$	1.96	0.00	9.39	0.00	1.65	0.50

Table 4.1: Average revenue and frequency of losses obtained for highly and moderately autocorrelated series under the four considered approaches (static, Scarf, classic $AR(1)$ forecasting method, and $ARUS(1)$), under Pareto, Lognormal and normally distributed error terms.

enue for heavy-tailed demands. In this case, in which data are not highly correlated, the static approach does not behave as poorly as in the previous example.

In conclusion, it could be said that the robust autoregressive approach is more stable than the static and $AR(1)$ approaches, and overcomes extreme conservativeness of the Scarf method: it usually performs better or equivalently to all methods in terms of average revenue and always outperforms or is equivalent when minimizing the frequency of losses. Last, it is interesting to note that the higher the profit of a product is, the worse the classic $AR(1)$ forecasting method performs.

As an alternative illustration of the different prediction methods' performance in the least favorable scenarios, we provide Figure 4.2, which depicts the predicted empirical cdf of the revenue in the interval of probabilities $[0, 0.2]$. For the sake of abbreviation, only the neutral products case, i.e., for $\frac{c}{v} = 0.5$, has been depicted. In the top of Figure 4.2 the error terms follow a $N(4,1)$ distribution, while in the central and bottom panels the errors are assumed to be $LN(0,3)$ and $Par(1,1)$ distributed, respectively.

Time series with moderate and high autocorrelation ($\boldsymbol{\theta} = 0.5$, $\boldsymbol{\theta} = 0.9$ respectively) are considered, and the results are given in the left and right column.

Several conclusions can be obtained from Figure 4.2. First, it can be deduced that for demand with normally distributed error the classic $AR(1)$ and robust $ARUS(1)$ approaches perform equivalently for highly correlated demand (right column), while for demand with lower correlation the $ARUS(1)$ approach outperforms the classic $AR(1)$. In both cases, the static and Scarf approaches present a poorer performance. Figure 4.2 illustrates the same phenomena reported on Table 4.1 for demand with heavy-tailed errors: the classic $AR(1)$ forecasting method performs poorly, obtaining highly negative revenues. Although the Scarf approach outperforms both the static and classic autoregressive approaches, it is worse than the $ARUS(1)$.

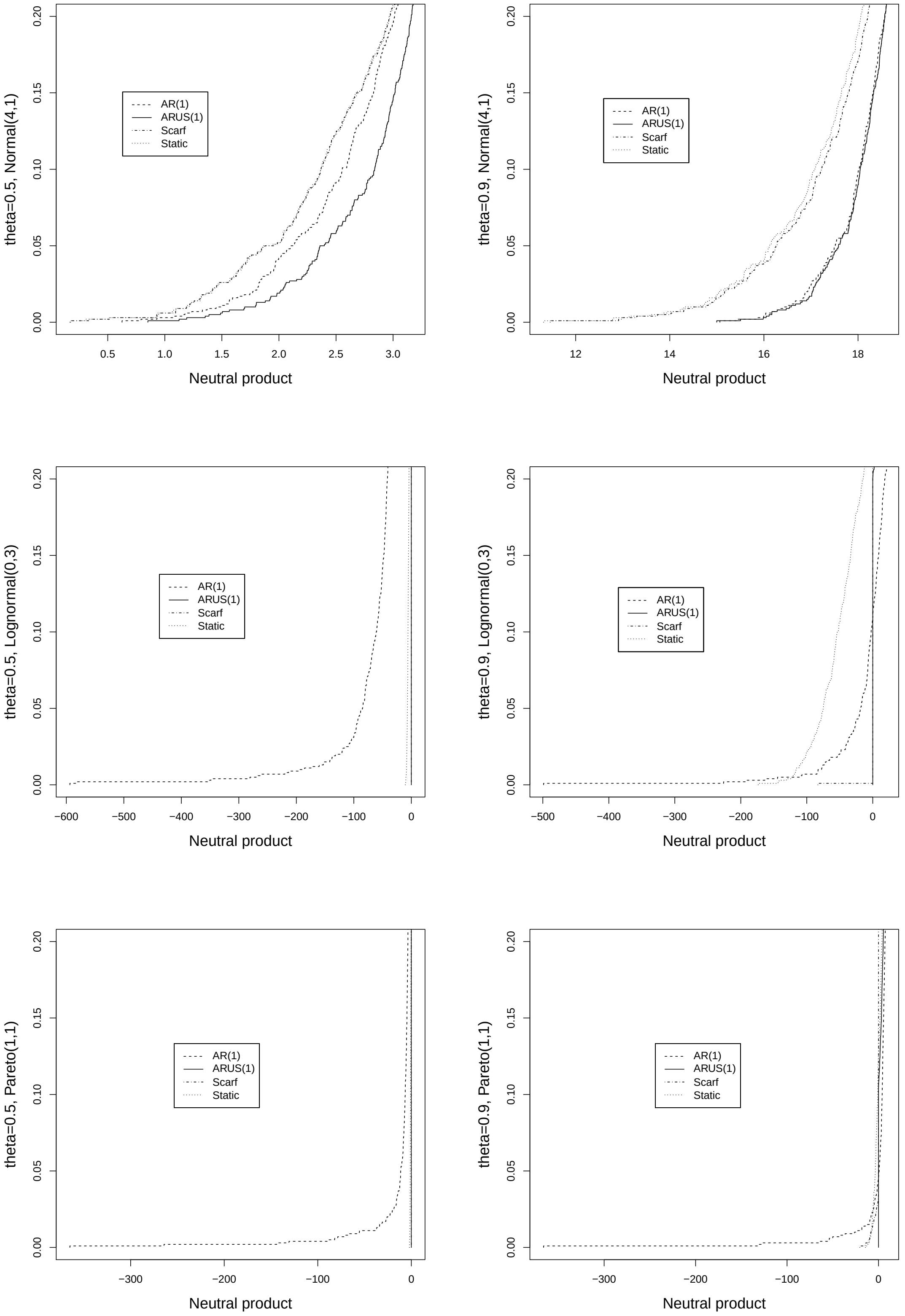


Figure 4.2: Empirical cdf of the revenue under the static (dotted line), $AR(1)$ forecasting method (dashed line), Scarf approach (dot-dashed line) and the $ARUS(1)$ approach (solid line), under normally (top), lognormally(central) and pareto (bottom) distributed error terms in the case of neutral products.

4.4 Concluding remarks and extensions

In this chapter we have considered a novel approach to the classic newsvendor problem. First, we incorporate temporal dependence by assuming that the demand follows an autoregressive process, and the forthcoming demand is to be forecasted from historical data. Second, the common assumption of normally distributed error terms in AR models is not made, as a distribution free counterpart is proposed. Moreover, a robust approach is used, and a closed form of the optimal solution is derived for the case $n = 1$. The performance of the proposed approach is compared to three traditional competing methods. The results show that the robust method outperforms the other approaches in terms of average revenue and obtains better results in term of robustness. In very few occasions runs with losses are obtained.

We have briefly explored a robust extension to the multi-period counterpart. We aim to integrate our robust autoregressive approach into the robust inventory model of Bertsimas and Thiele (2006), who assume uncertainty intervals for the demand are already given. We propose to provide such intervals, whose lengths depend on the profitability of the product (or equivalently on the risk aversion of the user), by applying the robust modeling for autoregressive demands $ARUS(n)$.

We have shown how to extend the model to a robust multi-item newsvendor problem with demands autocorrelated over time by products. Specifically, we have considered demands following a $VAR(n)$ process. Extensions to more general inventory models, such as the replaceable items case, deserve further attention and careful analysis. Future prospects concerning this work would also include to formulate robust versions of more sophisticated time series models. Moreover, in the newsvendor context, shortage penalties and salvage values per unit can be considered as possible extensions to include in the robust approaches proposed here.

Chapter 5

On the Effect of the Demand Forecasting Technique on Location-Allocation Decisions

In this chapter we investigate how the demand forecasting technique, as well as other factors, such as the correlation and variability of the demand and its geographical distribution, affect the facility location-allocation decisions. To do this, we first study location problems with demands correlated along time and between clients. Future demands are assumed to be uncertain but following a Vector Autoregressive process (VAR) (see expression (1.8) of Chapter 1), and the robust problems are defined. Theoretically, we show these robust problems are equivalent to a minimization problem with a regularization term over the coefficients of the VAR and the costs, and we also obtain some sensitivity results.

Secondly, in order to develop a numerical study we focus on a more simple model in which demands amongst clients are supposed to be independent. We analyze a location problem in which one single perishable product is to be produced at a number of facilities whose location is sought, to be later shipped to a set of users. The demand of the clients, defined by an uncertain autoregressive time series, is forecasted from historical data. Empirically, we show that: (i) embedding the forecasting technique into the location decision problem may lead to a facilities-clients allocation which does not necessarily correspond to the minimum cost allocation, but produces better revenues, (ii) taking into account the correlation of the demand can significantly improve the performance of the supply chain, and (iii) a highly variable demand may lead to different location and production decisions.

5.1 Introduction

Demand uncertainty affects decisions in supply chain management, at both strategic (location decisions) and tactical (production and inventory management) levels, see for example Snyder (2006); Melo et al. (2009). Although the demand forecast relies on the chosen forecasting technique, to the authors' knowledge no research has been devoted to analyze the effect of the choice of the forecasting technique on the location-allocation decisions.

Demand uncertainty has been addressed in the location and inventory literature by either considering a scenario analysis or assuming a perfect knowledge of the demand's distribution function (cdf) (see Snyder (2006); Gulpinar et al. (2013); An et al. (2014) in the location field, Levi et al. (2008); Aviv (2003, 2002); Shang and Song (2003); Wang et al. (2012); Reyman (1989) in inventory, and Barahona and Jensen (1998); Snyder et al. (2007) in joint location-production models). In particular, in some works the demand is modeled via a specific stochastic model as the normal distribution (Atamturk et al., 2012; Coullard et al., 2002; Liao et al., 2011; Snyder et al., 2007; Shen and Qi, 2007; Shen et al., 2003; Shu et al., 2005) or the Poisson process (Ozsen et al., 2009; Wang et al., 2007). All these approaches are contrary to the recommendations given by Bandi and Bertsimas (2012), who show that assuming a known cdf of the demand may be prohibitive and inaccurate in practice. Also See and Sim (2010); Bandi and Bertsimas (2012) show the drastic errors in the inventory policy caused by an inaccurate estimation of the demand from sample data.

Note that assuming a known cdf for the demand may not only lead to quite inaccurate results, but also implies that the consecutive demand values are considered as independent, an assumption that be may unrealistic in practice, see Lee et al. (2000); Graves (1999); Kahn (1987). In the inventory literature some authors have studied models with time-correlated demand; see, for instance, Dong and Lee (2003); Lu et al. (2006); Wang et al. (2012); See and Sim (2010); Aviv (2003); Dogan and Goetschalckx (1999); Snyder et al. (2007); Aghezzaf (2005). In particular, autoregressive processes have been widely used to model the demand (see Aviv (2002); Reyman (1989); Johnson and Thompson (1975)). A different approach is undertaken by Shahabi et al. (2014), who propose a location-inventory approach when demands are correlated amongst clients.

Despite most robust location problems are based in either distributional assumptions over the clients' demands or in scenario analysis, lately a new approach is raising: nominal values for the future demands are assumed to be given and uncertainty structures are built taking into account those values (see, for instance, Baron et al. (2011)). Although authors like to think no statistical procedures are involved in these approaches, the truth is that those nominal values for the demand must have been obtained somehow relying in prediction or estimation methods, and the user has no control over it. In the theoretical work developed here the parameters of the VAR are assumed to be given instead of those nominal values. The VAR definition (1.8) together with the information of the provided coefficients will be used to model uncertainty sets for the future demands. Moreover, Baron et al. (2011) proposes a multi-period setting in which the demand at future time instant t is assumed to belong to a l_1 or l_2-norm uncertainty set. Even although those uncertainty sets may be constructed taking into account the correlation of the demand, they might allow realizations that do not preserve such a correlation. In contrast, it will be seen that the feasible paths of the demands within the uncertainty sets that are proposed in this work manage to preserve the inner behaviour of the demands.

Several papers have been devoted to study the impact of taking into account demand correlation in supply chain management (Helper et al., 2010; Güllü, 1997). Moreover, much of this literature model the demand as an $AR(1)$ process (Ganesh et al., 2014; Raghunathan, 2003; So and Zheng, 2003). However, rather than studying the impact of updating the demand forecasts over the design of the supply chain, they state the benefits from considering demand correlation in a supply chain which is already set. An exception applies to Güllü (1997), which studies the effect of demand forecasting in the allocation using a Martingale Model of Forecast Evolution. The approach undertaken in the empirical study of this work is different, as we also analyze the impact of the demand prediction method over the location and allocation solutions.

In this chapter we first model robust location problems where we consider demands correlated not only amongst clients but also along time. Second, we focus on a more simple setting to empirically analyze the effect of the (time-correlated) demand forecasting technique in location and production decisions. Our test environment is a basic discrete p-facility location model in which the open facilities are to produce one single, perishable product. Demand at different time periods from different users is assumed to follow an uncertain autoregressive process, in the sense that the parameters defining the process are unknown, as well as the probability distribution of the error terms. The demand is forecasted from historical data. The problem we are addressing accomplishes three tasks: select the plants to be opened, allocate clients to such facilities, and decide the quantity of perishable product each plant must produce for each client at each single period. Our aim is to compare how different strategies for forecasting the demand affect such decisions.

The chapter is structured as follows. Next section introduces some robust locations problems with uncertain future demands following a VAR process. Some theoretical results are derived. In Section 5.3 we develop a data-driven approach for a location model, where the demand is estimated using a forecasting technique. This approach is used to analyze the effects of such forecasting method on the location-allocation decisions. The design of the experiments is addressed in Section 5.4. Specifically, the data sets used to illustrate the approach are introduced in Section 5.4.1, Section 5.4.2 deals with the stochastic processes that have been used for generating the demand time series values, and Section 5.4.3 reviews the demand forecasting techniques to be considered in the numerical experiments. Section 5.5 is devoted to present the obtained numerical results under two perspectives. First, we analyze the effects of the demand forecasting technique when considering the proposed approach. Second, the previously obtained results are compared to those obtained under a minimum cost allocation criterion. Finally, last section presents some remarks and extensions.

5.2 Robust location problems with Vector Autoregressive demands

Most robust location problems are based on either distributional assumptions over the clients' demands, or on scenario analysis. However, lately a new approach is becoming quite popular. It consists on constructing uncertainty sets around nominal values for future demands, assumed to be given. Nevertheless, the user may have no control over the estimation methods that are used to obtain those nominal values for the demand. Moreover, although either a statistician or an expert must have already harvested information about the behaviour of the demand in order to select a proper prediction method, unfortunately the user can incorporate such knowledge about the demand directly into the optimization problem only by making use of the estimated nominal values. For instance, consider the ambiguity sets (1.2) proposed by Bertsimas and Thiele (2006), discussed in Chapter 1.

A benchmark paper in this context is Baron et al. (2011), which proposes a multi-period setting in which the demand is assumed to belong to a l_1 (box) or l_2-norm (elliptical) uncertainty set. However, the work developed here presents substantial differences from Baron et al. (2011). First, nominal values for the future demand are assumed to be given in their approach, while here the parameters of the VAR are assumed to be given instead. The VAR definition (1.8) together with the information of the provided coefficients will be used to model uncertainty sets for the future demands. Finally, their demand realization for time t does not depend on the realization of the demand for time $t-1$; i.e., their uncertainty sets might contain paths that do not preserve the inner behaviour of the demands. This phenomenon will be illustrated in Section 5.2.1, as well as the feasible paths of the demands within the uncertainty sets that are proposed in this work.

Assume we possess the historical demands $\mathbf{X}_1, ..., \mathbf{X}_T$, where $\mathbf{X}_t = \left(X_t^1, ..., X_t^N\right)'$ are the demands of all clients at time instant t. We assume the series $\{\mathbf{X}_t\}_{t\geq 0}$ follows a VAR(n) of parameters $A_1, ..., A_n \in \mathbb{R}^{N\times N}$, $\boldsymbol{\alpha} \in \mathbb{R}^N$, i.e., it can be written as in (1.9). Our planning horizon is $T+1, ..., T+h$, and although demands $\mathbf{X}_{T+1}, ..., \mathbf{X}_{T+h}$ are uncertain they can be expressed in terms of a VAR(n) model for t in the planning horizon. Therefore, the future demands are required to belong to the following uncertainty set:

$$U_{\tilde{X}} = \{\tilde{X} : \quad \Pi\tilde{X} - \tilde{\mathbf{a}} = \mathbf{b}, \quad \tilde{X} \geq 0\} \tag{5.1}$$

where $\tilde{X} = \left(\mathbf{X}'_{T+1}, ..., \mathbf{X}'_{T+h}\right)' \in \mathbb{R}^{nh}$ and $\tilde{\mathbf{a}} = \left(\mathbf{a}'_{T+1}, ..., \mathbf{a}'_{T+h}\right)' \in \mathbb{R}^{nh}$ represents the uncertain parameters. The known parameters $\Pi \in \mathbb{R}^{nh\times nh}$ and $\mathbf{b} \in \mathbb{R}^{nh}$ are:

$$\mathbf{b} = \begin{bmatrix} \boldsymbol{\alpha} + \sum_{k=1}^{n} A_k \mathbf{X}_{T-k} \\ \boldsymbol{\alpha} + \sum_{k=2}^{n} A_k \mathbf{X}_{T+1-k} \\ \vdots \\ \boldsymbol{\alpha} + A_n \mathbf{X}_{T-1} \\ \boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha} \end{bmatrix}$$

$$\Pi = \begin{bmatrix} I_{n\times n} & 0 & \dots & & & 0 \\ -A_1 & I_{n\times n} & 0 & \dots & & 0 \\ -A_2 & -A_1 & I_{n\times n} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ -A_n & -A_{n-1} & -A_{n-2} & \dots & & \\ \vdots & & & & & \\ 0 & \dots & & & & I_{n\times n} \end{bmatrix}$$

We also define the following uncertainty set for the future prediction errors $\tilde{\mathbf{a}}$:

$$U_{\tilde{\mathbf{a}}} = \{\tilde{\mathbf{a}} : \quad \|\tilde{\mathbf{a}}\| \leq \delta\} \tag{5.2}$$

where $\|\cdot\|$ is a vector norm. Note that a matrix norm can also be used here so as to bound differently the errors between products and the errors along time. For instance, consider the $((s,\Lambda),a)$-norm, defined as

$$\|Z\|_{(s,\Lambda),q} = \|\Lambda\left(\|z_1\|_q, ..., \|z_h\|_q\right)\|_s \tag{5.3}$$

where $Z = [z_1, ..., z_h]$ is a $\mathbb{R}^{nh}$ vector, and $\Lambda \in \mathbb{R}^{h\times h}$. Then the uncertainty set (5.2) is a particular case, since it can be written as $U = \{\tilde{\boldsymbol{\epsilon}} : \quad \|\tilde{\boldsymbol{\epsilon}}\|_{(s,A),q} \leq 1\}$, with $A = diag\left(\frac{1}{h}\right)$.

In the multi-period setting it is reasonable to consider as an objective function to minimize the costs along the planning horizon. This can be formulated in terms of the robust p-median problem with correlated demands as:

$$\min_{\mathbf{x},\mathbf{y}\in R} \quad \max_{\substack{\tilde{X}\in U_{\tilde{X}} \\ \tilde{\boldsymbol{\epsilon}}\in U_{\tilde{\boldsymbol{\epsilon}}}}} \quad \tilde{X}'\mathbb{F}(\mathbf{x}) \tag{5.4}$$

where $\mathbb{F}(\mathbf{x}) \in \mathbb{R}^{nh}$ represents the unit costs of serving the clients :

$$\mathbb{F}(\mathbf{x}) = \begin{bmatrix} F(\mathbf{x}) \\ F(\mathbf{x}) \\ \vdots \\ F(\mathbf{x}) \end{bmatrix}, \quad \text{and} \quad F(\mathbf{x}) = \begin{bmatrix} \sum_{i=1}^{N} c_{1i}x_{1i} \\ \sum_{i=1}^{N} c_{2i}x_{2i} \\ \vdots \\ \sum_{i=1}^{N} c_{Li}x_{Li} \end{bmatrix}.$$

Analogously, the multi-period robust UFLP can be expressed as:

$$\min_{\mathbf{x},\mathbf{y}\in R'} \mathbf{o}'\mathbf{y} + \max_{\substack{\tilde{X}\in U_{\tilde{X}}\\ \tilde{\boldsymbol{\epsilon}}\in U_{\tilde{\boldsymbol{\epsilon}}}}} \quad \tilde{X}'\mathbb{F}(\mathbf{x}) \tag{5.5}$$

The following result shows that the previous minmax optimization problems are equivalent to simple minimization problems in which the feasible regions are the same as the original non-robust programs and regularization terms over the coefficients of the VAR and the costs have been added to the objective functions.

Theorem 2. *The robust p-median problem (5.4) is equivalent to the following optimization problem:*

$$\min_{\substack{\mathbf{x},\mathbf{y}\in R\\ \boldsymbol{\lambda}\geq 0}} \quad \mathbf{b}'G(\mathbf{x},\boldsymbol{\lambda}) + \delta\|G(\mathbf{x},\boldsymbol{\lambda})\|^* \tag{5.6}$$

where $G(\mathbf{x},\boldsymbol{\lambda}) = (\Pi^{-1})'(\mathbb{F}(\mathbf{x}) + \boldsymbol{\lambda})$ *and* $\boldsymbol{\lambda} \in \mathbb{R}^{nh}$*. Analogously, the robust UFLP is equivalent to:*

$$\min_{\substack{\mathbf{x},\mathbf{y}\in R'\\ \boldsymbol{\lambda}\geq 0}} \mathbf{o}'\mathbf{y} + \mathbf{b}'G(\mathbf{x},\boldsymbol{\lambda}) + \delta\|G(\mathbf{x},\boldsymbol{\lambda})\|^*. \tag{5.7}$$

Proof. Since the future error terms can be expressed as $\tilde{\mathbf{a}} = \Pi\tilde{X} - \mathbf{b}$, problem (5.4) can be rewritten:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}\in R} \quad \max_{\mathbf{X}} \quad & \tilde{X}\mathbb{F}(\mathbf{x}) \\ \text{s.t} & \begin{cases} \tilde{X} \geq 0 \\ \|\Pi\tilde{X} - \mathbf{b}\| \leq \delta \end{cases} \end{aligned} \tag{5.8}$$

As the original problem is convex (because the objective function is linear and any norm defines a convex set) and constraints verify Slater's condition, we can dualize only the positivity, obtaining that problem (5.8) is equivalent to:

$$\begin{aligned} \min_{\substack{\mathbf{x},\mathbf{y}\in R\\ \boldsymbol{\lambda}\geq 0}} \quad \max_{\mathbf{X}} \quad & \tilde{X}\mathbb{F}(\mathbf{x}) + \boldsymbol{\lambda}\tilde{X} \\ \text{s.t} & \left\{ \|\Pi\tilde{X} - \mathbf{b}\| \leq \delta \right. \end{aligned} \tag{5.9}$$

Then, the following change of variables $\mathbf{z} = \Pi\mathbf{X} - \mathbf{b}$ is effectuated, yielding:

$$\min_{\substack{\mathbf{x},\mathbf{y}\in R\\ \boldsymbol{\lambda}\geq 0}} \quad \max_{\|\mathbf{z}\|\leq\delta} \quad (\Pi^{-1}(\mathbf{z}+\mathbf{b}))'(\mathbb{F}(\mathbf{x})+\boldsymbol{\lambda}) \equiv \min_{\substack{\mathbf{x},\mathbf{y}\in R\\ \boldsymbol{\lambda}\geq 0}} \mathbf{b}'\Pi^{-1}(\mathbb{F}(\mathbf{x})+\boldsymbol{\lambda}) + \max_{\|\mathbf{z}\|\leq\delta} \mathbf{z}'(\Pi^{-1})'(\mathbb{F}(\mathbf{x})+\boldsymbol{\lambda})$$

Note that Π is invertible since $det(\Pi) = 1$ holds because Π is lower triangular with a diagonal of ones. Now, using the definition of dual norm it is straightforward to conclude that this is equivalent to problem (5.6). The proof for the UFLP is analogous.

□

Now a sensitivity analysis is performed by the following Corollary, in which the maximum bound for the norm of errors such that the solution to the robust p-median problem is the same as the original p-median problem is given.

Corollary 1. *Let* $(\mathbf{x}^0, \mathbf{y}^0)$ *be the solution of the deterministic p-median problem* (p-median). *Then, the maximum* δ *such that the solution to the robust p-median problem* (5.4) *is still* $(\mathbf{x}^0, \mathbf{y}^0)$ *is:*

$$\delta^0 = \min_{\substack{\mathbf{x},\mathbf{y}\in R\\ \boldsymbol{\lambda}\geq 0}} \frac{\hat{\mathbf{X}}'(\mathbb{F}(\mathbf{x}) - \mathbb{F}(\mathbf{x}^0))}{\|(\Pi^{-1})'(\mathbb{F}(\mathbf{x}^0) + \boldsymbol{\lambda})\|^* - \|(\Pi^{-1})'(\mathbb{F}(\mathbf{x}) + \boldsymbol{\lambda})\|^*}, \tag{5.10}$$

where $\hat{\mathbf{X}} = \Pi^{-1}\mathbf{b}$ *is the estimation of the demand via the VAR model* (1.8).

Proof. From Theorem 2 we know that δ^0 is the solution of the following problem:

$$\begin{aligned} &\max \delta \\ &\text{s.t} \left\{ \begin{array}{ll} \displaystyle\min_{\substack{\mathbf{x},\mathbf{y}\in R\\ \boldsymbol{\lambda}\geq 0}} & \hat{\mathbf{X}}'\mathbb{F}(\mathbf{x}) + \hat{\mathbf{X}}'\boldsymbol{\lambda} + \delta\|(\Pi^{-1})'(\mathbb{F}(\mathbf{x}) + \boldsymbol{\lambda})\|^* \geq \\ & \\ & \hat{\mathbf{X}}'\mathbb{F}(\mathbf{x}^0) + \hat{\mathbf{X}}'\boldsymbol{\lambda} + \delta\|(\Pi^{-1})'(\mathbb{F}(\mathbf{x}^0) + \boldsymbol{\lambda})\| \end{array} \right. \end{aligned}$$

which is equivalent to:

$$\begin{aligned} &\max \delta \\ &\text{s.t} \left\{ \delta \leq \frac{\hat{\mathbf{X}}'(\mathbb{F}(\mathbf{x}) - \mathbb{F}(\mathbf{x}^0))}{\|(\Pi^{-1})'(\mathbb{F}(\mathbf{x}^0) + \boldsymbol{\lambda})\|^* - \|(\Pi^{-1})'(\mathbb{F}(\mathbf{x}) + \boldsymbol{\lambda})\|^*}, \quad \forall (\mathbf{x}, \mathbf{y}) \in R, \quad \boldsymbol{\lambda} \geq 0 \right. \end{aligned}$$

which is equivalent to Problem (5.10). □

Since Problem (5.10) is a MILFP, *Dinkelbach* algorithm can be applied to obtain δ^0. This is a convergent iterative algorithm such that the solution of a problem $\min_z \frac{f(z)}{g(z)}$ can be solved iteratively $z_{k+1} = \arg\min_z f(z) - g(z)\frac{f(z_k)}{g(z_k)}$.

5.2.1 Illustration of demands' feasible paths

In order to test the possible realizations of the demands within the uncertainty sets defined in Baron et al. (2011), an autoregressive series as in (1.6) and a Moving Average series (the reader is referred to Box et al. (2008) for more information about MA models) both with parameters $\alpha = 100$ and $\theta = 0.9$ and errors following a $N(0, 1)$ have been generated. The series, with 100 observations, have been split up in train and validation sets with a proportion of 80% and 20%, respectively.

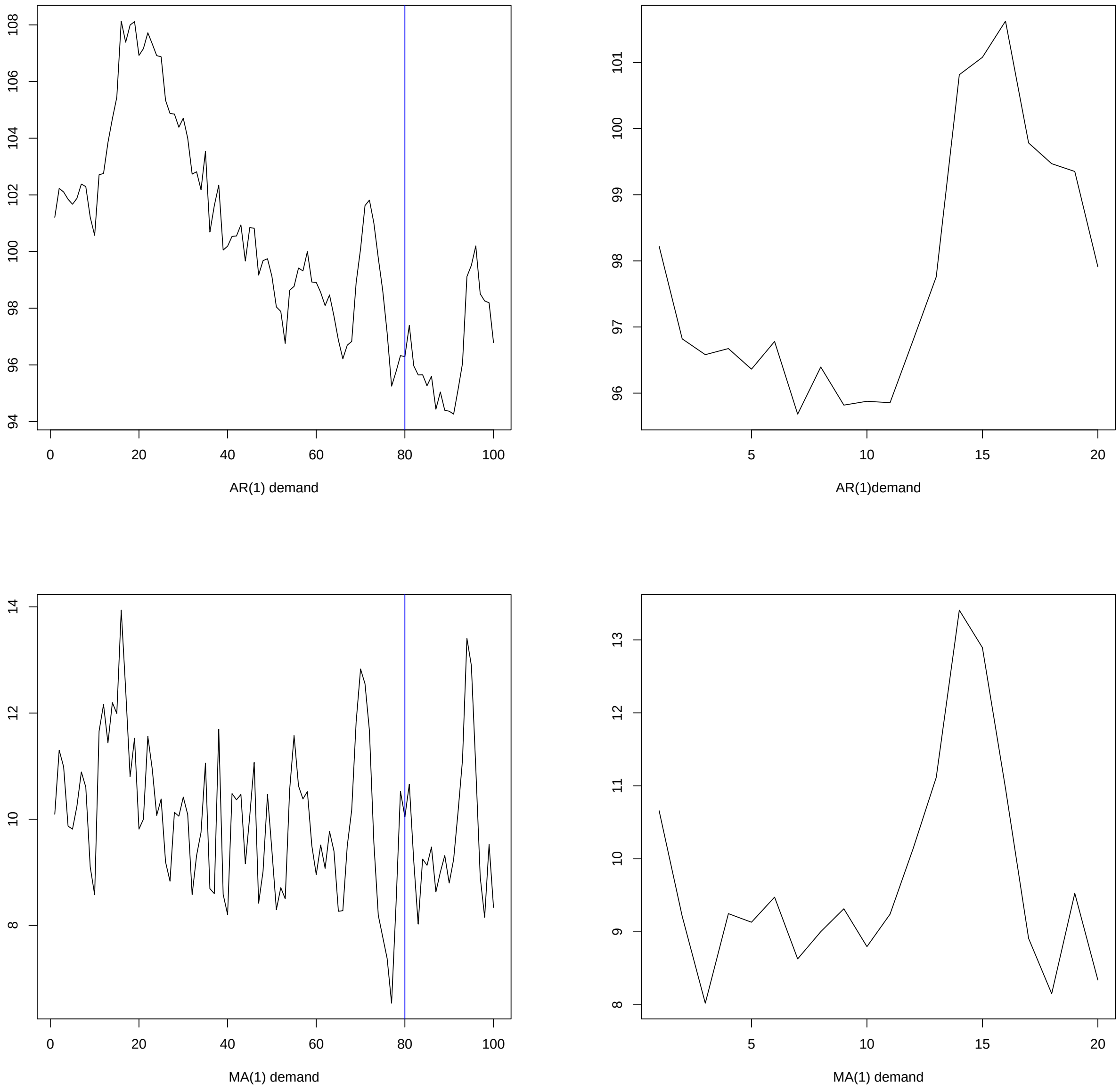


Figure 5.1: The top panels show highly correlated autoregressive demands, while the bottom panels show moving average demands, both generated with errors following $N(0,1)$ and parameters $\theta_1 = 0.9$, $\alpha = 100$. Left panel contain the whole series, right panel the *unknown* future values for the demand

Let us first focus on the box uncertainty sets proposed by Baron et al. (2011),

$$U_t = [\overline{X}_t(1-a_t), \overline{X}_t(1+a_t)], \quad t = 81, ..., 100,$$

where $\overline{X}_t$ are the nominal values for the future demand, assumed to be given. Once provided the real values of the future demand as $\overline{X}_t$, the uncertainty sets were calculated under the choice of parameters suggested in Baron et al. (2011), i.e., $a_t = \gamma + (1-\gamma)a_{t-1}$, with $\gamma = 0.15$ and $a_{80} = 0$. In order to assess the possible realizations of the demands within those uncertainty sets, values in those intervals were uniformly generated, i.e., $\hat{X}_t \sim U(\overline{X}_t(1-a_t), \overline{X}_t(1+a_t))$.

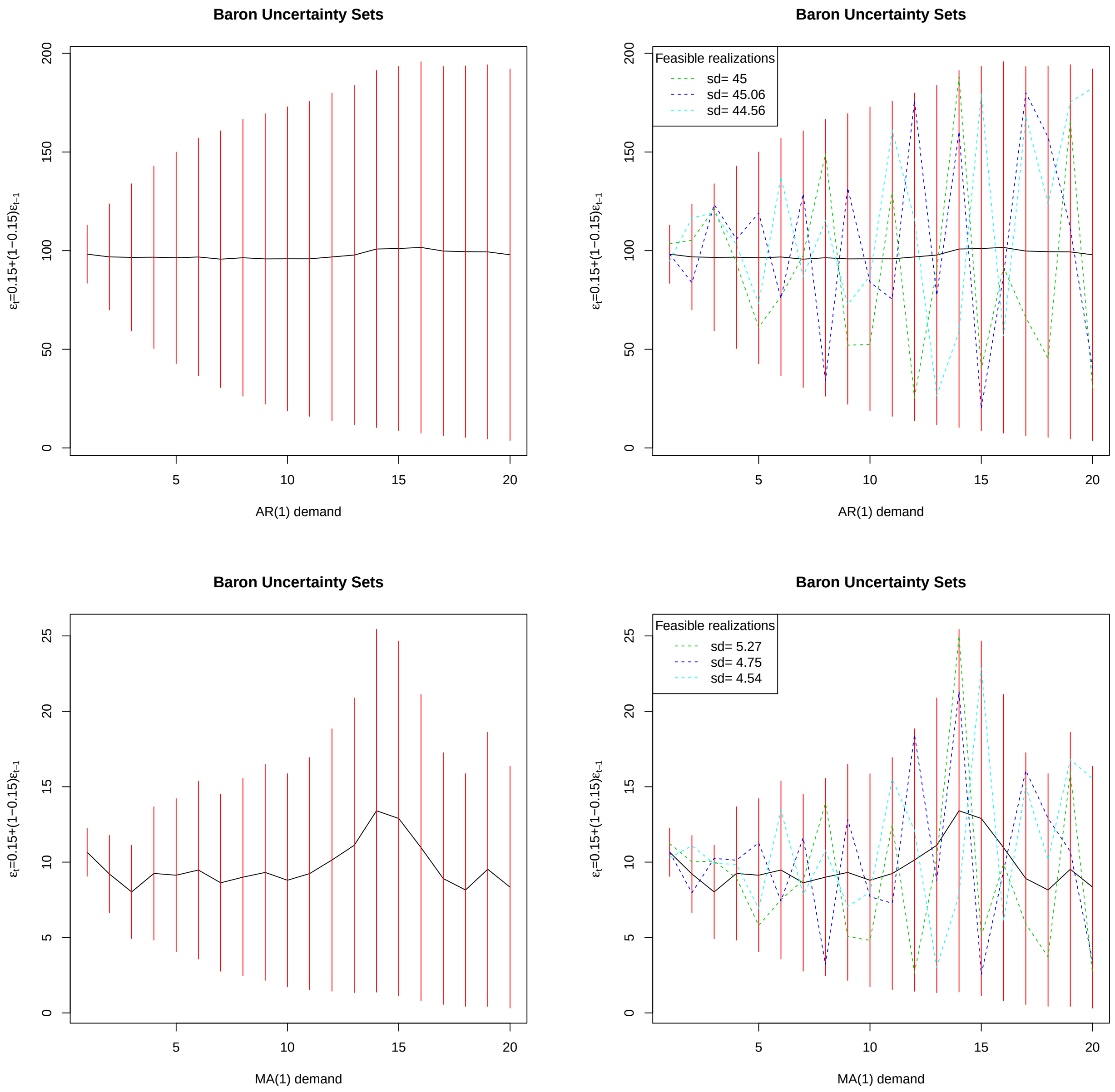


Figure 5.2: The red lines denote the interval uncertainty sets for each instant of time for AR (top panels) and MA (bottom panels) demands. On the right panel, some realizations for the demand are shown with their standard deviations

In order to test the implications of the uncertainty set (5.1) over the path of feasible demands, we estimate the coefficients $\alpha, \theta_1, .., \theta_n$ by OLS using the *known* demands

$X_1, ..., X_{80}$. To better compare with Baron et al. (2011), the next condition is required: $\|(a_{81}, ..., a_{100})\|_\infty \leq \max_{t=81,..,100} |a_t^{baron}|$. Once again, possible realizations for the demand at each period have been uniformly generated and depited in Figure 5.3. In comparison with Figures 5.2 and 5.3, the feasible realizations obtained by the proposed model seem to better maintain the inner behaviour of demand and generate less variable series, since the standard deviations obtained are closer to the real one.

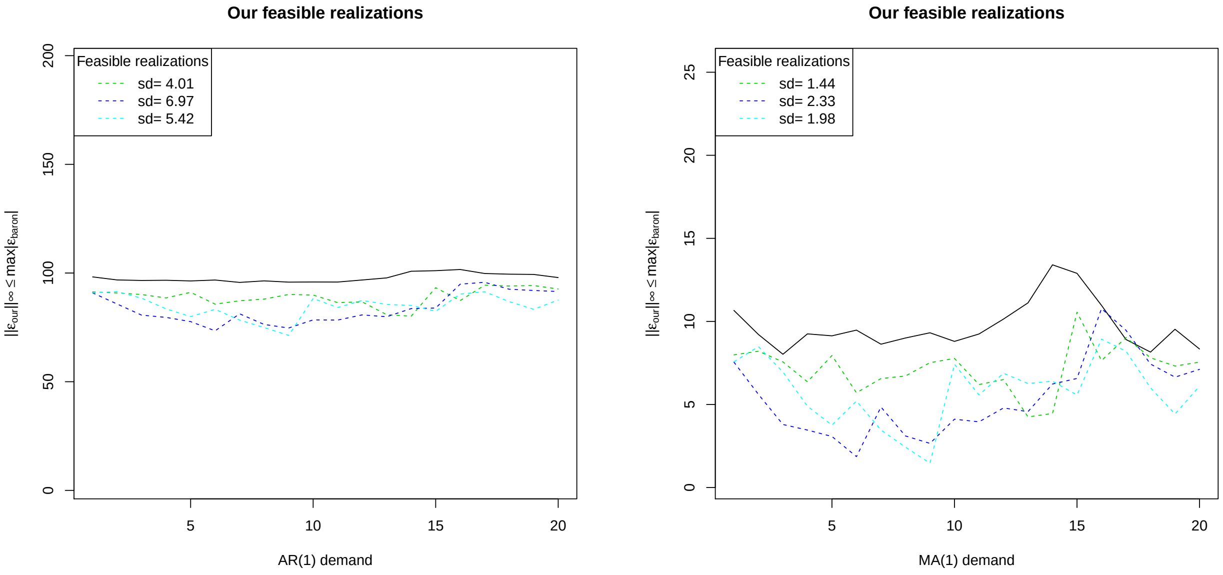


Figure 5.3: Some realizations for the demand for uncertainty sets (5.1) and (5.2), together with their standard deviations.

.

5.3 A data-driven approach for a single-period location-production model

In this section we describe a location-production data-driven procedure that shall be applied in Section 5.5 under the assumption of a correlated demand to illustrate the effect that the demand forecasting technique has on the location-allocation decisions.Due to the many paremeters to study, we have restricted to a single-item perishable product to improve the readability of the chapter. Therefore, no warehouses or stock levels are considered. The problem to be solved encompasses locating p facilities amongst the candidate locations to attend all customers, whose demand values $\{X_t^i\}_{t=1}^T$ are available up to time T. For simplicity such demand values are assumed to be independent amongst clients. Along with the location task, the quantity of products $Q_{li}(t)$ to produce at each location l to supply client i at each instant time t is to be determined. As the considered product is perishable, the demands can only be forecasted for one period ahead. For the sake of comprehension, we next present the notation used throughout the rest of this chapter.

$X^i = \{X_t^i\}_{t=1}^T \quad i \in \{1, ..., N\}$	time series of demands of clients up to time T
$Q_{li}(t) \quad l \in \{1, ..., L\}, \quad i \in \{1, ..., N\} \quad t > 0$	the quantity of product to produce at location l to supply client i, at time t
v_i	sale unit price of the perishable product for client i
$T + 1, ..., T + h$	planning horizon

If $b_{li}(t, X_t^i)$ denotes the benefit of serving client i from facility l at time t, under the assumption of a known demand X_t^i, then the benefit of serving client i from facility l per each period in the planning horizon $T + 1, ..., T + h$, is given by

$$B_{li}(T + 1, T + h, X^i) = \frac{1}{r} \sum_{t=T+1}^{T+h} b_{li}(t, X_t^i) \tag{5.11}$$

Coefficients $b_{li}(t, X_t^i)$ are deterministic when the demand X_t^i is known. Since the values of $\{X_t^i\}_{t>T}$ are not available in practice and no probabilistic model for the demand is assumed, a data-driven distribution-free recursive approach is proposed to estimate the benefit coefficients $b_{li}(t, X_t^i)$. The method consists in dividing each client's historical demand, given by a time series, into two sets: the training set (periods $1, ..., T_1$) and the testing set (periods $T_1 + 1, ..., T$). After having chosen a specific demand forecasting technique, then the train set will be used to estimate its parameters, and the test set to estimate the benefit coefficients $b_{li}(t, X_t^i)$, (and therefore, the total benefit as in Eq.(5.11)). This procedure is explained in more detail in Table 5.1.

Once the total benefit coefficients are estimated according to the procedure in Table 5.1, the location-allocation problem to be solved is formulated as the p-facility problem on (p-median) but maximizing the total profit as in (5.14)

1. Split the demand time series in train $\{X_t^i\}_{t=1}^{T_1}$ and test $\{X_t^i\}_{t=T_1+1}^{T}$ sets for all clients $i \in \{1, ..., N\}$.
2. For $t_0 = T_1 + 1, \ldots, T$ and for all $i \in \{1, ..., N\}$ repeat:
   (a) Use the known values of the demand $\{X_t^i\}_{t=1}^{t_0-1}$ to fit the parameters of the selected demand forecasting technique.
   (b) Use the (estimated) demand forecasting technique to obtain the prediction of the demand and the quantity of product to send to client i from every plant $i \in I$ at time t_0, $\hat{X}_{t_0}^i$ and $\hat{Q}_{li}(t_0)$.
   (c) Calculate the benefit coefficients as the obtained revenues
   $$b_{li}(t_0) = v_j \min\{X_{t_0}^i, \hat{Q}_{li}(t_0)\} - c_{li}\hat{Q}_{li}(t_0) \tag{5.12}$$
3. Estimate the benefit coefficients per period in the planning horizon as
$$\hat{B}_{li}(T+1, T+h) = \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} b_{li}(t) \tag{5.13}$$

Table 5.1: Recursive approach for estimating the parameters of the p-median problem.

$$
\begin{aligned}
\max \quad & \sum_{l=1}^{L}\sum_{i=1}^{N} \hat{B}_{li}(T+1, T+h)x_{li} \\
\text{s.t} \quad & \begin{cases}
\displaystyle\sum_{l=1}^{L} x_{li} = 1 & \forall i = 1, ..., N \\
x_{li} \leq y_l & \forall i = 1, ..., N, l = 1, ..., L \\
\displaystyle\sum_{l=1}^{L} y_l = p & \\
x_{li} \in \{0, 1\} & \forall i = 1, ..., N, l = 1, ..., L \\
y_l \in \{0, 1\} & \forall l = 1, ..., L
\end{cases}
\end{aligned}
\tag{5.14}
$$

Note that (5.14) is a p-median problem in which the costs (or, equivalently, the benefits $\hat{B}_{li}$) not only depend on the transportation and production costs (c_{li}^S and c_i^P), but they also depend on $\hat{Q}_{li}$, which is obtained through demand estimation. Hence, it is intuitively obvious that an erroneous forecast of the demand may strongly perturb the coefficients $\hat{B}_{li}$, and therefore the location and allocation decisions obtained by (5.14) may significantly differ from the solution of the standard p-median problem. This phenomenon will be clearly illustrated in Section 5.5.

5.4 Description of the experiments

In this section we describe in detail the design of the numerical experiments whose results are presented in Section 5.5. Specifically, the data sets, the stochastic models for the demand, the considered demand forecasting techniques as well as the structure of the profitability are discussed here.

5.4.1 Data sets

The data sets used in the experiments developed here closely follow those in Daskin and Owen (1999): the 55-node data set of Swain (1971) and the 88-node data set of Daskin (1995). The 88-node data set represents the fifty largest cities in the United States along with the state capitals, and its distance matrix was calculated using great circle distances, whereas for the Swain network the closest paths between nodes were calculated using Floyd's approach. Average demands for the U.S. cities data set were given, as for the nodes of the Swain network they were randomly generated following a uniform distribution, as done in Blanquero et al. (2014).

The number of facilities to be opened was also chosen accordingly to Daskin and Owen (1999): they tested the two data sets for $p = 3, ..., 15$. For the sake of abbreviation, we only considered the minimum, maximum and median of such set, therefore we will test the approach for three different values of p, namely $p = 3, 9, 15$.

5.4.2 Stochastic models for the demand

In the real data sets presented in Section 5.4.1, the nodes' demands $\{X_t^i\}_{t=1}^{T+h}$ were unavailable (only the average demands were known) and therefore we needed to simulate them artificially. Since we aim to study how the correlation and variability of the demand affect to the location-allocation decisions, we will assume that the demand follows an (uncertain) autoregressive process. Recall that the demand has been often modeled as an autoregressive process (Aviv, 2002; Reyman, 1989; Johnson and Thompson, 1975). However, any other stochastic process considering correlated demands may also be tested by the proposed data-driven approach.

In our experiments demands were generated from autoregressive processes as in (1.6) with $n = 1$, in such a way that the expected value of the series matches the average known values. We choose $\alpha^i = \alpha$ and $\theta_k^i = \theta_k$ for all $j \in J$. The choice $n = 1$ follows from Chapter 4, where proposed a robust forecasting technique for uncertain $AR(n)$ processes is proposed. Within that chapter, only results for $n = 1$ were included, due to the performance of the forecasting techniques being virtually unaffected by changing the value of n. The effect of different properties, as the distribution of errors $\{a_t^i\}_{t>0}$ (which determines the variability of the demand time series) and the strength of the correlation was analyzed. Two different distributions for the error terms were tested in our experiments: the normal and the lognormal distributions. First we considered standard normally distributed errors, as we aimed to test the performance under the commonly considered $AR(n)$ model. In addition, we found it of interest to check the effects of the forecasting techniques under a high variability in the demand time series, as done in Huh et al. (2011); Carrizosa et al. (2014). To such aim, the lognormal distribution $LN(0, 3)$ was selected. Another aspect to be considered when simulating the data is the strength of the temporal dependence. In our experiments, two values of $\theta_1 = \theta$ were set. Note that for $n = 1$ the coefficient θ_1 represents the lag-1 autocorrelation coefficient thus, in order to test the methods on highly and moderately correlated time series, $\theta_1 = 0.9$ and $\theta_1 = 0.5$ were fixed. Figure 5.4 shows a simulated demands time series generated under the two choices of errors distributions and correlation strengths.

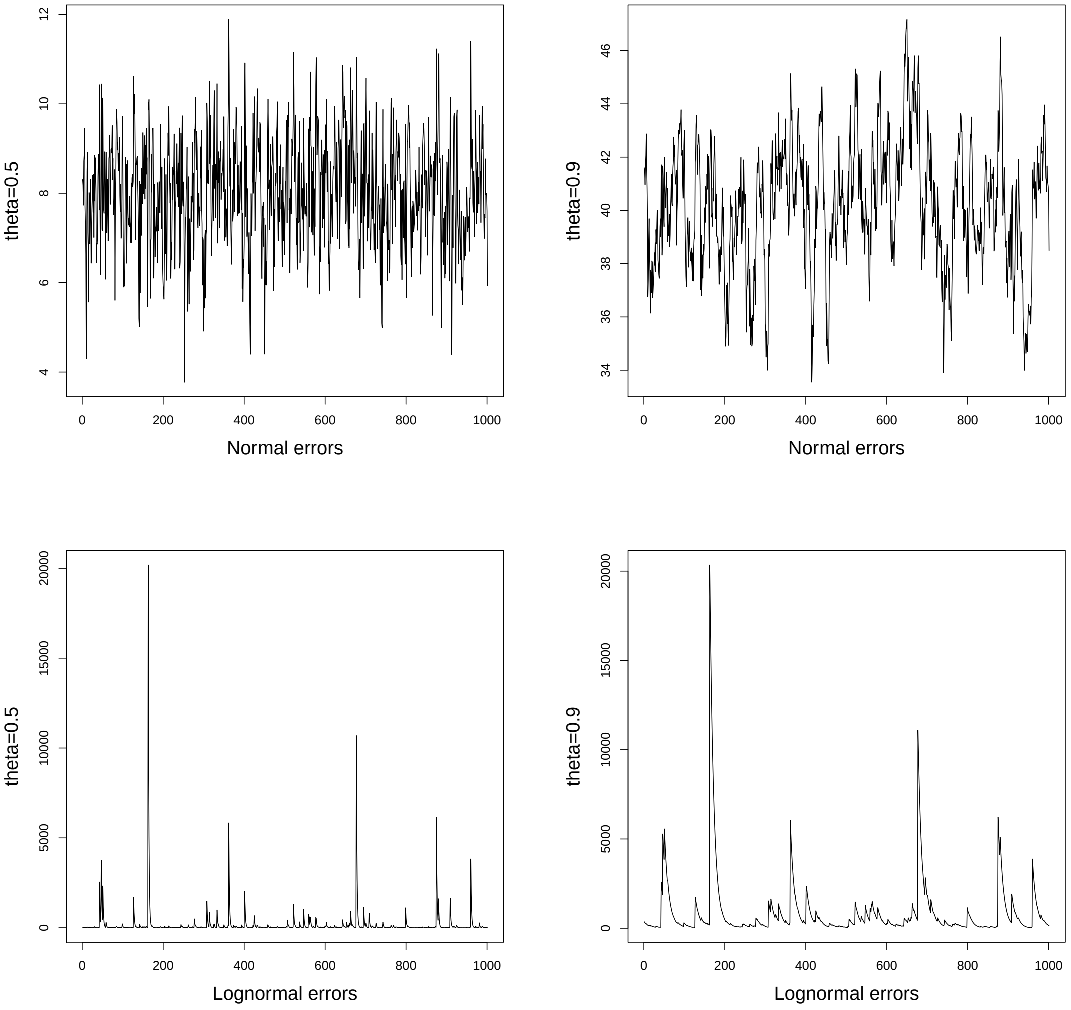


Figure 5.4: Examples of high and low correlated autoregressive demands generated with errors following $N(0, 1)$ and $LN(0, 3)$.

5.4.3 Forecasting techniques for the single-period inventory policy

Suppose the set L_0 of open facilities is already chosen, as well as the clients-facilities allocation. At each time period t we want to cover the demand of customer $i \in \{1, ..., N\}$, X_t^i, by plant $l \in L_0$. However, only the historical and possibly correlated demands for $t = 1, \ldots, T$ are available, thus a demand forecasting technique for estimating X_t^i, with $t > T$, needs to be considered. In order to address demands generated from the $AR(n)$ process, some of the demand forecasting techniques tested in Chapter 4 will be considered in our numerical experiments. Such forecasting techniques are the so called static, the classic $AR(n)$ approach and the robust $ARUS(n)$ approach.

5.4.4 Profitability structure

Note that the quantity of products $\hat{Q}_{li}(t)$ to produce at plant l to supply client i estimated via the forecasting techniques described in Section 5.4.3 depends on the total cost defined c_{li} and on the sale price v_j. Hence it is expected that the supply chain performance is affected by the profitability of the products, as the forecasting methods themselves do. Therefore, two types of profitability settings were considered for the perishable product: low- and high-profitability. For the sake of abbreviation we fix $v_j = v$ for all clients. For low-profit products, the sale price was stablished as

$$v = 1.25 \max_{\substack{l=1,\ldots,L \\ i=1,\ldots,N}} \{c_{li}^S, c_i^P\} \tag{5.15}$$

where c_{li}^S are the normalized transport costs, which are derived from normalizing the distance matrix, and c_i^P are the so as normalized production costs. The high-profit products are defined as in (5.15) but with a higher value:

$$v = 1.75 \max_{\substack{l=1,\ldots,L \\ i=1,\ldots,N}} \{c_{li}^S, c_i^P\}$$

5.5 Numerical results

In this section, the results obtained are compared under the different forecasting techniques outlined in Section 5.4.3, and also, under a benchmark approach, consisting in assuming perfect knowledge of the demand at each period. Under this perfect knowledge criterion, clients are supplied with their exact demand. Two different allocation strategies will be considered:

- The maximum-profit allocation, i.e., the one given by the optimal values of variables x_{li} when (5.14) is solved.

- The minimum-cost allocation, i.e., the one that allocates the clients to their minimum cost open facility ($x_{li} = 1$ if and only if $l = \arg\min_{l \in L_0} c_{li}$).

We consider of interest to compare the results with the classic minimum cost allocation as it does not take into account the demand of the clients and it has been reported to be very restrictive (Weaver and Church, 1986). It will be shown that the allocation decisions under the proposed data-driven approach do not necessarily coincide with that under the minimum cost strategy, and usually better total revenue values are obtained.

In our experiments 25 instances of the problem were simulated for each data set. Series of length $T+h = 288$ following an $AR(1)$ process with i.i.d errors were generated for each node. The train set is defined for time instants $t = 1, ..., 240$, while the test set considers the demand from $t = T_1 + 1 = 241$ to $t = T = 264$. The validation set is composed of the rest of time series values $\{X_t^i\}_{t=T+1}^{T+h}$, so the train set has a length equal to 240 and both the test and validation sets are of length 24. To manage the supply chain according to the proposed location-production approach, each demand forecasting technique is applied in the validation set where the location ($\{y_l\}_{l \in \{1,...,L\}}$) and allocation ($\{x_{li}\}_{l \in \{1,...,L\}, i \in \{1,...,N\}}$) solutions are obtained from the train and test sets, and where the total revenue for each instance as in (5.11) is computed similarly as in Table 5.1 using the validation set.

In the next section the influence of the forecasting technique, variability and correlation in the demand, type of profitability and geographical position of the nodes is analyzed under the maximum profit criterion. Section 5.5.2 will be devoted to examine in detail the effect of considering instead the minimum cost allocation. All the optimization problems were solved using `Xpress 7.2`.

5.5.1 Results under the maximum profit criterion

Figures 5.5 and 5.6 show the box plots of the 25 total revenues for all considered location-production approaches under the maximum profit criterion. The demand in all cases is generated from an $AR(1)$ process with normal $N(0, 1)$ and lognormal $LN(0, 3)$ errors, respectively. Note that this implies that a low (respectively, high) dispersion in the realizations of the demand is expected in Figure 2 (respectively, Figure 3). Our findings showed few differences between the behavior of the location approach under different number of plants to be opened. Therefore, to abbreviate only the case with $p = 3$ is illustrated here. The location-production strategies are named as the associated forecasting method, i.e., for instance $STATIC$ 0.9 denotes the location approach in which the demand has been generated with parameter $\theta = 0.9$ and forecasted by (1.10).

The results of the first set of experiments are presented in Figure 5.5. In this case, in which the $N(0, 1)$ model governs the distribution of the errors in the autoregressive process for the demand, the samples concentrate in an interval of small length, specially for low correlation. In addition, total revenues distribute similarly among the different forecasting techniques for a given value of θ. The best results are obviously obtained under the benchmark approach, as perfect knowledge of the demand is assumed. Concerning the static approach, a better performance is achieved for $\theta = 0.5$. Slightly better values, close to those of the benchmark approach, are obtained when the classic AR model is considered, a reasonable result since the AR process assumes normal errors. The results are similar for the 88-node data set (top part of Figure 5.5), though much worse results are obtained for $ARUS$. From this experiment it can be concluded that, when the demand variability is low, the results are highly dependent on both the spatial distribution of the nodes (since the obtained results are different for the two data sets) and the correlation in the demand. However, the effect of the forecasting technique and the type of profitability seems to be weaker.

Now we consider the variability of the samples represented by the box plots of Figure 5.6. As expected, the heavy-tailed behavior in the demand implies a high variability in the results obtained for both Swain and 88-node data sets, in particular under the benchmark approach. For the rest of approaches the variability increases with the correlation parameter θ. Also, it can be seen that the results for the 88-node data set are more dispersed than those for the Swain network, especially for high-profit products. Consider now the effect of the different forecasting techniques on the revenues. Similarly as in the previous situation, better results are obtained under the benchmark approach, this outperformance being more evident for low correlated demands. The performance of the static approach is similar in all cases. However, significant differences are observed for the combination of the location problem with autoregressive forecasting approaches. Although the robust approach never obtains losses, the model with the classic AR produces negative total revenues when $\theta = 0.5$ and for high-profit products, and also for low-profit products in the case of the 88-node database. Therefore in this case, where extreme values for the demand are considered, the performance of the supply chain under the maximum profit allocation criterion is influenced not only by the spatial distribution of the data and correlation in the demand, but also by the forecasting technique. The effect of the profitability seems to be negligible for the 88-node data sets. However, for the Swain data the type of profitability affects the results under the AR forecasting method.

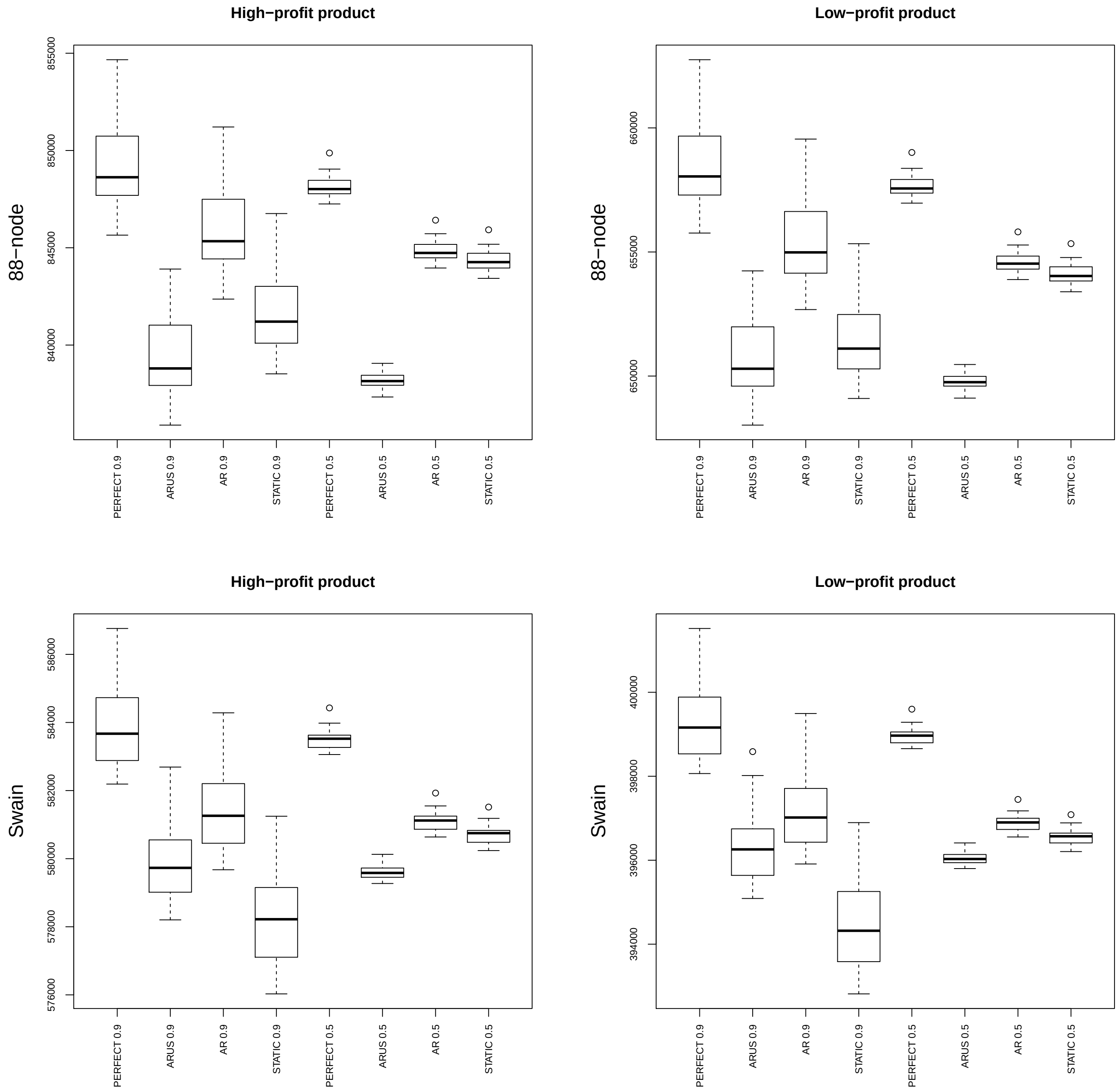


Figure 5.5: Box plots of the obtained total revenues for the 88-node and Swain databases under the maximum profit criterion. The nodes' demands have been generated from an $AR(1)$ process with errors distributed according to a $N(0, 1)$ model.

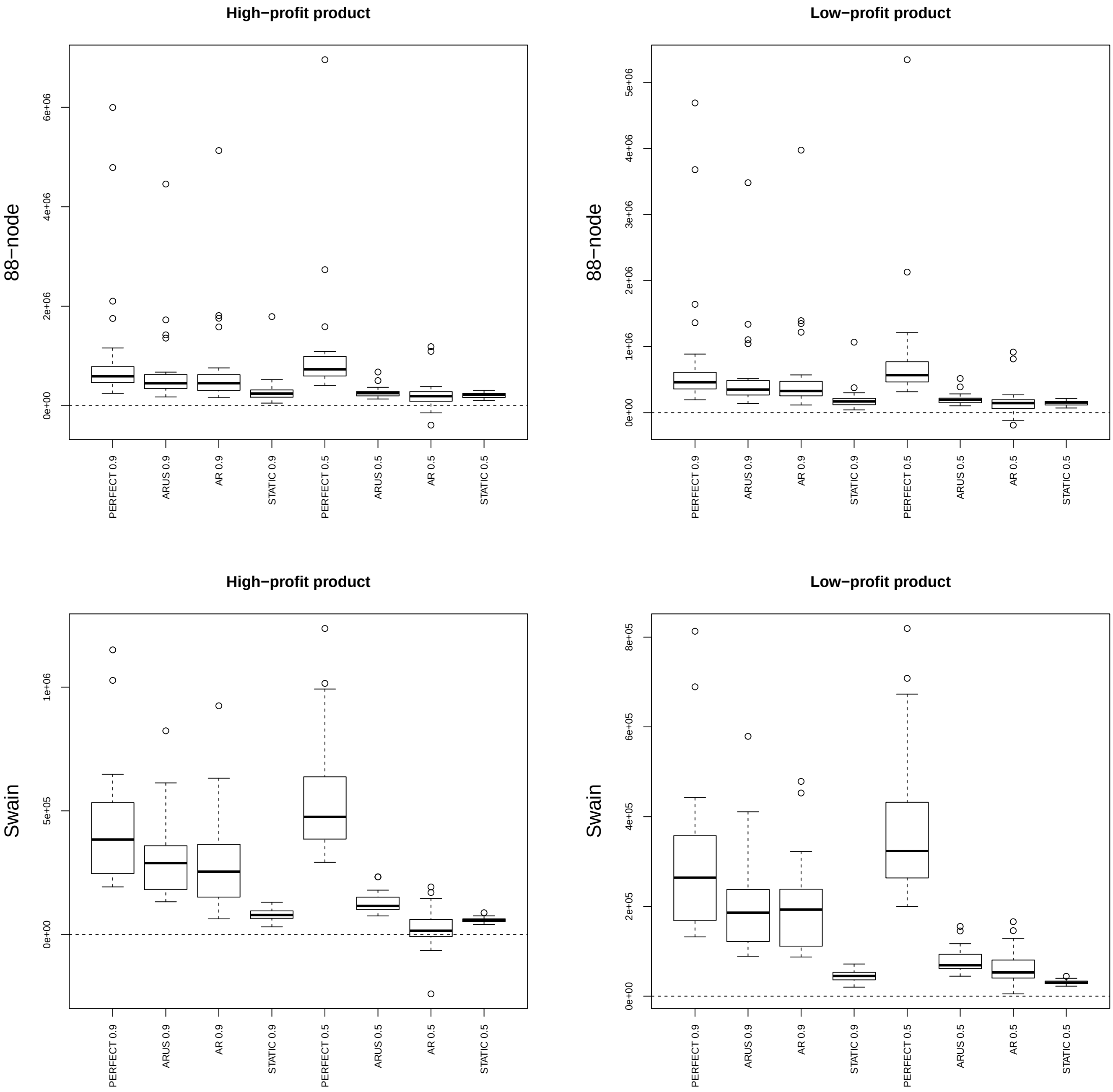


Figure 5.6: Box plots of the obtained total revenues for the 88-node (up) and Swain (bottom) databases under the maximum profit criterion. The nodes' demands have been generated from an $AR(1)$ process with errors distributed according a $LN(0,3)$ model.

5.5.2 The effect of the allocation

As commented at the beginning of Section 5.5, two location-allocation approaches have been considered in our experiments, namely, the one that considers maximum profit allocation of clients according to the forecasting technique, and the one that assigns clients to the minimum cost open facility. The results shown in the previous section were obtained under the maximum profit criterion. In this section, the effect of considering the alternative approach is analyzed.

Tables 5.2 and 5.3 report the average percentage of improvement of the minimum cost allocation over the maximum profit allocation, for the Swain and 88-node datasets, respectively. In order to test the effect of the number of plants to be opened, three values of p, $p = 3, 9, 15$, were considered in the experiment. In addition, since the results from Section 5.5.1 under the $N(0, 1)$ model were analogous, in this case the errors are assumed to be lognormally distributed.

		$\theta = 0.9$			$\theta = 0.5$		
Profitability	Forecasting	$p = 3$	$p = 9$	$p = 15$	$p = 3$	$p = 9$	$p = 15$
High	STATIC	2.62	1.91	1.24	6.18	8.06	8.94
	AR	-18.53	-18.62	-17.61	-8002.84	-1059.04	-938.30
	$ARUS$	0.98	1.30	1.27	2.15	2.30	2.51
Low	STATIC	3.11	4.11	4.12	6.79	8.42	8.88
	AR	-11.55	-17.61	-17.15	-374.73	-406.14	-380.50
	$ARUS$	1.62	0.94	0.75	2.84	2.83	3.11

Table 5.2: Average percentage of improvement of the minimum cost allocation over the maximum profit allocation for the **Swain** databases. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0, 3)$ model.

From Tables 5.2 and 5.3 it can be observed that the worst results are obtained under the classic autoregressive forecasting technique. For the AR, the minimum cost criterion always worsens the total revenue of the supply chain, especially for the Swain database. Note that in this case, when $p = 3$, $\theta = 0.5$ and for high profitability, the obtained loss is greater than 8000% of the total revenue. Note also that the average percentage of improvement gets worse as the autocorrelation of the demand decreases. The $ARUS$ is the least affected forecasting method since its average percentage of improvement varies between 0.07% and 3.11%. On the other hand the results under the static forecasting technique are more variable: for the Swain network it is the forecasting method that best responds to the change of allocation, but for the 88-node database and a high correlation this change results in a worsening of the total revenue between 10.66% and 18.31%.

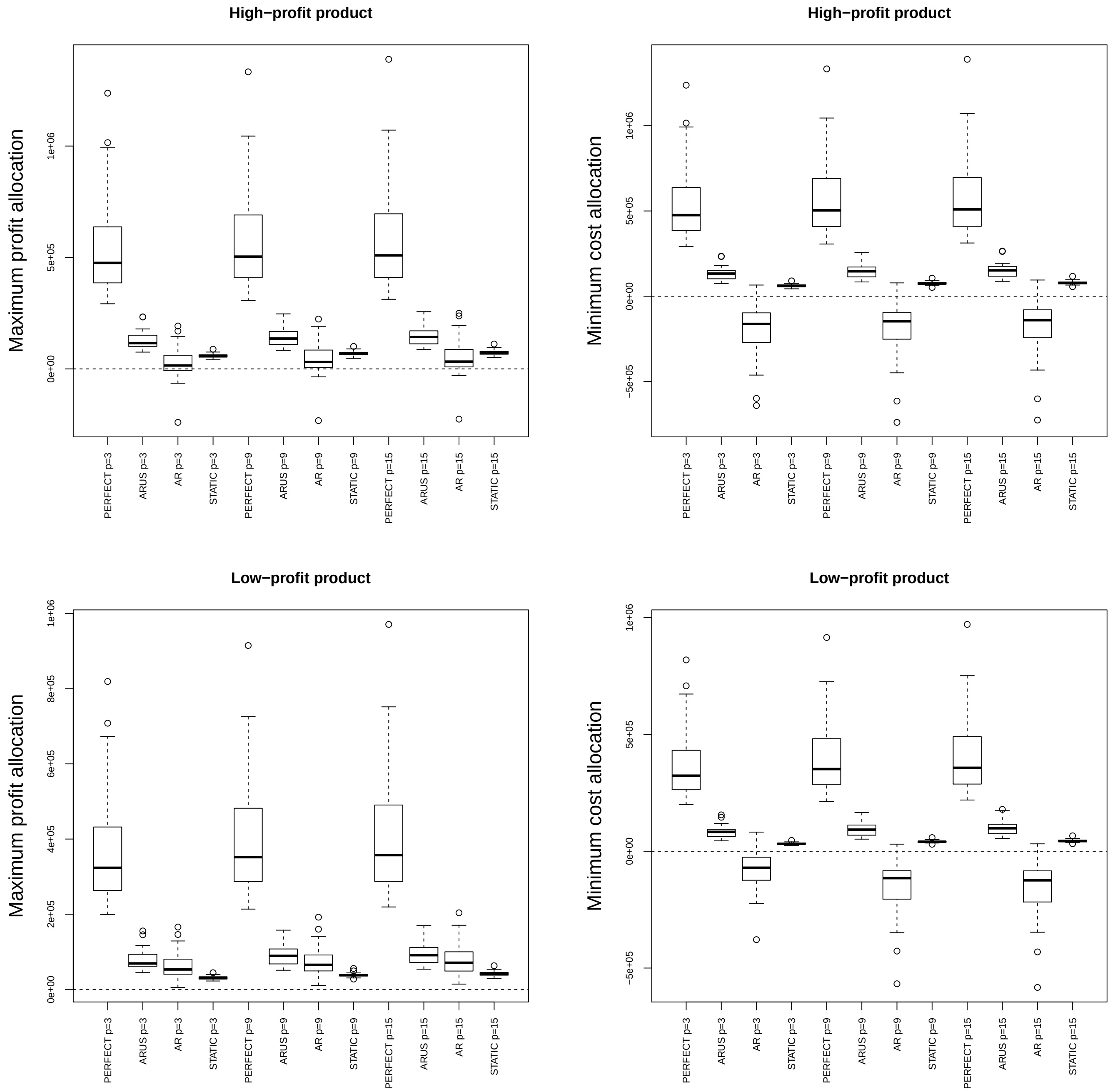
		$\theta = 0.9$			$\theta = 0.5$		
Profitability	Forecasting	$p = 3$	$p = 9$	$p = 15$	$p = 3$	$p = 9$	$p = 15$
High	STATIC	-13.45	-17.81	-18.31	1.29	1.81	2.42
	AR	-10.40	-8.94	-8.17	-123.60	-133.66	-178.07
	$ARUS$	0.10	0.08	0.12	0.08	0.07	0.08
Low	STATIC	-10.66	-11.89	-13.93	1.31	1.55	1.95
	AR	-11.71	-11.14	-10.63	-201.53	-146.40	-131.90
	$ARUS$	0.11	0.10	0.14	0.09	0.08	0.10

Table 5.3: Average percentage of improvement of the minimum cost allocation over the maximum profit allocation for the **88-node** databases. The nodes' demands have been generated from an $AR(1)$ process with errors distributed according a $LN(0, 3)$ model.

In conclusion, although a minimum cost allocation may be thought as a sensible criterion, the previous findings show that it can lead to worse revenues than those obtained under the maximum profit strategy. Moreover, substantial changes in the results are observed for the different data sets, forecasting techniques and correlation values. However, the behavior under different values of p does not seem to be affected.

In order to investigate in more detail the effect of the number of plants to be opened and the influence of the location-allocation approach, we depict in Figure 5.7 the box-plots of the revenues samples for different settings in the Swain data set. In particular, the results for $p = 3, 9, 15$, in combination with high and low profitability, the different forecasting techniques as well as the two considered location-allocation strategies are shown. In all cases the value of θ is set to 0.5 (since from Table 5.2 this is the case which produces the highest worsening) and the errors in the demand's autoregressive process are lognormal $LN(0, 3)$.

Several conclusions are obtained from Figure 5.7. First, with the exception of the benchmark approach, whose results improve with the value of p, the rest of forecasting techniques are unaffected by this parameter. Second, the type of profitability apparently has a minor effect when the minimum cost allocation is considered. Finally, the results under the AR forecasting method combined with the minimum cost allocation strategy differ significantly from those obtained under the rest of forecasting techniques for both allocation criteria. In order to better understand the reason why the minimum cost allocation criterion behaves differently under different forecasting techniques, we report in Table 5.4 the average percentage of clients (for $p = 3, 9, 15$) that were assigned with minimum cost when the maximum profit criterion was assumed. It is expected that the higher this percentage is, the less impact the minimum cost allocation criterion has.



High−profit product

High−profit product

Low−profit product

Low−profit product

Figure 5.7: Box plots of the obtained total revenues for the Swain database under both types of allocation and all values of p. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0,3)$ model and $\theta = 0.5$.

		Swain		88-node	
Profitability	Forecasting	$\theta = 0.9$	$\theta = 0.5$	$\theta = 0.9$	$\theta = 0.5$
High	STATIC	66.57	81.96	71.95	81.52
	AR	64.80	26.74	89.59	67.80
	ARUS	93.19	88.82	96.83	96.82
Low	STATIC	68.78	84.51	71.32	83.12
	AR	56.19	18.62	86.89	59.92
	ARUS	92.02	87.10	96.74	96.50

Table 5.4: Average percentage of clients allocated to the minimum cost open facility for the 88-node databases. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0, 3)$ model.

As observed, the robust approach $ARUS$ is the one which assigns more clients with minimum cost. In most cases all forecasting methods perform better for the 88-node database than for the Swain network. It is interesting to note that, unlike the autoregressive approaches AR and $ARUS$, the performance of the static forecasting technique improves when the correlation decreases. For the Swain network, AR never assigns more than 65% of clients to their nearest facility. Moreover, for low-correlated demand this percentage strongly decreases: for low-profit products only the 18.62% of clients are assigned with minimal cost, and this case corresponds to the worst average percentage of improvement derived from applying minimum cost allocation instead of the original assignment of clients in Table 5.2.

In order to better understand this phenomenon, we depict in Figures 5.8-5.9 the locations of the opened plants for the case $p = 3$ and the assignment of clients in the 88-node data set for two randomly selected instances (top and bottom panels) under both allocation criteria (left and right panels). In all cases the demand is low-correlated and heavy-tailed, and final total revenue is displayed for each case. Figures 5.8 and 5.9 show the results for high and low profit product, respectively.

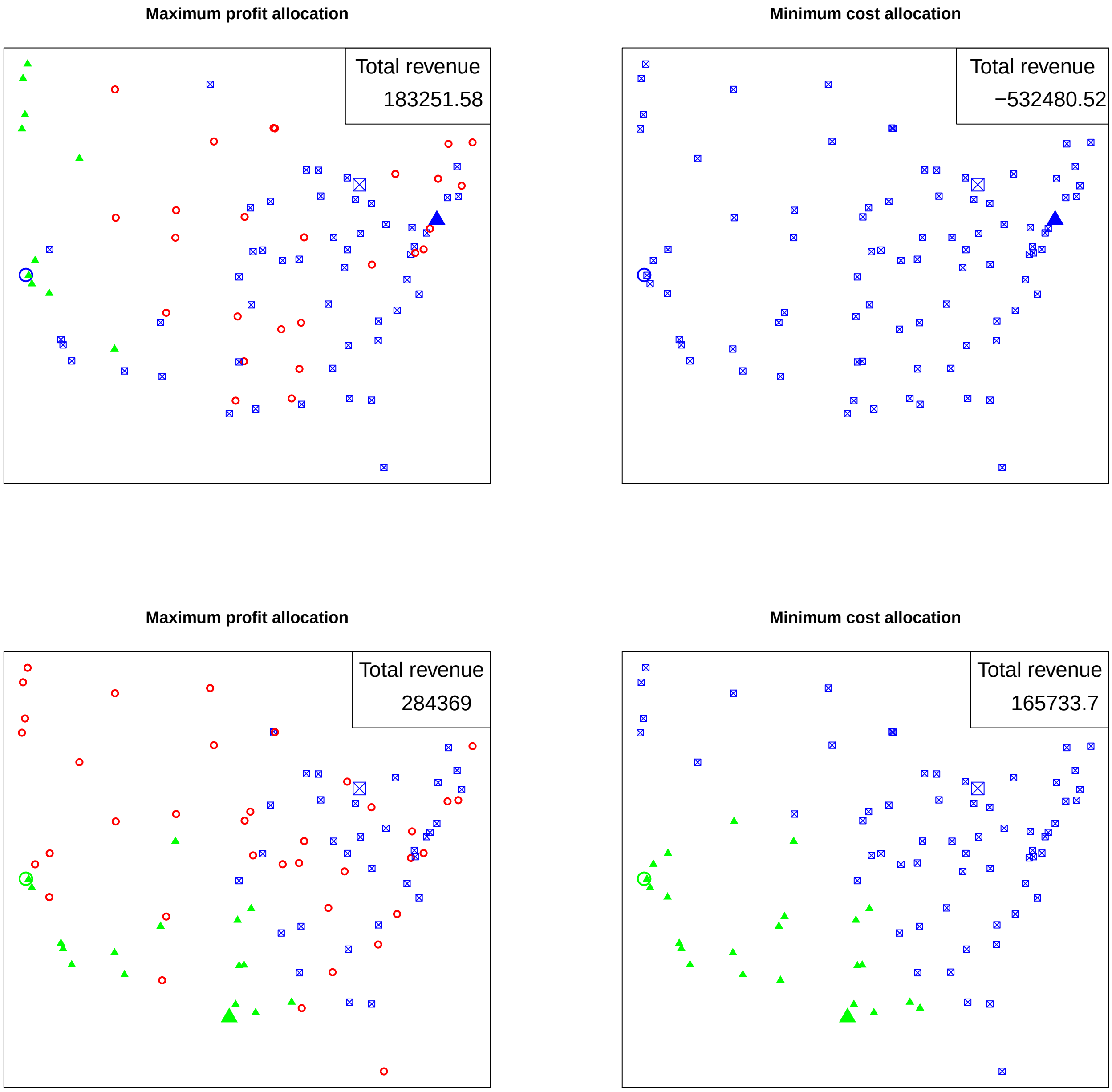


Figure 5.8: Location and allocation of clients for AR forecasting technique of two randomly selected instances for **high-profit** products in the 88-node database under both criteria: maximum profit and minimum cost (production+transportation) allocation. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0,3)$ model and $\theta = 0.5$.

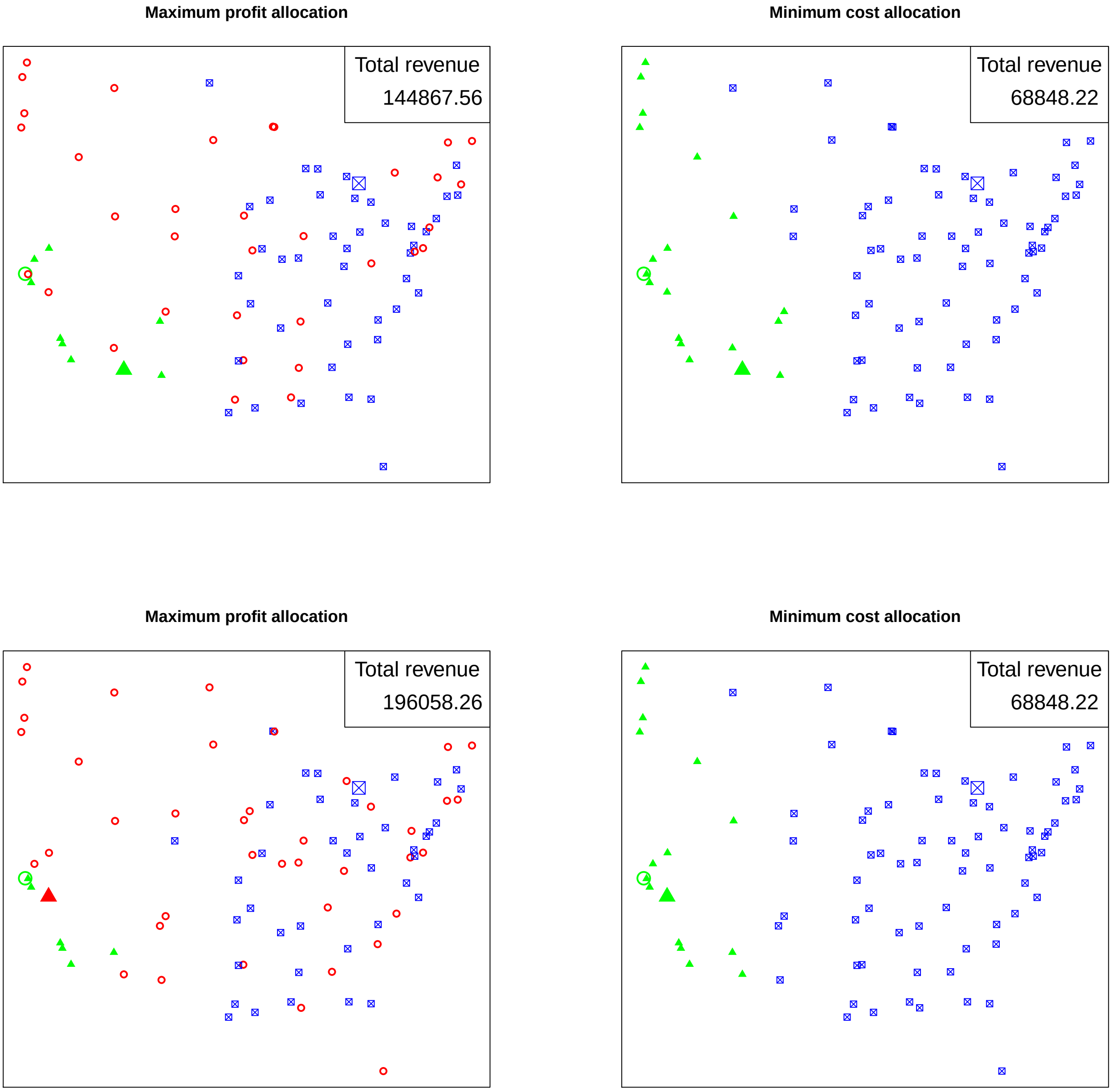


Figure 5.9: Location and allocation of clients for AR forecasting technique of two randomly selected instances for **low-profit** products in the 88-node database under both criteria: maximum profit and minimum cost (production+transportation) allocation. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0, 3)$ model and $\theta = 0.5$.

The three open facilities are represented within the figures by either a large solid green triangle, an empty red circle or a crossed blue square. Their assigned clients have been represented by smaller but analogous symbols with the same color. It may be the case that an open facility is not served by itself but by another open plant. In such a case, only the color (not the symbol) changes to denote which plant will supply it.

Figures 5.8-5.9 show a number of interesting issues. In particular:

1. First, the minimum cost allocation is usually suboptimal. Second, in some cases the clients are allocated to their furthest open facility, and geographically close clients may be optimally allocated to facilities far from each other. Moreover, we have observed that the facilities opened may have high production costs.

2. The optimal location of plants may vary according to the realization of the demand (see the plant denoted by a triangle in both figures).

3. In some cases the plants are not served by themselves (see the plant denoted by a circle in the left panel of both figures). This is due to extremely high production costs: it is more profitable to be served by a plant with higher transportation costs but lower overall expenses.

4. Under the minimum cost allocation, some open plants do not serve any client (see the plant denoted by a circle in the right panels of both figures).

These results may be a consequence of the behavior observed in Carrizosa et al. (2014): the AR approach performs better for low-profit products than for high-profit products. Hence, the AR forecasting method produces the requirement for location-allocation solutions that increase the costs of the problem (thus diminishing the profitability) by opening facilities with high production costs and serving clients that are far away. This will usually lead to a better joint performance of the location-production solution. From this experiment it can be concluded that the forecasting technique deeply influences the location-allocation solution.

5.6 Concluding remarks and extensions

In this chapter we have analyzed the impact of numerous factors over the location-allocation decisions, with special emphasis on the effects of the demand forecasting technique. Theoretically, we have proposed robust location problems considering demands correlated along time and between clients. Results stating the tractability and sensitivity of such models have been proved. Empirically, several conclusions have been obtained from the experiments considered. Firstly, the variability of the demand strongly affects the performance of the joint location-production model, in such a way that an unsuitable choice of the demand forecasting technique can lead to losses, mainly when the demand is affected by a strong variability. Secondly, the geographical distribution of the nodes as well as the correlation of the demand also influences the performance of the joint location-production model too. Therefore, as shown by our numerical experience, taking into account the correlation of the demand when managing the supply chain can significantly improve the final revenue. Thirdly, when the cost is optimized, some clients may not be allocated to their minimum cost open facility. Furthermore, the number of clients optimally allocated with minimal cost depends on the forecasting technique chosen for the proposed joint location-production model. In conclusion, in order to improve the total revenue of a supply chain the choice of the demand forecasting technique must take into account the correlation and variability of the demand.

All conclusions obtained in our experiments are constrained to one single facility model (p facilities, with p fixed) and strong assumptions of demand distribution (demands of different clients are considered as independent, and thus the spatial correlation is ignored). Furthermore, the simplest inventory model has been used in order not to add more degrees of freedom to the problem. Extensions to more realistic models, including multi-item inventory or non-perishable storable products deserve further study. Moreover, it would be interesting to empirically test the performance of the proposed robust location problems with Vector Autoregressive demands agains the robust counterparts existent in the current literature, with special interest on the comparison with Baron et al. (2011).

Chapter 6

A Robust Perspective on Transaction Costs in Portfolio Optimization

In this chapter we show how to use a transaction cost term in a portfolio optimization problem to compute portfolios that are both efficient in terms of transaction costs and robust to estimation error. Theoretically, we prove that the portfolio problem with transaction costs is equivalent to three different problems designed to alleviate the impact of estimation error: a robust portfolio optimization problem, a robust regression problem, and a Bayesian portfolio problem. Motivated by these results, we propose a data-driven approach to portfolio optimization that calibrates the transaction cost term taking into account both transaction costs and estimation error. Our empirical results demonstrate that the data-driven portfolios perform favorably because they strike an optimal trade-off between rebalancing the portfolio to capture the information in recent historical return data, and avoiding the large transaction costs and impact of estimation error associated with excessive trading. Essentially, the data-driven portfolios induce an slow and steady rate of trade that results in superior performance.

6.1 Introduction

Markowitz (1952) showed that an investor who cares only about the mean and variance of portfolio returns should choose a portfolio on the efficient frontier. Mean-variance portfolios continue to be the workhorse of much of the investment management industry, but two crucial aspects in their successful implementation are estimation error and transaction costs. In this chapter we show that these two aspects are intimately related, and use this relation to propose a data-driven approach to compute portfolios that are both efficient with respect to transaction costs and resilient with respect to estimation error.

Estimation error is important because to implement the mean-variance portfolios in practice one has to estimate the mean and the covariance matrix of asset returns, and it is well known that the resulting portfolios often perform poorly out of sample; see Michaud (1989); Chopra and Ziemba (1993); Broadie (1993), and DeMiguel et al. (2009b). A popular approach to alleviate the impact of estimation error is to use robust portfolio optimization; see, for instance, Goldfarb and Iyengar (2003); Garlappi et al. (2007), and Lu (2011a,b). This approach captures the uncertainty about the mean and covariance matrix of asset returns by assuming they may lie anywhere inside the so-called uncertainty sets. The robust portfolio is the one that maximizes the mean-variance utility with respect to the worst-case mean and covariance matrix of asset returns.

Another approach to combat estimation error in portfolio selection is to use Bayesian portfolios. The tenet behind this approach is that the investor has a prior belief about the distribution of asset returns. Then the investor uses Bayes rule to combine her prior belief with the observed data and obtain the portfolio that maximizes the mean-variance utility with respect to the posterior distribution. The prior distribution can be diffuse Barry (1974); Bawa et al. (1979), empirical Jorion (1986), or based on an asset pricing model Pástor (2000); Pástor and Stambaugh (2000) .

Transaction costs are important because they can easily erode the gains from a trading strategy. For instance, DeMiguel et al. (2014) show that the gains from a trading strategy that exploits serial dependence in stock returns do not survive even modest proportional transaction costs of ten basis points. Transaction costs can be generally modelled with the p-norm of the portfolio trade vector. For small trades, which do not impact the market price, the transaction cost is generally assumed to be proportional to the amount traded, and thus it can be approximated by the 1-norm of the portfolio trade vector. For larger trades, the literature has traditionally assumed that they have a linear impact on the market price, and thus they result in quadratic transaction costs that are captured by the 2-norm. Finally, several authors have recently argued that market impact costs grow as the square root of the amount traded (Torre, 1997; Almgren et al., 2005; Farmer et al., 2013; Frazzini et al., 2015), and thus they are captured by the p-norm with $p = 1.5$.

We make a theoretical contribution and an empirical contribution. The theoretical contribution is to show that estimation error and transaction costs are intimately related. Specifically, we show that the portfolio optimization problem with p-norm transaction costs can be equivalently reformulated as three different portfolio problems designed to alleviate the impact of estimation error: (i) a robust portfolio optimization problem, (ii) a robust regression problem, and (iii) a Bayesian portfolio problem. First, we show that the portfolio optimization problem with p-norm transaction costs is equivalent to a robust portfolio optimization problem where the mean of asset returns can take any value in an uncertainty set defined by the q-norm, where $1/p + 1/q = 1$. For instance, a mean-variance portfolio problem with proportional (quadratic) transaction costs can be equivalently rewritten as a robust portfolio optimization problem where the mean can take any value in an uncertainty set defined by a box (hypersphere) centered at the nominal mean return.

Second, we show that the portfolio optimization problem with p-norm transaction costs is equivalent to a robust regression formulation of the mean-variance problem. It is well known that the mean-variance portfolio optimization problem can be equivalently reformulated as a linear regression problem; see, for instance, Britten-Jones (1999). We demonstrate that including a p-norm transaction cost term in the mean-variance portfolio optimization problem is equivalent to solving a robust linear regression problem where the historical returns may take any value in an uncertainty set defined by a certain matrix norm. Third, we show that the portfolio optimization problem with p-norm transaction costs is equivalent to a Bayesian portfolio problem where the investor has a prior belief that the portfolio weights are distributed as a Multivariate Exponential

Power Distribution centered around the starting portfolio.

Our theoretical results demonstrate that incorporating a p-norm transaction cost in the mean-variance optimization problem may help to reduce the impact of estimation error. Therefore it is reasonable to assume that one could *calibrate* the transaction cost parameter to produce portfolios that are not only efficient in terms of transaction costs, but also resilient to estimation error. From a real-world perspective[1], combating estimation error by calibrating the transaction cost term has two advantages. First, a transaction cost term has a natural economic interpretation, and this facilitates the task of selecting a reasonable range of parameters to calibrate from. Second, practitioners are used to incorporate transaction cost terms in their portfolio selection frameworks, and thus it may be easier for them to simply calibrate the transaction cost parameter instead of using a more sophisticated approach based on uncertainty sets or prior beliefs. These observations motivate our second contribution.

Our empirical contribution is to propose a data-driven approach to portfolio selection with estimation error and transaction costs. Concretely, we propose using cross-validation to calibrate the transaction cost parameter and compute portfolios that perform well in a realistic scenario with both estimation error and transaction costs. We compare the out-of-sample performance of the proposed data-driven portfolios on five empirical datasets with that of the mean-variance portfolios that ignore transaction costs, as well as mean-variance portfolios that capture the nominal proportional transaction costs. We find that the proposed data-driven portfolios outperform the traditional portfolios in terms of their out-of-sample Sharpe ratio net of transaction costs. The data-driven portfolios perform well because they calibrate the transaction cost parameter to achieve intermediate levels of turnover that strike an optimal (data-driven) trade-off between two goals: (i) rebalancing the portfolio to capture the information in recent historical return data, and (ii) avoiding the large transaction costs and impact of estimation error associated with excessive trading.

The remainder of this manuscript is organized as follows. Sections 6.2, 6.3, and 6.4 show that the portfolio problem with transaction costs is equivalent to a robust portfolio problem, a robust regression problem, and a Bayesian portfolio problem, respectively. In Section 6.5, we propose the data-driven approach to portfolio selection and evaluate its out-of-sample performance. Section 6.6 concludes.

The notation followed in this chapter is analogous to that introduced in Section 1.3.2.

[1] Fabozzi et al. (2007) review the practical modelling aspects of portfolio optimization and management.

6.2 The robust optimization problem

This section shows that the mean-variance portfolio problem with p-norm transaction costs can be equivalently rewritten as a robust portfolio optimization problem where the mean asset returns can take any value in an uncertainty set defined by the q-norm, where $1/p + 1/q = 1$. (Gotoh and Takeda, 2011, Proposition 1) provide a similar result for the case without transaction costs. We adapt their result and provide interpretation for the case with transaction costs.

We first introduce some basic definitions and preliminary results. For a given vector norm $\|\cdot\|$, its dual norm $\|\cdot\|^*$ is:

$$\|\mathbf{x}\|^* = \max_{\|\mathbf{y}\|\leq 1} \mathbf{y}^T\mathbf{x}.$$

It is easy to show that the dual norm of the p-norm is the q-norm, where $1/p+1/q = 1$; see (Higham, 2002, Section 6.1). Let Λ be a symmetric and positive definite matrix; we define the (p, Λ)-norm of vector $\mathbf{x}$ as $\|\mathbf{x}\|_{p,\Lambda} := \|\Lambda x\|_p$. The following lemma shows that $\|\cdot\|_{p,\Lambda}$ is indeed a vector norm and characterizes its dual norm.

Lemma 4. *The (p, Λ)-norm, $\|\cdot\|_{p,\Lambda}$, is a vector norm and $\|\cdot\|_{q,\Lambda^{-1}}$ is its dual norm, where $1/p + 1/q = 1$.*

Proof. To prove that $\|\cdot\|_{p,\Lambda}$ is a norm we exploit the fact that $\|\cdot\|_p$ is a norm:

1. $\|\mathbf{x}\|_{p,\Lambda} = \|\Lambda\mathbf{x}\|_p \geq 0$.

2. $\|\alpha\mathbf{x}\|_{p,\Lambda} = \|\alpha\Lambda\mathbf{x}\|_p = |\alpha| \cdot \|\mathbf{x}\|_{p,\Lambda}$.

3. $\|\mathbf{x} + \mathbf{y}\|_{p,\Lambda} = \|\Lambda(\mathbf{x} + \mathbf{y})\|_p \leq \|\Lambda\mathbf{x}\|_p + \|\Lambda\mathbf{y}\|_p = \|\mathbf{x}\|_{p,\Lambda} + \|\mathbf{y}\|_{p,\Lambda}$.

We now prove that $\|\cdot\|_{q,\Lambda^{-1}}$ is the dual norm of $\|\cdot\|_{p,\Lambda}$; that is, we prove that

$$\|\mathbf{y}\|_{q,\Lambda^{-1}} = \max_{\|\mathbf{x}\|_{p,\Lambda}\leq 1} \mathbf{y}^T\mathbf{x}. \tag{6.1}$$

Let $\tilde{\mathbf{y}} = \Lambda^{-1}\mathbf{y}$ and $\tilde{\mathbf{x}} = \Lambda x$. Because $\tilde{\mathbf{y}}^T\tilde{\mathbf{x}} = \mathbf{y}^T\mathbf{x}$, Equation (6.1) is equivalent to:

$$\|\tilde{\mathbf{y}}\|_q = \max_{\|\tilde{\mathbf{x}}\|_p\leq 1} \tilde{\mathbf{y}}^T\tilde{\mathbf{x}}. \tag{6.2}$$

Equation (6.2) holds because the q-norm is the dual of the p-norm. □

The following proposition gives our main result.

Proposition 3. *For every risk-aversion parameter* $\gamma > 0$ *and transaction cost parameter* $\kappa \geq 0$*, there exists* $\delta > 0$ *such that the mean-variance problem with p-norm transaction costs, Problem* (1.13)*, is equivalent to the following robust optimization problem:*

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{\gamma}{2}\mathbf{w}^T\Sigma\mathbf{w} - \mu^T\mathbf{w} + \max_{\hat{\mu}\in U(\delta)} (\mu - \hat{\mu})^T(\mathbf{w} - \mathbf{w}_0), \\ s.t. \quad & \mathbf{w}^T\mathbf{1}_N = 1, \end{aligned} \tag{6.3}$$

where the uncertainty set for mean asset returns is

$$U(\delta) = \{\hat{\mu} : \|\mu - \hat{\mu}\|_{q,\Lambda^{-1}} \leq \delta\}.$$

Proof. From Lemma 4, it follows that the third term in the objective function of Problem (6.3) can be rewritten as :

$$\max_{\hat{\mu}:\|\mu-\hat{\mu}\|_{q,\Lambda^{-1}}\leq\delta} (\mu - \hat{\mu})^T(\mathbf{w} - \mathbf{w}_0) = \delta\|\mathbf{w} - \mathbf{w}_0\|_{p,\Lambda}. \tag{6.4}$$

This implies that Problem (6.3) is equivalent to the following problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{\gamma}{2}\mathbf{w}^T\Sigma\mathbf{w} - \mu^T\mathbf{w} + \delta\|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p, \\ \text{s.t.} \quad & \mathbf{w}^T\mathbf{1}_N = 1. \end{aligned} \tag{6.5}$$

Note, however, that the p-norm transaction cost term in Problem (1.13) is raised to the power of p. It is easy to show, however, that for any $\kappa \geq 0$, there exists $\delta > 0$ such that Problem (6.5) is equivalent to Problem (1.13). To see this, note that there exists $\eta \geq 0$ such that Problem (6.5) can be equivalently rewritten as:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{\gamma}{2}\mathbf{w}^T\Sigma\mathbf{w} - \mu^T\mathbf{w}, \\ \text{s.t.} \quad & \|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p \leq \eta, \\ & \mathbf{w}^T\mathbf{1}_N = 1. \end{aligned} \tag{6.6}$$

Moreover, constraint $\|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p \leq \eta$ is equivalent to constraint $\|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p^p \leq \eta^p$. Penalty term theory then guarantees under mild regularity conditions that the constraint $\|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p^p \leq \eta^p$ can be moved to the objective function multiplied by a certain penalty coefficient. Finally, there exists a $\delta > 0$ such that the penalty parameter equals κ. □

A few comments are in order. First, Proposition 3 shows that the p-norm transaction cost can be interpreted as the maximum *regret* the investor may experience (in terms of mean return) by trading from the starting portfolio $\mathbf{w}_0$ to portfolio $\mathbf{w}$, assuming the true mean belongs to the uncertainty set $U(\delta)$, which is defined in terms of the dual q-norm. To see this, note that Proposition 3 essentially shows that the transaction cost term can be rewritten as

$$\kappa\|\Lambda(\mathbf{w}-\mathbf{w}_0)\|_p^p = \kappa\|\mathbf{w}-\mathbf{w}_0\|_{p,\Lambda}^p = \max_{\hat{\mu}:\|\mu-\hat{\mu}\|_{q,\Lambda^{-1}}\leq\delta}(\mu-\hat{\mu})^T(\mathbf{w}-\mathbf{w}_0),$$

where μ is the estimated mean asset return vector, and $\hat{\mu}$ is the worst-case mean asset return vector for the given portfolio $\mathbf{w}$.

Second, Proposition 3 shows that the use of the p-norm to capture transaction costs corresponds to the use of the q-norm to represent the uncertainty set for mean returns, where $1/p+1/q=1$. To understand the implications of this correspondence, assume for the sake of simplicity that $\Lambda = I$. For this case, proportional transaction costs, which model small transactions, correspond to a box-shaped uncertainty set for the mean; and quadratic transaction costs, which are used to model larger trades, correspond to a hyperspherical uncertainty set. One interpretation of this correspondence is that for small trades, the investor assumes that the uncertainty around the estimate of mean return for each asset is independent from other assets, and thus the investor models the uncertainty set for the means as a box. For large trades, the investor assumes there is a certain dependence between the uncertainty around the means of different assets, and thus the investor models the uncertainty set for the means as a hypersphere.[2]

The correspondence between different types of transaction costs and uncertainty sets for mean returns is illustrated in Figure 6.1 for the case with two assets. The left graph in Figure 6.1 depicts the level sets for a proportional transaction cost term (solid red line) and a quadratic transaction cost term (dashed green) in the space of portfolio weights. The graph also shows the starting portfolio $\mathbf{w}_0$ at the origin (blue dot), the optimal portfolio in the presence of proportional transaction costs (red square), and the optimal portfolio with quadratic costs (green triangle). The right graph depicts the corresponding uncertainty sets for the case with proportional costs (solid red line) and quadratic costs (dashed green line) in the space of mean returns. The graph also shows the nominal mean return at the origin (blue dot), the set of worst-case mean returns for proportional costs (red squares), and the worst-case mean return for quadratic costs (green triangle).

[2] The insights for the general case with $\Lambda \neq I$ are similar up to a linear transformation.

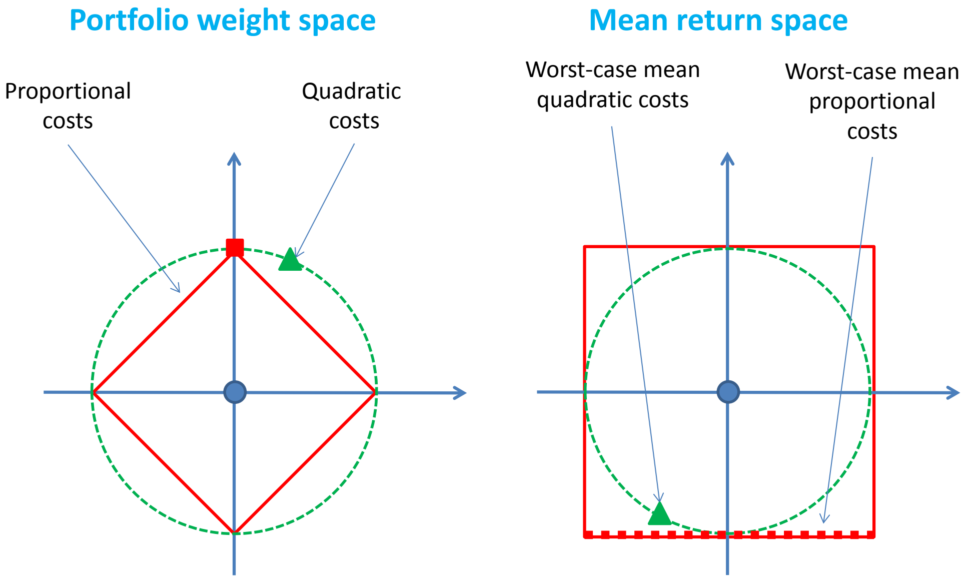


Figure 6.1: Transaction cost level sets and mean return uncertainty sets

Finally, our results in this section are related to those by Natarajan et al. (2009), who show that there is an equivalence between uncertainty sets in robust portfolio optimization and financial *risk measures*. The distinguishing feature of our analysis in this section is that we focus on the equivalence between uncertainty sets and *transaction costs* rather than risk measures.

6.3 The robust regression problem

It is well known that the mean-variance problem can be rewritten as a linear regression problem. In this section, we show that including a p-norm transaction cost term in the mean-variance problem is equivalent to robustifying the equivalent linear regression problem with respect to uncertainty in the historical return data.

This result is related to that by Caramanis et al. (2012), who show that including a p-norm regularization term in a linear regression problem is equivalent to solving a robust version of this linear regression problem. We adapt their result to the specific context of the mean-variance optimization problem with transaction costs. In particular, we consider the case where the portfolio trade vector is transformed via a matrix Λ before computing its p-norm.

Britten-Jones (1999) showed that the tangency mean-variance portfolio is the scaled solution of the following regression problem:

$$\min_{\mathbf{w}} \quad \|\mathbf{1}_T - R\mathbf{w}\|_2^2,$$

where $\mathbf{1}_T \in \mathbb{R}^T$ is the vector of ones, and $R \in \mathbb{R}^{T\times N}$ is the matrix whose columns contain the historical returns for each of the N assets. The following proposition shows how Britten-Jones' result can be extended to rewrite the general mean-variance portfolio problem with transaction costs as a linear regression problem with a regularization term.

Proposition 4. *For every risk-aversion parameter $\gamma > 0$ and transaction cost parameter $\kappa \geq 0$, there exist μ_0 and $\kappa' \geq 0$ such that the mean-variance problem with p-norm transaction costs, Problem* (1.13)*, can be equivalently rewritten as the following constrained regularized linear regression problem*

$$\begin{aligned} \min_{\mathbf{w}} \quad & \|\mathbf{1}_T - R\mathbf{w}\|_2^2 + \kappa'\|\Lambda(\mathbf{w} - \mathbf{w_0})\|_p^p, \\ s.t \quad & \mathbf{w}^T\mu = \mu_0, \\ & \mathbf{w}^T\mathbf{1}_N = 1. \end{aligned} \tag{6.7}$$

Proof. The constrained OLS problem (6.7) can be rewritten as

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{1}_T^T\mathbf{1}_T + \mathbf{w}^T R^T R\mathbf{w} - 2\cdot \mathbf{1}_T^T R\mathbf{w} + \kappa'\|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p^p, \\ \text{s.t.} \quad & \mathbf{w}^T\mu = \mu_0, \\ & \mathbf{w}^T\mathbf{1}_N = 1. \end{aligned}$$

Moreover, because $\Sigma = R^T R - \mu\mu^T$ and $\mathbf{1}_T^T R = T\cdot\mu$, we have that this problem is equivalent to

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{1}_T^T\mathbf{1}_T + \mathbf{w}^T\Sigma\mathbf{w} + \mathbf{w}^T\mu\mu^T\mathbf{w} - 2\cdot T\cdot \mu^T\mathbf{w} + \kappa'\|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p^p, \\ \text{s.t.} \quad & \mathbf{w}^T\mu = \mu_0, \\ & \mathbf{w}^T\mathbf{1}_N = 1. \end{aligned}$$

Furthermore, because $\mathbf{w}^T\mu$ is constant in the feasible region, this problem is equivalent to

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T\Sigma\mathbf{w} + \kappa'\|\Lambda(\mathbf{w}-\mathbf{w}_0)\|_p^p, \\ \text{s.t.} \quad & \mathbf{w}^T\mu = \mu_0, \\ & \mathbf{w}^T\mathbf{1}_N = 1. \end{aligned}$$

It is easy to show that for any $\gamma > 0$ and $\kappa \geq 0$, there exists a μ_0 and $\kappa' \geq 0$ such that this problem is equivalent to Problem (1.13). □

A couple of comments about regression problem (6.7). First, the term $\kappa'\|\mathbf{w}-\mathbf{w_0}\|_p^p$ can be interpreted as a regularization term that reduces the impact of estimation error. Second, the constraints guarantee that the portfolio mean return is μ_0 and the portfolio weights add up to one.

Finally, the following proposition shows that the mean-variance problem with p-norm transaction costs is equivalent to a robust linear regression problem where the historical returns can take any value in an uncertainty set defined by a certain matrix norm.

Proposition 5. *Let Λ be a positive-definite diagonal matrix. Then, for every risk-aversion parameter $\gamma > 0$ and transaction cost parameter $\kappa \geq 0$, there exist μ_0 and $\delta > 0$ such that the mean-variance problem with p-norm transaction costs, Problem* (1.13)*, is equivalent to the following robust regression problem:*

$$\begin{aligned} \min_{\mathbf{w}} \max_{\Delta R \in U(\delta)} \quad & \|\mathbf{1}_T - R\mathbf{w} - \Delta R(\mathbf{w}-\mathbf{w_0})\|_2, \\ s.t \quad & \mathbf{w}^T\mu = \mu_0, \\ & \mathbf{w}^T\mathbf{1}_N = 1, \end{aligned} \tag{6.8}$$

where

$$U(\delta) = \{\Delta R = [\xi_1|...|\xi_N] : \|\Lambda^{-1}\left(\|\xi_1\|_2, ..., \|\xi_N\|_2\right)\|_q \leq \delta\}.$$

Proof. Given Proposition 4, it suffices to show that for every $\kappa' \geq 0$, there exists a $\delta > 0$ such that the optimal objective value of the inner optimization problem in (6.8) is equivalent (in terms of minimizers with respect to $\mathbf{w}$) to the objective function of Problem (6.7).

We prove this result in three steps. First, we prove that the inner optimization problem in (6.8) is equivalent to:

$$\begin{aligned} \max_{\mathbf{c}} \quad & \left(\|\mathbf{1}_T - R\mathbf{w}\|_2 + \sum_{i=1}^{N} c_i |\mathbf{w}_i - \mathbf{w}_{0i}| \right), \\ \text{s.t} \quad & \|\mathbf{c}\|_{q,\Lambda^{-1}} \leq \delta, \\ & \mathbf{c} \geq 0. \end{aligned} \tag{6.9}$$

Second, we prove that the optimal objective value of Problem (6.9) is equal to $\|\mathbf{1}_T - R\mathbf{w}\|_2 + \delta \|\Lambda(\mathbf{w} - \mathbf{w_0})\|_p$.

Third, we show that for every $\kappa' > 0$, there exists $\delta > 0$ such that problem $\min_{\mathbf{w}} \|\mathbf{1}_T - R\mathbf{w}\|_2 + \delta \|\Lambda(\mathbf{w} - \mathbf{w_0})\|_p$ is equivalent to $\min_{\mathbf{w}} \|\mathbf{1}_T - R\mathbf{w}\|_2^2 + \kappa' \|\Lambda(\mathbf{w} - \mathbf{w_0})\|_p^p$.

Step 1: Inner optimization problem equivalent to Problem (6.9).

We follow a similar argument to that in the proof of Theorem 14.8 of Caramanis et al. (2012). Given a fixed portfolio $\mathbf{w}$, we first show that for every matrix $\Delta R = [\xi_1|...|\xi_N]$ such that $\Delta R \in U(\delta)$, there exists a rank-one matrix $\Delta R^* \in U(\delta)$ such that

$$\|\mathbf{1}_T - R\mathbf{w} - \Delta R(\mathbf{w} - \mathbf{w_0})\|_2 \leq \|\mathbf{1}_T - R\mathbf{w} - \Delta R^*(\mathbf{w} - \mathbf{w_0})\|_2. \tag{6.10}$$

Specifically, we show that this inequality holds with the rank-one matrix $\Delta R^* = [\xi_1^*|...|\xi_N^*]$, where

$$\boldsymbol{\xi_i}^* = \begin{cases} -c_i \text{sgn}(\mathbf{w}_i - \mathbf{w}_{0i})\mathbf{v} & \text{if } \mathbf{w}_i \neq \mathbf{w}_{0i}, \\ 0 & \text{otherwise,} \end{cases}$$

where $c_i = \|\xi_i\|_2$ and the vector $\mathbf{v}$ is

$$\mathbf{v} = \begin{cases} \frac{\mathbf{1}_T - R\mathbf{w}}{\|\mathbf{1}_T - R\mathbf{w}\|_2} & \text{if } R\mathbf{w} \neq \mathbf{1}_T, \\ \text{any vector with unit 2-norm} & \text{otherwise.} \end{cases}$$

To see that Inequality (6.10) holds, note that by the triangular inequality

$$\|\mathbf{1}_T - R\mathbf{w} - \Delta R(\mathbf{w} - \mathbf{w_0})\|_2 \leq \|\mathbf{1}_T - R\mathbf{w}\|_2 + \sum_{i=1}^{N} |w_i - w_{0i}| \cdot \|\xi_i\|_2. \tag{6.11}$$

Moreover, because $\boldsymbol{\xi_i}^*$ is proportional to $\mathbf{1}_T - R\mathbf{w}$, we have that the triangular inequality is binding for ΔR^* and thus

$$\|\mathbf{1}_T - R\mathbf{w} - \Delta R^*(\mathbf{w} - \mathbf{w_0})\|_2 = \|\mathbf{1}_T - R\mathbf{w}^*\|_2 + \sum_{i=1}^{N} c_i|\mathbf{w}_i^* - \mathbf{w}_{0i}| \tag{6.12}$$

$$= \|\mathbf{1}_T - R\mathbf{w}\|_2 + \sum_{i=1}^{N} |w_i - w_{0i}| \cdot \|\xi_i\|_2. \tag{6.13}$$

Inequality (6.10) shows that the maximizer to the inner optimization problem in (6.7) must be a rank-one matrix in $U(\delta)$ whose columns are proportional to $\mathbf{1}_T - R\mathbf{w}$. Because all columns of these rank-one matrices are proportional to $\mathbf{1}_T - R\mathbf{w}$, there is a one-to-one correspondence between these rank-one matrices and the vectors $\mathbf{c}$ such that $\|\mathbf{c}\|_{q,\Lambda^{-1}} \leq \delta$ and $\mathbf{c} \geq 0$. This correspondence is specifically given by $\mathbf{c}_i = \|\xi_i^*\|_2$. This together with Equations (6.12) and (6.13) imply that the inner optimization problem in (6.8) is equivalent to Problem (6.9).

Step 2: the optimal value of Problem (6.9) **is** $\|\mathbf{1}_T - R\mathbf{w}\|_2 + \delta\|\Lambda(\mathbf{w} - \mathbf{w_0})\|_p$.

Caramanis et al. (2012) state a similar result without proof in their Corollary 14.11. Note that $\|\mathbf{1}_T - R\mathbf{w}\|_2$ does not depend on $\mathbf{c}$ and thus to simplify the exposition we focus thereafter on problem:

$$\begin{aligned} \max_{\mathbf{c}} \quad & \sum_{i=1}^{N} c_i|\mathbf{w}_i - \mathbf{w}_{0i}|, \\ \text{s.t} \quad & \|\mathbf{c}\|_{q,\Lambda^{-1}} \leq \delta, \\ & \mathbf{c} \geq 0. \end{aligned} \tag{6.14}$$

We prove the result in two sub-steps. First, we show that the nonnegativity constraints in Problem (6.14) are redundant. Second, we show the result.

Step 2.1: the nonnegativity constraints in Problem (6.14) **are redundant**

Consider the problem without shortsale constraints:

$$\max_{\|\mathbf{c}\|_{q,\Lambda^{-1}} \leq \delta} \mathbf{c}|\mathbf{w} - \mathbf{w}_0|. \tag{6.15}$$

Let $\mathbf{c}^*$ be the maximizer of Problem (6.15), and assume for contradiction that there exists j such that $c_j^* < 0$. Then, there exists c^{**} such that $c^{**} \geq 0$ and $\sum_{i=1}^N c_i^{**}|\mathbf{w}_i - \mathbf{w}_{0i}| > \sum_{i=1}^N c_i^*|\mathbf{w}_i - \mathbf{w}_{0i}|$. Indeed, it suffices to choose $c_j^{**} = -c_j^*$. Because Λ is assumed to be a diagonal positive definite matrix, its diagonal elements $\lambda_1, ..., \lambda_N$ are strictly positive. Moreover, because the objective function is linear, the constraint in (6.15) is binding, thus

$$\sum_{i=1}^N \left(\frac{1}{\lambda_i}|c_i^*|\right)^q = \sum_{i=1}^N \left(\frac{1}{\lambda_i}|c_i^{**}|\right)^q = \delta^q,$$

and $\sum_{i=1}^N c_i^{**}|\mathbf{w}_i - \mathbf{w}_{0i}| > \sum_{i=1}^N c_i^*|\mathbf{w}_i - \mathbf{w}_{0i}|$.

Step 2.2: the optimal value of problem (6.15) is $\delta\|\Lambda(\mathbf{w} - \mathbf{w_0})\|_p$.

We use Lemma 4 and the definition of dual norm. Specifically, we can deduce that the solution of problem (6.15) is $\delta\||\mathbf{w} - \mathbf{w}_0|\|_{p,\Lambda}$, but as Λ is a diagonal positive definite matrix we have:

$$\delta\|\Lambda|\mathbf{w} - \mathbf{w}_0|\|_p = \left(\sum_{i=1}^N |\lambda_i|w_i - w_{0i}||^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^N |\lambda_i(w_i - w_{0i})|^p\right)^{\frac{1}{p}} = \delta\|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p.$$

Step 3: problem $\min_{\mathbf{w}} \|\mathbf{1}_T - R\mathbf{w}\|_2 + \delta\|\Lambda(\mathbf{w} - \mathbf{w_0})\|_p$ is equivalent to $\min_{\mathbf{w}} \|\mathbf{1}_T - R\mathbf{w}\|_2^2 + \kappa'\|\Lambda(\mathbf{w} - \mathbf{w_0})\|_p^p$

This step follows using a similar reasoning to that behind the last paragraph of the proof of Proposition 3 and taking into account that the minimizer of $\|\mathbf{1}_T - R\mathbf{w}\|_2$ is the same as the one of $\|\mathbf{1}_T - R\mathbf{w}\|_2^2$. □

A few comments are in order. First, Proposition 5 shows that the portfolio problem with p-norm transaction costs is equivalent to a robust regression problem where the true historical returns can take any value in the uncertainty set $U(\delta)$. To see this, note that the term $\Delta R(\mathbf{w} - \mathbf{w}_0)$ in the objective function (6.8) captures the regret (in terms of regression mean squared error), that the investor may experience by trading from the starting portfolio, given that there is uncertainty about the true historical returns.

Second, the uncertainty set $U(\delta)$ is defined in terms of the $((q, \Lambda^{-1}), 2)$-matrix-norm:

$$\|\mathbf{X}\|_{(q,\Lambda^{-1}),2} = \|\Lambda^{-1}\left(\|\mathbf{x}_1\|_2, ..., \|\mathbf{x}_N\|_2\right)\|_q.$$

Kowalski (2009) used a similar norm in the context of regularized regression. Finally, to simplify the exposition we assume in Proposition 5 that the transaction cost matrix Λ is diagonal, but it is possible to show that the result in Proposition 5 holds for the general case with symmetric positive definite Λ.

6.4 The Bayesian portfolio problem

The following proposition shows that the mean-variance problem with p-norm transaction costs is equivalent to a Bayesian portfolio problem where the investor believes a priori that the portfolio weights are jointly distributed as a Multivariate Exponential Power (MEP) distribution. The method of proof is similar to that used by DeMiguel et al. (2009a), who provide a Bayesian interpretation for the 1-norm, 2-norm and A-norm constrained portfolios. We extend their analysis to the general case with p-norm transaction cost, with $p \in (1, 2)$.

Proposition 6. *For every risk-aversion parameter $\gamma > 0$ and transaction cost parameter $\kappa \geq 0$, there exists $\alpha > 0$ such that the weights that solve the mean-variance portfolio optimization problem with p-norm transaction costs, Problem* (1.13)*, are the mode of the posterior distribution of the mean-variance portfolio weights when the investor believes a priori that the variance of the mean-variance portfolio return, denoted by σ^2, has an independent distribution $\pi(\sigma^2)$, that asset returns are normally distributed, and that the mean-variance portfolio weights are jointly distributed as an MEP distribution, whith probability density function:*

$$\pi(\mathbf{w}) = \frac{p^N|\Lambda|}{2^N\alpha^N\Gamma(1/p)^N} e^{-\frac{\|\Lambda(\mathbf{w}-\mathbf{w}_0)\|_p^p}{\alpha^p}}, \tag{6.16}$$

where α is the scale parameter and $\Gamma(\cdot)$ is the gamma function.

Proof. **Step 1: Equation (6.16) defines a distribution.**
To see this, note that if the portfolio weights were independently distributed as an Exponential Power distribution, the joint prior distribution for a portfolio $\mathbf{w}$ would be

$$\pi'(\mathbf{w}) = \prod_{i=1}^{N} \pi'_0(\mathbf{w}_i) = \prod_{i=1}^{N} \frac{p}{2\alpha\Gamma(1/p)} e^{-\frac{|\mathbf{w}_i - \mathbf{w}_{i0}|^p}{\alpha}} = \frac{p^N}{2^N\alpha^N\Gamma(1/p)^N} e^{-\frac{\|\mathbf{w}-\mathbf{w}_0\|_p^p}{\alpha^p}},$$

and thus we know that

$$\int_{\mathbb{R}^N} \frac{p^N}{2^N \alpha^N \Gamma(1/p)^N} e^{-\frac{\|\mathbf{w}-\mathbf{w}_0\|_p^p}{\alpha^p}} = 1.$$

The difference between the distribution $\pi'(\mathbf{w})$ and the distribution in Equation (6.16) is simply a linear transformation of the variables $h(\mathbf{w}-\mathbf{w}_0) = \Lambda(\mathbf{w}-\mathbf{w}_0)$. Then, by a well-known theorem of elementary analysis, see Schwartz (1954), we have that:

$$\int_{\mathbb{R}^N} \frac{p^N}{2^N \alpha^N \Gamma(1/p)^N} e^{-\frac{\|\mathbf{w}-\mathbf{w}_0\|_p^p}{\alpha^p}} = \int_{\mathbb{R}^N} \frac{p^N |\Lambda|}{2^N \alpha^N \Gamma(1/p)^N} e^{-\frac{\|\Lambda(\mathbf{w}-\mathbf{w}_0)\|_p^p}{\alpha^p}} = 1.$$

Step 2: mean-variance portfolio weights are mode of posterior distribution. We now show that the mean-variance portfolio weights with p-norm transaction cost are the mode of the posterior distribution. Bayes theorem guarantees that the posterior distribution of the portfolio weights and portfolio return variance conditional on the observed sample returns is

$$\pi(\mathbf{w}, \sigma^2 | \{r_t\}_{t=1}^T) = \frac{\pi(\mathbf{w})\pi(\sigma^2) L(\mathbf{w}, \sigma^2 | \{r_t\}_{t=1}^T)}{\int_{\mathbf{w},\sigma^2} \pi(\mathbf{w})\pi(\sigma^2) L(\mathbf{w}, \sigma^2 | \{r_t\}_{t=1}^T) d\mathbf{w} d\sigma^2},$$

where $L(\mathbf{w}, \sigma^2 | \{r_t\}_{t=1}^T)$ stands for the likelihood function of $\mathbf{w}$ and σ^2 given the observed sample returns $\{r_t\}_{t=1}^T$

$$L(\mathbf{w}, \sigma^2 | \{r_t\}_{t=1}^T) = \frac{1}{\sigma^N \sqrt{2^N \pi^N}} e^{\frac{-\sum_{t=1}^T (\mathbf{w}^T r_t - \mathbf{w}^T \hat{\mu})^2}{2\sigma^2}}.$$

Hence

$$\pi(\mathbf{w}, \sigma^2 | \{r_t\}_{t=1}^T) \propto \pi(\mathbf{w})\pi(\sigma^2) L(\mathbf{w}, \sigma^2 | \{r_t\}_{t=1}^T);$$

that is,

$$\pi(\mathbf{w}, \sigma^2 | \{r_t\}_{t=1}^T) \propto \frac{\pi(\sigma^2)}{\sigma^N} e^{\frac{-\sum_{t=1}^T (\mathbf{w}^T r_t - \mathbf{w}^T \hat{\mu})^2}{2\sigma^2} - \frac{\|\Lambda(\mathbf{w}-\mathbf{w}_0)\|_p^p}{\alpha^p}}. \tag{6.17}$$

From (6.17), it is straightforward to show that for a given portfolio return variance σ^2, there exist μ_0 such that the portfolio that maximizes the posterior distribution of the mean-variance portfolio weights subject to the condition that the portfolio weights add up to one, is the solution to the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T \Sigma \mathbf{w} + \kappa \|\Lambda(\mathbf{w} - \mathbf{w}_0)\|_p^p, \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{1}_N = 1, \\ & \mu^T \mathbf{w} \geq \mu_0, \end{aligned}$$

where

$$\kappa = \frac{2\sigma^2}{\alpha^p (T-1)}.$$

□

The proof of Proposition 6 establishes that for a given portfolio return variance σ^2, the following relation between the transaction cost parameter and the investor prior beliefs holds:

$$\kappa = \frac{2\sigma^2}{\alpha^p (T-1)}, \tag{6.18}$$

where T is the estimation window length. We make three observations from Equation (6.18). First, the higher the portfolio return variance σ^2, the higher the equivalent transaction cost parameter. The intuition behind this result is that the less certain the investor is about the portfolio return, the more reticent she will be to move from her prior belief portfolio weights. This reticence is equivalent to facing large transaction costs. Second, the longer the estimation window, the smaller the equivalent transaction cost parameter. The intuition behind this result is that the longer the estimation window, the more accurate the estimators obtained from the historical data, and thus the more willing the investor is to move away from her prior portfolio weights. Third, the larger the scale parameter of the prior distribution of portfolio weights, the smaller the equivalent transaction cost parameter. Intuitively, the less confident the investor is in her prior beliefs, the more willing she will be to move away from the prior belief weights.

Note that the MEP prior distribution includes as a particular case the Multivariate Normal prior distribution for the case when $p = 2$ and $\Lambda = \Sigma^{1/2}$. Proposition 6 implies that assuming this Multivariate Normal prior is equivalent to capturing a quadratic transaction cost with $\Lambda = \Sigma^{1/2}$. Gârleanu and Pedersen (2013) point out that this type of quadratic transaction cost is reasonable for the case with large trades. Also, the MEP prior includes as a particular case (when $p = 1$ and $\Lambda = I$) a prior belief that the portfolio weights are independently and identically distributed as a Laplace distribution. Proposition 6 implies that assuming this Laplace prior is equivalent to capturing a proportional transaction cost. For values of $p \in (1, 2)$, the MEP prior generalizes the Normal and Laplacian priors, and provides a Bayesian interpretation for the p-norm transaction costs beyond the cases with $p = 1$ and $p = 2$.

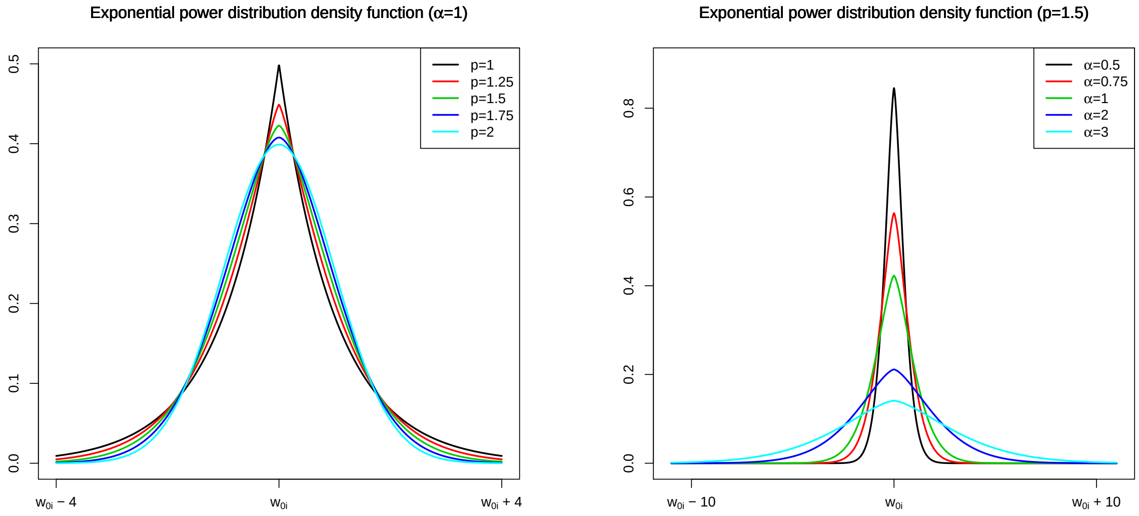


Figure 6.2: Power exponential density functions for different values of p and α.

This figure depicts power exponential density functions for different values of p and α. The left graph depicts the power exponential density functions for different values of p and $\alpha = 1$, and the right graph for different values of α and fixed $p = 1.5$.

Figure 6.2 depicts the Exponential Power distribution for the case with a single asset. The left graph depicts the Exponential Power distribution for different values of $p \in [1, 2]$ with $\alpha = 1$. The case with $p = 1$ corresponds to the Laplace distribution, and the case with $p = 2$ to the Normal distribution. The graph illustrates that for values of $p < 2$, the Exponential Power distribution has sharper peak and fatter tails than the Normal distribution. The right graph in Figure 6.2 depicts the Exponential Power distribution for different values of the scale parameter α with $p = 1.5$. The graph illustrates that the scale parameter is a proxy for the dispersion of the distribution; that is, the higher the α, the higher the dispersion of the distribution.

Table 6.1 summarizes our main theoretical results for the general case with p-norm transaction costs as well as for the particular cases with quadratic and proportional transaction costs. The first column indicates the type of transaction cost, the second column gives the mathematical formulation of the transaction cost term, the third, fourth, and fifth columns give the corresponding Bayesian prior, robust optimization uncertainty set, and robust regression uncertainty set, respectively.

Transaction	Transaction	**Bayesian**	**Robust optim.**	**Robust regression**
cost	cost	prior dist.	Uncertainty set	Uncertainty set
type	formulation	for weights	for mean	for historical ret.
p-norm	$\kappa\lVert\Lambda\Delta\mathbf{w}\rVert_p^p$	MEP	$\lVert\Lambda^{-1}\Delta\mu\rVert_q \leq \delta$	$\lVert\Delta R\rVert_{(q,\Lambda^{-1}),2} \leq \delta$
quadratic	$\kappa\lVert\Lambda\Delta\mathbf{w}\rVert_2^2$	Normal	$\lVert\Lambda^{-1}\Delta\mu\rVert_2 \leq \delta$	$\lVert\Delta R\rVert_{(2,\Lambda^{-1}),2} \leq \delta$
proportional	$\kappa\lVert\Delta\mathbf{w}\rVert_1$	Laplace	$\lVert\Delta\mu\rVert_\infty \leq \delta$	$\lVert\Delta R\rVert_{\infty,2} \leq \delta$

Table 6.1: Summary of main theoretical results.

6.5 Data-driven portfolios

The theoretical results in the previous sections demonstrate that incorporating a p-norm transaction cost term in the mean-variance portfolio problem may help to reduce the impact of estimation error. Motivated by these findings, we now propose a data-driven approach to portfolio selection with estimation error and transaction costs. Concretely, we propose using a nonparametric approach to calibrate the transaction cost parameter to obtain portfolios that are not only efficient from a transaction cost perspective, but also robust to estimation error.

We evaluate the out-of-sample performance of the proposed data-driven portfolios on the five empirical datasets with US stock monthly return data listed in Table 6.2, which are similar those used in the literature; see DeMiguel et al. (2009a) and the references therein. Specifically, we consider two datasets with returns on industry portfolios downloaded from Ken French's website, two datasets with returns on portfolios of stocks sorted by size and book-to-market from Ken French's website, and one dataset with returns on individual stocks downloaded from the CRSP database. We assume proportional transaction costs of 50 basis points, which are similar to those considered in the literature; see, for instance, Balduzzi and Lynch (1999); DeMiguel et al. (2009b), and the references therein.

The remainder of this section is organized as follows. Section 6.5.1 describes the benchmark and data-driven portfolios that we compare. Section 6.5.2 describes the rolling-horizon methodology used to compare the different portfolio policies. Finally, Section 6.5.3 discusses the performance of the different portfolio policies.

Abbrv.	Dataset	N	Time period	Source
10Ind	Ten industry portfolios	10	07/1963-12/2013	K.French
48Ind	Forty eight industry portfolios	48	07/1963-12/2013	K.French
6FF	6 portfolios sorted by size and book-to-market	6	07/1963-12/2013	K.French
25FF	25 portfolios of firms sorted by size and b-t-m	25	07/1963-12/2013	K.French
CRSP	500 randomized stocks from CRSP	500	04/1968-04/2005	CRSP

Table 6.2: Datasets used for comparison of portfolios.

6.5.1 Description of the portfolios

Table 6.3 contains four panels corresponding to the four different types of portfolios we compare. Panel A lists benchmark portfolios that ignore transaction costs and estimation error. These are computed by solving Problem (1.13) with transaction cost parameter $\kappa = 0$. Panel B lists benchmark portfolios that capture the nominal proportional transaction costs, but ignore estimation error. These are computed by solving Problem (1.13) with $p = 1$, $\Lambda = I$, where I is the identity matrix, and $\kappa = 50$ basis points. Panel C lists the data-driven portfolios with calibrated proportional transaction costs. These are computed by solving Problem (1.13) with $p = 1$, $\Lambda = I$, and $\kappa = \kappa_{cv}$ calibrated via cross-validation. Finally, Panel D lists the data-driven portfolios with calibrated quadratic transaction costs. These are computed by solving Problem (1.13) with $p = 2$, $\Lambda = \Sigma^{1/2}$, where Σ is the estimated covariance matrix, and $\kappa = \kappa_{cv}$ calibrated via cross-validation.

Each panel lists four different portfolios depending on the variant of mean-variance portfolio considered. The mean-variance portfolio (MEAN) is computed by solving Problem (1.13) with risk-aversion parameter $\gamma = 5$.[3] The minimum-variance portfolio (MIN) is computed by solving Problem (1.13) after dropping the term $\mu^T \mathbf{w}$. We consider the minimum-variance portfolio because it is well known that it often outperforms the mean-variance portfolio out of sample due to the difficulty estimating mean returns.[4] Finally, we consider shortsale-constrained versions of the mean variance (MEANC) and minimum-variance (MINC) portfolios computed by imposing the constraint $\mathbf{w} \geq 0$. We consider shortsale constrained portfolios because as documented by Jagannathan and Ma (2003) these constraints often help to improve out-of-sample performance.

A few comments are in order. First, why consider data-driven portfolios with cali-

[3] We have considered other values of the risk aversion parameter ($\gamma = 2, 10$), but the results are qualitatively similar and thus we do not report them to conserve space.

[4] (Jagannathan and Ma, 2003, pp. 1652Ű-1653), for instance, explain that: “The estimation error in the sample mean is so large nothing much is lost in ignoring the mean altogether when no further information about the population mean is available”.

Portfolio	Abbreviation
Panel A. Benchmark portfolios that ignore transaction costs	
Minimum-variance portfolio, shortsale unconstrained, $\kappa = 0$	MIN
Minimum-variance portfolio, shortsale constrained, $\kappa = 0$	MINC
Mean-variance portfolio, shortsale unconstrained, $\kappa = 0$	MEAN
Mean-variance portfolio, shortsale constrained, $\kappa = 0$	MEANC
Panel B. Benchmark portfolios with nominal transaction costs	
Minimum-variance portfolio, shortsale unconstrained, $0.005\|\Delta\mathbf{w}\|$	
Minimum-variance portfolio, shortsale constrained, $0.005\|\Delta\mathbf{w}\|$	
Mean-variance portfolio, shortsale unconstrained, $0.005\|\Delta\mathbf{w}\|$	
Mean-variance portfolio, shortsale constrained, $0.005\|\Delta\mathbf{w}\|$	
Panel C. Data-driven portfolios with calibrated proportional transaction costs	
Minimum-variance portfolio, shortsale unconstrained, $\kappa_{cv}\|\Delta\mathbf{w}\|$	
Minimum-variance portfolio, shortsale constrained, $\kappa_{cv}\|\Delta\mathbf{w}\|$	
Mean-variance portfolio, shortsale unconstrained, $\kappa_{cv}\|\Delta\mathbf{w}\|$	
Mean-variance portfolio, shortsale constrained, $\kappa_{cv}\|\Delta\mathbf{w}\|$	
Panel D. Data-driven portfolios with calibrated quadratic transaction costs	
Minimum-variance portfolio, shortsale unconstrained, $\kappa_{cv}\|\Sigma^{1/2}\Delta\mathbf{w}\|_2^2$	
Minimum-variance portfolio, shortsale constrained, $\kappa_{cv}\|\Sigma^{1/2}\Delta\mathbf{w}\|_2^2$	
Mean-variance portfolio, shortsale unconstrained, $\kappa_{cv}\|\Sigma^{1/2}\Delta\mathbf{w}\|_2^2$	
Mean-variance portfolio, shortsale constrained, $\kappa_{cv}\|\Sigma^{1/2}\Delta\mathbf{w}\|_2^2$	

Table 6.3: Portfolios evaluated, where κ_{cv} is calibrated with cross-validation.

brated *quadratic* transaction costs when the nominal transaction costs are *proportional*? The answer is that the data-driven portfolios are designed to address not only transaction costs, but also estimation error, and it is therefore possible that the data-driven portfolios with calibrated *quadratic* costs offer a good trade-off between these two objectives. Second, note that the transaction cost parameters corresponding to the data-driven portfolios with proportional versus quadratic costs are not easy to compare. The reason for this is that the parameters for the two different types of data-driven portfolios have different units: one multiplies a 1-norm term and the other a squared 2-norm term. Fortunately, the data-driven portfolios can be equivalently calibrated in terms of trading volume or turnover, which can be directly compared for the data-driven portfolios with *proportional* and *quadratic* transaction costs. To see this, note that for the portfolios with proportional costs, it is easy to show that there is a one-to-one correspondence between the minimizers to the problem with proportional transaction costs for different values of the transaction cost parameter κ, and the problem with turnover constraint $\|\mathbf{w} - \mathbf{w}_0\|_1 \leq \tau$ for different values of the trading threshold τ. Likewise, for the data-driven portfolios with quadratic transaction costs and $\Lambda = \Sigma^{1/2}$, Kourtis (2015) shows that the optimal portfolio is a convex combination of the starting portfolio and the mean-variance portfolio in the absence of transaction costs. Consequently, for these portfolios one can also calibrate the trading volume τ instead of the transaction cost parameter κ. Summarizing, to facilitate the comparison between the two types of data-driven portfolios (with proportional and quadratic costs), we calibrate these portfolios by selecting their trading volume τ from 0%, 0.5%, 1%, 2.5%, 5% and 10%.[5]

We calibrate the data-driven portfolios using the bootstrap methodology of 10-fold cross-validation; Efron and Gong (1983). Specifically, we divide the estimation window of M returns into ten intervals of $M/10$ returns each. For j from 1 to 10, we remove the j^{th}-interval from the estimation window, and use the remaining sample returns to compute the data-driven portfolio for each value of the trading volume τ from 0%, 1%, 2.5%, 5% and 10%. We then evaluate the return of the resulting portfolios (net of transaction costs of 50 basis points) on the j^{th}-interval. After completing this process for each of the 10 intervals, we have the M "out-of-sample" portfolio returns for each value of τ. Finally, we compute the variance of these out-of-sample returns and select the value of τ that corresponds to the portfolio with smallest variance.[6]

Finally, we have also used 10-fold cross-validation as well as generalized cross-validation as defined by (Fu, 1998, Section 5) to calibrate simultaneously the transaction

[5] For computational convenience, we approximate the shortsale-constrained data-driven portfolios with quadratic transaction costs by taking a convex combination of the starting portfolio, which is a long-only portfolio, and the *shortsale-constrained* mean-variance portfolio in the absence of transaction costs.

[6] We have also tried using the Sharpe ratio of returns net of transaction costs as the calibration criterion, but the results are qualitatively similar to those obtained from the variance criterion, and thus we do not report them to conserve space.

cost parameter (value of κ) and the type of transaction cost (proportional or quadratic). We find that such calibrations do not result in a substantial improvement in terms of out-of-sample performance (compared to the case where we only calibrate the transaction cost parameter κ with 10-fold cross-validation) and thus we do not report the results to conserve space.

6.5.2 Comparison methodology

We use a rolling-horizon methodology similar to that used in DeMiguel et al. (2009b) and DeMiguel et al. (2009a) to compare the performance of the different portfolios. We use an estimation window of $M = 120$ monthly returns. We start by using the first M returns in the historical sample to compute the different portfolios and evaluate their out-of-sample return on period $M + 1$. We then "roll" the estimation window one month forward by dropping the earliest return and adding a new return, and repeat the process. At the end of this iterative process we obtain $T - M$ out-of-sample returns for each portfolio that allow us to compare performance. Table 6.4 outlines the proposed methodology.

To test the statistical significance of the differences between the out-of-sample variance and Sharpe ratio of the different portfolios with those of the benchmark minimum-variance portfolio, we use the bootstrap methodology employed by DeMiguel et al. (2009a), which is based on the work by Ledoit and Wolf (2008).

6.5.3 Discussion

Tables 6.5, 6.6, 6.7, and 6.8 report the out-of-sample Sharpe ratio, turnover, variance, and mean, respectively, for the different portfolios. The quantity corresponding to the best performing portfolio is highlighted in bold for each dataset.

Sharpe ratio.

Table 6.5 reports the out-of-sample Sharpe ratio for each of the 16 portfolio policies considered, together with the p-value indicating whether the differences with the benchmark minimum-variance portfolio are significant. Panel A reports the Sharpe ratios for the benchmark portfolios that ignore transaction costs. This panel shows that the minimum-variance portfolios generally outperform mean-variance portfolios. This is explained by the well-known difficulties associated with estimating mean returns from historical data. Imposing shortsale constraints on the minimum-variance portfolio helps only for the two datasets with largest number of assets (48Ind and CRSP). This makes sense as estimating the covariance matrix of asset returns is harder for datasets with many assets, and under these circumstances the shortsale constraints will help to alleviate the impact of estimation error. Imposing shortsale constraints on the mean-variance

1. Initialization: set the first estimation window as $[\mathbf{r}_1, \mathbf{r}_M]$ and calculate the starting portfolio weights for each strategy η, $\mathbf{w}_0^\eta$
2. For the current estimation window, $[\mathbf{r}_{t-M+1}, \mathbf{r}_t]$:
   - (i) For every data-driven portfolio η perform 10-fold cross-validation to select $\tau_t^\eta \in \{0.00, 0.01, 0.025, 0.05, 0.1\}$ to minimize portfolio variance
   - (ii) Compute optimal portfolio weights for every strategy η, $\mathbf{w}_t^\eta$
   - (iii) Use the next return $\mathbf{r}_{t+1}$ to compute the out-of-sample return net of transaction costs at time $t+1$ for every strategy η:

     $$R_{t+1}^\eta = \mathbf{r}_{t+1}^T \mathbf{w}_t^\eta - (1 + \mathbf{r}_{t+1}^T \mathbf{w}_t^\eta) \times 0.005 \|\mathbf{w}_t^\eta - \mathbf{w}_0^\eta\|_1$$

   - (iv) Use the next return $\mathbf{r}_{t+1}$ to compute the starting portfolio for the next period $\mathbf{w}_0^\eta$
3. Roll the estimation window by dropping the first return ($\mathbf{r}_{t-M+1}$) and adding the next return ($\mathbf{r}_{t+1}$) to obtain the next estimation window $[\mathbf{r}_{t-M+2}, \mathbf{r}_{t+1}]$
4. Repeat steps 2-3 until $t = T - 1$
5. We have a sample of $T - M$ portfolio returns for each strategy η. Compute:
   - Mean
   - Variance
   - Sharpe ratio
   - Turnover

Table 6.4: Methodology used to compare portfolios

Strategy	10Ind	48Ind	6FF	25FF	CRSP
Panel A. Benchmark portfolios that ignore transaction costs					
MIN	0.3027	0.1220	0.3525	0.3195	0.3243
	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
MINC	0.2954	0.2642	0.2490	0.2461	**0.3741**
	(0.34)	(0.00)	(0.00)	(0.02)	(0.07)
MEAN	0.0756	-0.1015	0.2407	0.0941	-0.1124
	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
MEANC	0.2120	0.1754	0.2492	0.2370	0.1272
	(0.01)	(0.15)	(0.00)	(0.04)	(0.00)
Panel B. Benchmark portfolios with nominal transaction costs					
MIN	0.2899	0.0972	0.3319	0.3274	0.3614
	(0.34)	(0.36)	(0.17)	(0.39)	(0.02)
MINC	0.2442	0.2652	0.2370	0.2317	0.3684
	(0.06)	(0.00)	(0.00)	(0.02)	(0.13)
MEAN	0.2124	-0.0893	0.3129	0.1455	-0.0506
	(0.02)	(0.00)	(0.20)	(0.01)	(0.00)
MEANC	0.2211	0.1989	0.2478	0.2586	0.1806
	(0.02)	(0.06)	(0.00)	(0.08)	(0.03)
Panel C. Data-driven portfolios with calibrated proportional transaction cost					
MIN	**0.3301**	0.1282	0.3438	0.3921	0.3595
	(0.06)	(0.44)	(0.31)	(0.00)	(0.03)
MINC	0.2992	**0.3000**	0.2484	0.2567	0.3696
	(0.46)	(0.00)	(0.00)	(0.06)	(0.12)
MEAN	0.2487	-0.0668	0.3112	0.1586	-0.0629
	(0.09)	(0.01)	(0.15)	(0.02)	(0.00)
MEANC	0.2707	0.2462	0.2498	0.2460	0.1754
	(0.17)	(0.00)	(0.00)	(0.06)	(0.02)
Panel D. Data-driven portfolios with calibrated quadratic transaction cost					
MIN	0.3278	0.2486	**0.3554**	**0.3998**	0.3607
	(0.02)	(0.00)	(0.42)	(0.00)	(0.02)
MINC	0.3013	0.2812	0.2432	0.2522	0.3700
	(0.49)	(0.00)	(0.00)	(0.04)	(0.11)
MEAN	0.2466	0.0627	0.3442	0.1304	-0.0710
	(0.07)	(0.14)	(0.43)	(0.04)	(0.00)
MEANC	0.2767	0.2633	0.2502	0.2523	0.1731
	(0.19)	(0.00)	(0.00)	(0.06)	(0.02)

Table 6.5: Sharpe ratios

This table reports the monthly out-of-sample Sharpe ratio and the corresponding p-value that the Sharpe ratio for each of the portfolios is different from that for the minimum-variance portfolio

portfolio helps for every dataset because the unconstrained mean-variance portfolio is very sensitive to estimation error.

Panel B shows that capturing nominal proportional transaction costs generally helps to improve the performance of the mean-variance portfolios, but it only helps to improve the minimum-variance portfolio for two of the five datasets (25FF and CRSP), and the difference is statistically significant only for the CRSP dataset. The reason for this is that the nominal transaction cost term helps to combat estimation error to a certain extend, and this is helpful for mean-variance portfolios, which are very sensitive to estimation error. Minimum-variance portfolios, on the other hand, are more resilient to estimation error, and a nominal transaction cost parameter is not sufficient to strike the right balance between estimation error and transaction costs. This seems to indicate that using a data-driven approach to calibrate the transaction cost parameter may help to improve the performance.

From Panels A and B we conclude that the shortsale-unconstrained minimum-variance portfolio that ignores transaction costs is the best of our benchmark portfolios, and we now compare its performance to that of the data-driven portfolios in Panels C and D.

Panel C shows that the data-driven approach based on *proportional* transaction costs generally helps to improve the performance of the traditional portfolios. Specifically, the data-driven shortsale-unconstrained minimum-variance portfolio outperforms its benchmark counterpart for every dataset except 6FF, with an improvement in Sharpe ratio that ranges from 9% to 23% for the different datasets, and is statistically significant for two of the five datasets.

Finally, Panel D shows that the data-driven approach based on quadratic transaction costs also helps to improve the performance of the traditional portfolio. Specifically, the data-driven shortsale-unconstrained minimum-variance portfolio with calibrated quadratic transaction costs outperforms its benchmark counterpart for every dataset, with an improvement in Sharpe ratio that ranges from 1% to 104%, and is statistically significant for four of the five datasets.

Turnover.

Table 6.6 reports the turnovers for the different portfolio policies. The table shows that taking transaction costs into account helps to reduce turnover. For instance, for the 25FF dataset, the monthly turnover of the shortsale-unconstrained minimum-variance portfolio is 75.32%. Including a calibrated quadratic transaction cost term reduces this turnover to 5.14%, including a calibrated proportional transaction cost term to 0.89%, and including a nominal proportional transaction cost term to 0.02%.

Comparing the turnover of the portfolios with nominal proportional transaction costs (Panel B) with that of the data-driven portfolios with calibrated proportional

Strategy	10Ind	48Ind	6FF	25FF	CRSP
Panel A. Benchmark portfolios that ignore transaction costs					
MIN	0.1422	0.7450	0.1945	0.7532	0.1905
MINC	0.0494	0.0711	0.0474	0.0765	0.0714
MEAN	0.9651	11.0694	1.4017	8.3947	1.3716
MEANC	0.1523	0.2138	0.1187	0.2388	0.1516
Panel B. Benchmark portfolios with nominal transaction costs					
MIN	**0.0000**	0.0103	0.0001	0.0002	**0.0000**
MINC	0.0003	**0.0010**	**0.0000**	**0.0001**	0.0001
MEAN	0.0106	55.5685	0.0069	0.1989	0.1412
MEANC	0.0054	0.0114	0.0003	0.0002	0.0009
Panel C. Data-driven portfolios with calibrated proportional transaction cost					
MIN	0.0098	0.0135	0.0107	0.0089	0.0014
MINC	0.0087	0.0079	0.0118	0.0092	0.0004
MEAN	0.0134	0.0865	0.0170	0.0190	0.0306
MEANC	0.0056	0.0083	0.0087	0.0101	0.0175
Panel D. Data-driven portfolios with calibrated quadratic transaction cost					
MIN	0.0181	0.0490	0.0219	0.0514	0.0019
MINC	0.0113	0.0141	0.0136	0.0129	0.0016
MEAN	0.0215	0.0966	0.0301	0.0594	0.0301
MEANC	0.0093	0.0079	0.0114	0.0145	0.0276

Table 6.6: Turnover

This table reports the monthly turnover of the different portfolios. Turnover is the average percentage of wealth traded in each period and is equal to the sum of the absolute value of the rebalancing trades across the N available assets and over the $T - M - 1$ trading dates, normalized by the total number of trading dates.

transaction costs (Panel C), we observe that the nominal transaction cost term induces an *all-or-nothing* trading pattern, whereas the data-driven portfolios are associated with intermediate levels of turnovers. For instance, the unconstrained minimum-variance portfolio with nominal costs is effectively a *buy-and-hold* portfolio (with almost zero turnover) for every dataset except 48Ind, whereas the counterpart data-driven portfolios with calibrated proportional costs have reasonable monthly turnovers ranging between 0.14% and 1.35% for the different datasets. On the other hand, the unconstrained mean-variance portfolio with nominal costs has very large monthly turnover for 25FF and CRSP of around 15–20%, and huge of 5,556% for 48Ind, whereas the counterpart data-driven portfolios with calibrated proportional costs have reasonable turnovers ranging between 1.34% and 8.65%.

The mathematical intuition behind why the nominal transaction cost term induces an all-or-nothing trading pattern is that the proportional transaction cost term is a piecewise linear term, which when combined in the objective function with the linear-quadratic mean-variance objective, results in policies that advice either large trading or no trading. This all-or-nothing trading pattern is indeed optimal in the absence of estimation error. Constantinides (1979, 1986); Davis and Norman (1990) and Muthuraman and Kumar (2006), amongst others, show that the optimal portfolio policy in the presence of proportional transaction costs is characterized by a no-trade region: if the portfolio is inside this region, then it is optimal not to trade, and if it is outside, then it is optimal to trade to the boundary of this region.

The Sharpe ratio results in Table 6.5, however, show that this *all-or-nothing* trading pattern leads to poor performance when in addition to transaction costs the investor is also facing estimation error. *No trading* results in poor performance because buy-and-hold policies essentially ignore the information available in recent historical data—they stick to the portfolio weights obtained from the earliest estimation window. *All trading* leads to poor performance because the resulting portfolio policies are too sensitive to recent historical data, which leads to large transaction costs and sensitivity to estimation error. The data-driven portfolios, on the other hand, allow reasonable amounts of turnover that strike an optimal trade-off between incorporating the information in recent historical return data, and avoiding the large transaction costs and impact of estimation error associated with large turnovers.[7]

[7] Note that this is a counterintuitive result as one would expect that the data-driven portfolios would always result in smaller turnover compared to the portfolios that capture nominal transaction costs. Concretely, one would expect that the data-driven portfolios would result in a larger calibrated transaction cost parameter in order to robustify the portfolios and reduce the impact of estimation error. Our results, however, show that from a data-driven perspective, it is optimal to calibrate the transaction cost parameter to achieve intermediate levels of turnover.

Strategy	10Ind	48Ind	6FF	25FF	CRSP
Panel A. Benchmark portfolios that ignore transaction costs					
MIN	**0.1242**	3.8689	**0.0553**	0.9104	0.9630
	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
MINC	0.1259	**2.9976**	0.0686	1.1468	**0.8717**
	(0.35)	(0.00)	(0.00)	(0.00)	(0.08)
MEAN	0.5280	218.7537	0.2515	20.7243	9.5164
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
MEANC	0.2275	7.4006	0.0876	1.4930	1.6589
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Panel B. Benchmark portfolios with nominal transaction costs					
MIN	0.1684	9.1902	0.0616	1.0374	0.9664
	(0.00)	(0.00)	(0.00)	(0.03)	(0.48)
MINC	0.1870	4.3855	0.0713	1.2203	0.8804
	(0.00)	(0.08)	(0.00)	(0.00)	(0.15)
MEAN	0.2775	11100.0558	0.1115	10.6297	10.3566
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
MEANC	0.1955	6.2474	0.0780	1.3973	1.7651
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Panel C. Data-driven portfolios with calibrated proportional transaction cost					
MIN	0.1354	7.2630	0.0619	**0.8949**	0.9643
	(0.01)	(0.00)	(0.00)	(0.38)	(0.50)
MINC	0.1342	3.4277	0.0722	1.1827	0.8796
	(0.04)	(0.06)	(0.00)	(0.00)	(0.12)
MEAN	0.2182	337842.7232	0.1086	14.9258	11.1818
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
MEANC	0.1572	4.5914	0.0759	1.3007	1.7030
	(0.00)	(0.02)	(0.00)	(0.00)	(0.00)
Panel D. Data-driven portfolios with calibrated quadratic transaction cost					
MIN	0.1315	3.6690	0.0615	0.9378	0.9644
	(0.02)	(0.18)	(0.00)	(0.27)	(0.47)
MINC	0.1317	3.3144	0.0728	1.1973	0.8795
	(0.08)	(0.02)	(0.00)	(0.00)	(0.13)
MEAN	0.2026	1088237.2067	0.1014	39.6984	11.6139
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
MEANC	0.1531	4.1090	0.0757	1.3166	1.6819
	(0.00)	(0.27)	(0.00)	(0.00)	(0.00)

Table 6.7: Variance

This table reports the monthly out-of-sample variances and the corresponding p-value that the variance for each of these strategies is different from that for the minimum-variance portfolio

Variance and mean.

Finally, Tables 6.7 and 6.8 show the out-of-sample variance and mean for the different portfolios. We observe that considering transaction costs generally does not help to reduce the variance of portfolio returns, but it helps to increase the mean. Thus the gains from using the proposed data-driven approaches are obtained from improvements in out-of-sample means, rather than variances. Intuitively, transaction costs are likely to impact mean returns more than return variance. To see this, assume the trading volume of a portfolio policy is relatively constant over time, then transaction costs will shift every portfolio return down by an almost constant amount. This "shift" will result in a reduction in mean returns, but may not have a large impact on variance. Finally, comparing the variances of the data-driven portfolios to those of the portfolios capturing nominal transaction costs, we conclude that the data-driven portfolios usually attain a lower out-of-sample variance.

Summary.

The proposed data-driven portfolios outperform the traditional portfolios in terms of Sharpe ratio net of transaction costs. Capturing a nominal proportional transaction cost term helps to improve the performance of the mean-variance portfolios, but not that of the minimum-variance portfolios. Comparing the data-driven and nominal portfolios, we conclude that calibrating the transaction cost term with cross-validation results in better out-of-sample performance because it leads to optimal data-driven levels of portfolio turnover. Finally, the gains from capturing transaction costs come in the form of higher mean returns, rather than lower variance of returns.

6.6 Concluding remarks

We show that there is an intimate relation between transaction costs and estimation error in portfolio selection. In particular, we show that a mean-variance problem with p-norm transaction costs is equivalent to three different robust formulations of the portfolio problem: (i) a robust optimization problem where mean returns can take any value in an uncertainty set, (ii) a robust regression problem where historical returns can take any value in an uncertainty set, and (iii) a Bayesian portfolio problem where the investor holds a certain prior on the portfolio weights. Table 6.1 summarizes the relation between these four formulations.

Motivated by our theoretical findings, we propose a data-driven approach to portfolio selection with transaction costs and estimation error. The approach consists of using the nonparametric technique of cross-validation to calibrate the transaction cost parameter in a mean-variance problem in order to improve the out-of-sample performance in the

Strategy	10Ind	48Ind	6FF	25FF	CRSP
Panel A. Benchmark portfolios that ignore transaction costs					
MIN	0.1067	0.2399	0.0829	0.3049	0.3182
MINC	0.1048	0.4575	0.0652	0.2635	0.3492
MEAN	0.0549	-1.5015	**0.1207**	0.4286	-0.3467
MEANC	0.1011	0.4772	0.0738	0.2896	0.1638
Panel B. Benchmark portfolios with nominal transaction costs					
MIN	0.1190	0.2948	0.0824	0.3335	**0.3553**
MINC	0.1056	0.5554	0.0633	0.2560	0.3457
MEAN	0.1119	-9.4064	0.1045	0.4743	-0.1627
MEANC	0.0978	0.4971	0.0692	0.3057	0.2400
Panel C. Data-driven portfolios with calibrated proportional transaction cost					
MIN	**0.1215**	0.3456	0.0855	0.3710	0.3530
MINC	0.1096	0.5555	0.0668	0.2791	0.3467
MEAN	0.1162	-38.8372	0.1025	0.6127	-0.2102
MEANC	0.1073	0.5275	0.0688	0.2806	0.2290
Panel D. Data-driven portfolios with calibrated quadratic transaction cost					
MIN	0.1189	0.4762	0.0882	0.3872	0.3542
MINC	0.1093	0.5119	0.0656	0.2759	0.3470
MEAN	0.1110	**65.3606**	0.1096	**0.8215**	-0.2420
MEANC	0.1083	0.5337	0.0688	0.2895	0.2244

Table 6.8: Mean

This table reports the monthly out-of-sample mean of the different portfolios presence of both transaction costs and estimation error. Our numerical results on five empirical datasets show that the data-driven approach helps to improve the out-of-sample performance in terms of Sharpe ratio of returns net of transaction costs. The explanation for the favorable performance of the data-driven portfolios is that they strike an optimal (data-driven) trade-off between rebalancing the portfolio to capture the information in recent historical return data, and avoiding the large transaction costs and impact of estimation error associated with excessive trading. Essentially, the data-driven portfolios induce and slow and steady rate of trade that results in superior performance.

List of Figures

1.1 A simulated graph (left panel) together with its approximated counterpart (right panel), estimated by OLS. 4

1.2 Graph estimated using the LAR methodology. 6

1.3 Example of an uncertainty set for a non-deterministic parameter **u**. . . . 8

1.4 Demand time series generated from an autoregressive process with heavy-tailed errors, and real and predicted revenues under two classic approaches (classic newsvendor and AR). 14

1.5 Portfolio frontier 18

2.1 Graphical representation of a simulated VAR and its sparse counterparts. Naive algorithm (center) versus the proposed SC-VAR method (right). . 27

2.2 Median F_1 measure for Lasso/Group Lasso, greedy and SC-VAR with $\epsilon = 0, 0.2$ for simulated databases with $n = 1$ and $n = 2$ and different levels of densities. 37

2.3 Real graphs of randomly selected instances of processes with different densities, represented along with their VAR solutions. SC-VAR and Lasso solutions are represented for a couple of levels of sparsity, $V_T = 1$ and $V_T = 10$, and the SC-VAR is also depicted fixing solely $VA_{=}10$. 38

2.4 Median elapsed times in seconds taken by the SC-VAR to be solved for different requirements of the sparsity in VAR(1) and VAR(2) processes with various levels of density. 40

2.5 MSE (Panel A) and some heatmaps and graphs (Panel B) representing some solutions of the VAR, SC-VAR, and standard or Group Lasso for Flu and Airpollution databases. For the Flu database the SC-VAR and the standard Lasso are depicted for $V_T = 1$, while for the Airpollution dataset the SC-VAR and the Group Lasso outputs are allowing one ($V_S = 1$) or two ($V_S = 2$) causal features. 43

2.6 Heatmap representing the paths of solution for a fixed $V_T = 10$ and different values of ϵ together with its MSEs (right and left plots of Panel A, respectively), and heatmaps for different V_T with fixed $\epsilon = 0, 0.2$ (left and right plots of Panel B, respectively). 45

3.1 Path of solutions of the Lasso for the diabetes database. The size of the coefficients $\boldsymbol{\beta}$ are depicted against the number of non zeroes of the solution (NZ). 55

3.2 Path of solutions of the tightened Lasso for the diabetes database. . . . 58

4.1 Examples of highly correlated autoregressive demands generated with errors following $N(4, 1)$ and $LN(0, 3)$. 84

4.2 Empirical cdf of the revenue under the static (dotted line), $AR(1)$ forecasting method (dashed line), Scarf approach (dot-dashed line) and the $ARUS(1)$ approach (solid line), under normally (top), lognormally(central) and pareto (bottom) distributed error terms in the case of neutral products. 88

5.1 The top panels show highly correlated autoregressive demands, while the bottom panels show moving average demands, both generated with errors following $N(0, 1)$ and parameters $\theta_1 = 0.9$, $\alpha = 100$. Left panel contain the whole series, right panel the *unknown* future values for the demand . 98

5.2 The red lines denote the interval uncertainty sets for each instant of time for AR (top panels) and MA (bottom panels) demands. On the right panel, some realizations for the demand are shown with their standard deviations 99

5.3 Some realizations for the demand for uncertainty sets (5.1) and (5.2), together with their standard deviations. 100

5.4 Examples of high and low correlated autoregressive demands generated with errors following $N(0, 1)$ and $LN(0, 3)$. 104

5.5 Box plots of the obtained total revenues for the 88-node and Swain databases under the maximum profit criterion. The nodes' demands have been generated from an $AR(1)$ process with errors distributed according to a $N(0, 1)$ model. 108

5.6 Box plots of the obtained total revenues for the 88-node (up) and Swain (bottom) databases under the maximum profit criterion. The nodes' demands have been generated from an $AR(1)$ process with errors distributed according a $LN(0, 3)$ model. 109

5.7 Box plots of the obtained total revenues for the Swain database under both types of allocation and all values of p. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0, 3)$ model and $\theta = 0.5$. 112

5.8 Location and allocation of clients for AR forecasting technique of two randomly selected instances for **high-profit** products in the 88-node database under both criteria: maximum profit and minimum cost (production+transportation) allocation. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0, 3)$ model and $\theta = 0.5$. 114

5.9 Location and allocation of clients for AR forecasting technique of two randomly selected instances for **low-profit** products in the 88-node database under both criteria: maximum profit and minimum cost (production+transportation) allocation. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0, 3)$ model and $\theta = 0.5$. . . 115

6.1 Transaction cost level sets and mean return uncertainty sets 125

6.2 Power exponential density functions for different values of p and α. . . . 134

List of Tables

2.1 Normalized median MSE (and standard deviations) under the SC-VAR, greedy and the Lasso for VAR(1) processes, and the SC-VAR, greedy and the Group Lasso for VAR(2) processes, all generated with a 10, 50 and 90% density, for different requirements of sparsity in terms of number of causal features (V_S=1,...,10) 35

2.2 Summary of real databases used for numerical illustrations 41

2.3 Results after tuning the parameters for the adaptative lasso, SC-VAR with $\epsilon = 0.2$ and greedy approaches in real and simulated databases. Normalized median MSE and median number of non-zeroes are displayed for the simulated databases. 46

3.1 Real data sets specifications and sources 62

3.2 Median MSE and number of non-zeroes (NZ) of the ten random shuffles for the real data sets. 64

3.3 Median MSE and number of non-zeroes (NZ) of the ten simulated instances. 66

4.1 Average revenue and frequency of losses obtained for highly and moderately autocorrelated series under the four considered approaches (static, Scarf, classic $AR(1)$ forecasting method, and $ARUS(1)$), under Pareto, Lognormal and normally distributed error terms. 86

5.1 Recursive approach for estimating the parameters of the p-median problem. 102

5.2 Average percentage of improvement of the minimum cost allocation over the maximum profit allocation for the **Swain** databases. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0, 3)$ model. 110

5.3 Average percentage of improvement of the minimum cost allocation over the maximum profit allocation for the **88-node** databases. The nodes' demands have been generated from an $AR(1)$ process with errors distributed according a $LN(0,3)$ model. 111

5.4 Average percentage of clients allocated to the minimum cost open facility for the 88-node databases. The nodes' demands have been generated from an $AR(1)$ process with errors distributed following a $LN(0,3)$ model.113

6.1 Summary of main theoretical results. 135

6.2 Datasets used for comparison of portfolios. 136

6.3 Portfolios evaluated, where κ_{cv} is calibrated with cross-validation. . . . 137

6.4 Methodology used to compare portfolios 140

6.5 Sharpe ratios 141

6.6 Turnover 143

6.7 Variance 145

6.8 Mean 147

References

Abdel-Malek, L. L. and Montanari, R. (2005). On the multi-product newsboy problem with two constraints. *Computers & Operations Research*, 32(8):2095–2116.

Abegaz, F. and Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, 14(3):586–599.

Aghezzaf, E. (2005). Capacity planning and warehouse location in supply chains with uncertain demands. *Journal of the Operational Research Society*, 56:453–462.

Almgren, R., Thum, C., Hauptmann, E., and Li, H. (2005). Direct estimation of equity market impact. *Risk*, 18(7):58–62.

An, Y., Zeng, B., Zhang, Y., and Zhao, L. (2014). Reliable p-median facility location problem: two-stage robust models and algorithms. *Transportation Research Part B: Methodological*, 64:54–72.

Arnold, A., Liu, Y., and Abe, N. (2007). Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 66–75. ACM.

Atamturk, A., Berenguer, G., and Shen, Z. (2012). A conic integer programming approach to stochastic joint location-inventory problems. *Operations Research*, 60:366–381.

Aviv, Y. (2002). Gaining benefits from joint forecasting and replenishment processes: The case of auto-correlated demand. *Manufacturing & Service Operations Management*, 4(1):55–55.

Aviv, Y. (2003). A time-series framework for supply-chain inventory managemet. *Operations Research*, 51:210–227.

Baker, K. and Scudder, G. (1990). Sequencing with earliness and tardiness penalties. *Operations Research*, 38:22–36.

Balduzzi, P. and Lynch, A. W. (1999). Transaction costs and predictability: Some utility cost calculations. *Journal of Financial Economics*, 52(1):47–78.

Bandi, C. and Bertsimas, D. (2012). Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming*, 134:23–70.

Barahona, F. and Jensen, D. (1998). Plant location with minimum inventory. *Mathematical Programming*, 83:101–111.

Baron, O., Milner, J., and Naseraldin, H. (2011). Facility location: A robust optimization approach. *Production and Operations Management*, 20(5):772–785.

Barry, C. B. (1974). Portfolio analysis under uncertain means, variances, and covariances. *The Journal of Finance*, 29(2):515–522.

Bawa, V. S., Brown, S. J., and Klein, R. W. (1979). Estimation risk and optimal portfolio choice. *Amsterdam and New York: North Holland Publishing Co.*

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*. Princeton University Press.

Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805.

Ben-Tal, A. and Nemirovski, A. (1999). Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13.

Ben-Tal, A. and Nemirovski, A. (2000). Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88(3):411–424.

Ben-Tal, A. and Nemirovski, A. (2002). Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480.

Bertsimas, D., Brown, D. B., and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501.

Bertsimas, D. and Copenhaver, M. S. (2014). Characterization of the equivalence of robustification and regularization in linear, median, and matrix regression. `arXiv: 1411.6160`.

Bertsimas, D. and King, A. (2015). Or forum: An algorithmic approach to linear regression. *Operations Research.*

Bertsimas, D. and Mazumder, R. (2014). Least quantile regression via modern optimization. *The Annals of Statistics*, 42(6):2429–2525.

Bertsimas, D. and Thiele, A. (2006). A robust optimization approach to inventory theory. *Operations Research*, 54:150–168.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bimpikis, K. and Markakis, M. G. (2015). Inventory pooling under heavy-tailed demand. `http://www.econ.upf.edu/$\sim$mmarkakis/inventory_pooling.pdf`.

Blanquero, R., Carrizosa, E., Nogales-Gómez, A., and Plastria, F. (2014). Single-facility huff location problems on networks. *Annals of Operations Research*, 222(1):175–195.

Box, E., Jenkins, G., and Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance*, 54(2):655–671.

Broadie, M. (1993). Computing efficient frontiers using estimated parameters. *Annals of Operations Research*, 45:21–58.

Burer, S. and Letchford, A. N. (2012). Non-convex mixed-integer nonlinear programming: A survey. *Surveys in Operations Research and Management Science*, 17(2):97–106.

Camm, J. D., Raturi, A. S., and Tsubakitani, S. (1990). Cutting big M down to size. *Interfaces*, 20(5):61–66.

Cao, G., Guo, Y., and Bouman, C. A. (2010). High dimensional regression using the sparse matrix transform (smt). In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1870–1873. IEEE.

Caramanis, C., Mannor, S., and Xu, H. (2012). Robust optimization in machine learning. In Sra, S., Nowozin, S., and Wright, S. J., eds., *Optimization for machine learning*, pages 369–402. MIT Press, Michigan.

Carrizosa, E. and Guerrero, V. (2014). rs-sparse principal component analysis: A mixed integer nonlinear programming approach with vns. *Computers & Operations Research*, 52:349–354.

Carrizosa, E., Nogales-Gómez, A., and Morales, D. R. (2016a). Strongly agree or strongly disagree?: Rating features in support vector machines. *Information Sciences*, 329:256–273.

Carrizosa, E., Olivares-Nadal, A., and Ramírez-Cobo, P. (2014). Robust newsvendor problem with autoregressive demand. *Computers & Operations Research*, 68:123–133.

Carrizosa, E., Olivares-Nadal, A. V., and Ramírez-Cobo, P. (2013). Time series interpolation via global optimization of moments fitting. *European Journal of Operational Research*, 230(1):97 – 112.

Carrizosa, E., Olivares-Nadal, A. V., and Ramírez-Cobo, P. (2016b). A sparsity-controlled vector autoregressive model. `https://www.researchgate.net/publication/285576539_A_Sparsity-Controlled_Vector_Autoregressive_Model`.

Chatterjee, S. and Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Chen, L. and Zhang, L. (2010). A better three-factor model that explains more anomalies. *Journal of Finance*, 65(2):563–595.

Chen, Y. F., Xu, M., and Zhang, Z. G. (2009). Technical note: a risk-averse newsvendor model under the CVaR criterion. *Operations Research*, 57(4):1040–1044.

Cheng, H., Tan, P.-N., Gao, J., and Scripps, J. (2006). Multistep-ahead time series prediction. In *Advances in Knowledge Discovery and Data Mining*, pages 765–774. Springer.

Choi, S., Ruszczyński, A., and Zhao, Y. (2011). A multiproduct risk-averse newsvendor with law-invariant coherent measures of risk. *Operations Research*, 59(2):346–364.

Chopra, V. K. and Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19(2):6–11.

Constantinides, G. M. (1979). Multiperiod consumption and investment behavior with convex transactions costs. *Management Science*, 25(11):1127–1137.

Constantinides, G. M. (1986). Capital market equilibrium with transaction costs. *The Journal of Political Economy*, 94(4):842–862.

Cornuéjols, G., Nemhauser, G. L., and Wolsey, L. A. (1983). The uncapacitated facility location problem. Technical report, DTIC Document.

Coullard, C., Daskin, M., and Shen, Z. (2002). An inventory-location model: Formulation, solution algorithm and computational results. *Annals of Operations Research*, 100:86–106.

Dana, J. D. and Petruzzi, N. C. (2001). Note: The newsvendor model with endogenous demand. *Management Science*, 47(11):1488–1497.

Daskin, M. (1995). *Network and Discrete Location: Models, Algorithms and Applications*. John Wiley & Sons.

Daskin, M. S. and Owen, S. H. (1999). Two new location covering problems: The partial p-center problem and the partial set covering problem. *Geographical Analysis*, 31(3):217–235.

Davis, M. H. and Norman, A. R. (1990). Portfolio selection with transaction costs. *Mathematics of Operations Research*, 15(4):676–713.

Davis, R. A., Zang, P., and Zheng, T. (2012). Sparse vector autoregressive modeling. `arXiv:1207.0520`.

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009a). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009b). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953.

DeMiguel, V., Nogales, F. J., and Uppal, R. (2014). Stock return serial dependence and out-of-sample portfolio performance. *Review of Financial Studies*, 27(4):1031–1073.

Ding, X., Puterman, M. L., and Bisi, A. (2002). The censored newsvendor and the optimal acquisition of information. *Operations Research*, 50(3):517–527.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212.

Dogan, K. and Goetschalckx, M. (1999). A primal decomposition method for the integrated design of multi-period production-distribution systems. *IIE Transactions*, 31:1027–1036.

Dominici, F., Zeger, S. L., and Samet, J. M. (2000). A measurement error model for time-series studies of air pollution and mortality. *Biostatistics*, 1(2):157–175.

Dong, L. and Lee, H. L. (2003). Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand. *Operations Research*, 51(6):969–980.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46.

Draper, N. R., Smith, H., and Pownell, E. (1998). *Applied regression analysis*, volume 3rd edition. Wiley New York.

Eeckhoudt, L., Gollier, C., and Schlesinger, H. (1991). Increases in risk and deductible insurance. *Journal of Economic Theory*, 55:435–440.

Eeckhoudt, L., Gollier, C., and Schlesinger, H. (1995). The risk-averse (and prudent) newsboy. *Management Science*, 41(5):786–794.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Eichler, M. (2012). Causal inference in time series analysis. In *Causality*, chapter 22, pages 327–354. Wiley Series in Probability and Statistics.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669.

Fabozzi, F. J., Kolm, P. N., Pachamanova, D., and Focardi, S. M. (2007). *Robust portfolio optimization and management*. John Wiley & Sons.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101.

Farmer, J. D., Gerig, A., Lillo, F., and Waelbroeck, H. (2013). How efficiency shapes market impact. *Quantitative Finance*, 13(11):1743–1758.

Fourer, R., Gay, D., and Kernighan, B. W. (2002). *The AMPL book*. Duxbury Press, Pacific Grove.

Frazzini, A., Israel, R., and Moskowitz, T. J. (2015). Trading costs of asset pricing anomalies. *Fama-Miller Working Paper*.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, 16:499–511.

Gabrel, V., Murat, C., and Thiele, A. (2014). Recent advances in robust optimization: An overview. *European Journal of Operational Research*, 235(3):471–483.

Gallego, G. and Moon, I. (1993). The distribution free newsboy problem: review and extensions. *Journal of Operational Research Society*, 44:825–834.

Ganesh, M., Raghunathan, S., and Rajendran, C. (2014). The value of information sharing in a multi-product, multi-level supply chain: Impact of product substitution, demand correlation, and partial information sharing. *Decision Support Systems*, 58:79–94.

Garlappi, L., Uppal, R., and Wang, T. (2007). Portfolio selection with parameter and model uncertainty: A multi-prior approach. *Review of Financial Studies*, 20(1):41–81.

Gârleanu, N. and Pedersen, L. H. (2013). Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, 68(6):2309–2340.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.

Godfrey, G. A. and Powell, W. B. (2001). An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. *Management Science*, 47(8):1101–1112.

Goldfarb, D. and Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38.

Gorrostieta, C., Fiecas, M., Ombao, H., Burke, E., and Cramer, S. (2013). Hierarchical vector auto-regressive models and their applications to multi-subject effective connectivity. *Frontiers in computational neuroscience*, 7.

Gotoh, J.-Y. and Takeda, A. (2011). On the role of norm constraints in portfolio selection. *Computational Management Science*, 8(4):323–353.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438.

Graves, S. C. (1999). A single-item inventory model for a nonstationary demand process. *Manufacturing & Service Operations Management*, 1(1):50–61.

Griffin, M. P., Lake, D. E., and Moorman, J. R. (2005). Heart rate characteristics and laboratory tests in neonatal sepsis. *Pediatrics*, 115(4):937–941.

Güllü, R. (1997). A two-echelon allocation model and the value of information under correlated forecasts and demands. *European Journal of Operational Research*, 99(2):386–400.

Gulpinar, N., Pachamanova, D., and Canakoglu, E. (2013). Robust strategies for facility location under uncertainty. *European Journal of Operational Research*, 225:21–35.

Hadley, G. and Whitin, T. M. (1963). *Analysis of inventory systems.* Prentice Hall.

Han, Q., Du, D., and Zuluaga, L. F. (2014). Technical note: A risk-and-ambiguity-averse extension of the max-min newsvendor order formula. *Operations Research*, 62(3):535–542.

Hanasusanto, G. A., Kuhn, D., Wallace, S. W., and Zymler, S. (2014). Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming*, pages 1–32.

Hastie, T. and Efron, B. (2013). Least Angle Regression, Lasso and Forward Stagewise. `http://cran.r-project.org/web/packages/lars/lars.pdf`.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations.* CRC Press.

Haufe, S., Nolte, G., Mueller, K.-R., and Krämer, N. (2010). Sparse causal discovery in multivariate time series. *JMLR W&CP*, 6:97–106.

Helper, C. M., Davis, L. B., and Wei, W. (2010). Impact of demand correlation and information sharing in a capacity constrained supply chain with multiple-retailers. *Computers & Industrial Engineering*, 59(4):552–560.

Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C. (2008). Least angle and ℓ_1 penalized regression: A review. *Statistics Surveys*, 2:61–93.

Heyman, D. P. and Sobel, M. J. (2003). *Stochastic models in operations research: Stochastic optimization*, volume 2. Courier Corporation.

Higham, N. J. (2002). *Accuracy and stability of numerical algorithms.* SIAM.

Hsu, N.-J., Hung, H.-L., and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657.

Hu, J. and Hu, F. (2009). Estimating equation–based causality analysis with application to microarray time series data. *Biostatistics*, 10(3):468–480.

Huh, W. T., Levi, R., Rusmevichientong, P., and Orlin, J. B. (2011). Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research*, 59(4):929–941.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1684.

Jiang, H., Netessine, S., and Savin, S. (2011). Technical note: Robust newsvendor competition under asymmetric information. *Operations Research*, 59(1):254–261.

Johnson, G. and Thompson, H. (1975). Optimality of myopic inventory policies for certain dependent demand processes. *Management Science*, 21:1303–1307.

Jorion, P. (1986). Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21(03):279–292.

Kabak, I. and Schiff, A. (1978). Inventory models and management objectives. *Sloan Management Review*, 10:53–59.

Kahn, J. A. (1987). Inventories and the volatility of production. *The American Economic Review*, 77(7):667–679.

Kalli, M. and Griffin, J. E. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779–793.

Khouja, M. (1999). The single-period (news-vendor) problem: literature review and suggestions for future research. *Omega*, 27(5):537–553.

Kogan, L. and Tian, M. H. (2015). Firm characteristics and empirical factor models: a model-mining experiment. *FRB International Finance Discussion Paper*.

Kojima, K., Yamaguchi, R., Imoto, S., Yamauchi, M., Nagasaki, M., Yoshida, R., Shimamura, T., Ueno, K., Higuchi, T., Gotoh, N., et al. (2009). A state space representation of VAR models with sparse learning for dynamic gene networks. *Genome Informatics*, 22:56–68.

Kourtis, A. (2015). A stability approach to mean-variance optimization. *The Financial Review*, 50:301–330.

Kowalski, M. (2009). Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324.

Lau, H. (1980). The newsboy problem under alternative optimization objectives. *Journal of Operational Research Society*, 31:525–535.

Lau, H.-S. and Lau, A. H.-L. (1995). The multi-product multi-constraint newsboy problem: Applications, formulation and solution. *Journal of Operations Management*, 13(2):153–162.

LeBlanc, M. and Tibshirani, R. (1998). Monotone shrinkage of trees. *Journal of Computational and Graphical Statistics*, 7(4):417–433.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859.

Lee, H. L., So, K. C., and Tang, C. S. (2000). The value of information sharing in a two-level supply chain. *Management Science*, 46(5):626–643.

Lee, J. and Leyffer, S. (2012). *Mixed integer nonlinear programming*. Springer.

Levi, R., Roundy, R. O., Shmoys, D. B., et al. (2008). Approximation algorithms for capacitated stochastic inventory control models. *Operations Research*, 56(5):1184–1199.

Li, J. (2012). Monetary policy analysis based on Lasso-Assisted Vector Autoregression (Lavar). SSRN:2017877.

Liao, S., Hsieh, C., and Lai, P. (2011). An evolutionary approach for multi-objective optimization of the integrated location-inventory distribution network problem in vendor-managed inventory. *Expert Systems with Applications*, 38:6768–6776.

Lichman, M. (2016). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Lin, J. and Ng, T. S. (2011). Robust multi-market newsvendor models with interval demand data. *European Journal of Operational Research*, 212(2):361–373.

Lozano, A. C., Abe, N., Liu, Y., and Rosset, S. (2009). Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110–i118.

Lu, X., Song, J., and Regan, A. (2006). Inventory planning with forecast upadtes: Approximate solutions and cost error bounds. *Operations Research*, 54:1079–1097.

Lu, Z. (2011a). A computational study on robust portfolio selection based on a joint ellipsoidal uncertainty set. *Mathematical Programming*, 126(1):193–201.

Lu, Z. (2011b). Robust portfolio selection based on a joint ellipsoidal uncertainty set. *Optimization Methods & Software*, 26(1):89–104.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis.* Springer Science & Business Media.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256.

Melo, M., Nickel, S., and da Gama, F. S. (2009). Facility location and supply chain management: A review. *European journal of Operational Research*, 196:401–412.

Michaud, R. O. (1989). The Markowitz optimization enigma: is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42.

Miller, A. (2002). *Subset selection in regression*, volume 2nd edition. Chapman & Hall.

Mirchandani, P. and Francis, R. (1990). *Discrete Location Theory*, volume 1. Wiley-Interscience, New York.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2015). *Introduction to linear regression analysis.* John Wiley & Sons.

Muthuraman, K. and Kumar, S. (2006). Multidimensional portfolio optimization with proportional transaction costs. *Mathematical Finance*, 16(2):301–335.

Natarajan, K., Pachamanova, D., and Sim, M. (2009). Constructing risk measures from uncertainty sets. *Operations Research*, 57(5):1129–1141.

Ozsen, L., Daskin, M., and Coullard, C. (2009). Facility location modeling and inventory management with multisourcing. *Transportation Science*, 43:455–472.

Pástor, Ľ. (2000). Portfolio selection and asset pricing models. *The Journal of Finance*, 55(1):179–223.

Pástor, Ľ. and Stambaugh, R. F. (2000). Comparing asset pricing models: An investment perspective. *Journal of Financial Economics*, 56(3):335–381.

Peña, D. and Poncela, P. (2006). Dimension reduction in multivariate time series. In *Advances in Distribution Theory, Order Statistics, and Inference*, pages 433–458. Springer.

Peña, D. and Sánchez, I. (2007). Measuring the advantages of multivariate vs. univariate forecasts. *Journal of Time Series Analysis*, 28(6):886–909.

Perakis, G. and Roels, G. (2008). Regret in the newsvendor model with partial information. *Operations Research*, 56:188–203.

Petruzzi, N. C. and Dada, M. (1999). Pricing and the newsvendor problem: A review with extensions. *Operations Research*, 47(2):183–194.

Qin, Y., Wang, R., Vakharia, A. J., Chen, Y., and Seref, M. M. (2011). The newsvendor problem: Review and directions for future research. *European Journal of Operational Research*, 213(2):361–374.

Raghunathan, S. (2003). Impact of demand correlation on the value of and incentives for information sharing in a supply chain. *European Journal of Operational Research*, 146(3):634–649.

Reyman, G. (1989). State reduction in a dependent demand inventory model given by a time series. *European Journal of Operational Research*, 41:174–180.

Rish, I. and Grabarnik, G. (2014). *Sparse modeling: theory, algorithms, and applications*. CRC Press.

Rosenfield, D. (1986). Optimal management of tax-sheltered employee reimbursement programs. *Interfaces*, 16:68–72.

Saltelli, A., Chan, K., Scott, E. M., et al. (2000). *Sensitivity analysis*, volume 1. Wiley New York.

Scarf, H. (1958). A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209.

Schwartz, J. (1954). The formula for change in variables in a multiple integral. *American Mathematical Monthly*, pages 81–85.

Schweitzer, M. and Cachon, G. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science*, 46(3):404–420.

See, C.-T. and Sim, M. (2010). Robust approximation to multiperiod inventory management. *Operations Research*, 58(3):583–594.

Shahabi, M., Unnikrishnan, A., Jafari-Shirazi, E., and Boyles, S. D. (2014). A three level location-inventory problem with correlated demand. *Transportation Research Part B: Methodological*, 69:1–18.

Shang, K. H. and Song, J.-S. (2003). Newsvendor bounds and heuristic for optimal policies in serial supply chains. *Management Science*, 49(5):618–638.

Shen, Z., Coullard, C., and Daskin, M. (2003). A joint location-inventory model. *Transportation Science*, 37:40–55.

Shen, Z. and Qi, L. (2007). Incorporating inventory and routing costs in strategic location models. *European Journal of Operational Research*, 179:372–389.

Shojaie, A. and Michailidis, G. (2010). Discovering graphical Granger causality using the truncating Lasso penalty. *Bioinformatics*, 26(18):i517–i523.

Shu, J., Teo, C., and Shen, Z. (2005). Stochastic transportation-inventory network design problem. *Operations Research*, 53:48–60.

Sjöstrand, K., Clemmensen, L. H., Larsen, R., and Ersbøll, B. (2012). Spasm: A matlab toolbox for sparse statistical modeling.

Snyder, L., Daskin, M., and Teo, C. (2007). The stochastic location model with risk pooling. *European Journal of Operational Research*, 179:1221–1238.

Snyder, L. V. (2006). Facility location under uncertainty: a review. *IIE Transactions*, 38(7):547–564.

So, K. C. and Zheng, X. (2003). Impact of supplier's lead time and forecast demand updating on retailer's order quantity variability in a two-level supply chain. *International Journal of Production Economics*, 86(2):169–179.

Song, S. and Bickel, P. J. (2011). Large vector auto regressions. `arXiv:1106.3915`.

Swain, R. (1971). *A decomposition algorithm for a class of facility location algorithms.* Tesis Doctoral, Cornell University, Ithaca, N.Y.

Thakkar, R., Finley, D., and Liao, W. (1983). A stochastic demand CVP model with return on investment criterion. *Contemporary Accounting Research*, 1:77–86.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Torgo, L. (2016). Regression data sets. `http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html`.

Torre, N. (1997). Barra market impact model handbook. *BARRA Inc., Berkeley.*

Vairaktarakis, G. L. (2000). Robust multi-item newsboy models with a budget constraint. *International Journal of Production Economics*, 66(3):213–226.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., and Canales-Rodríguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):969–981.

Wang, T., Atasu, A., and Kurtulus, M. (2012). A multiordering newsvendor model with dynamic forecast evolution. *Manufacturing & Service Operations Management*, 14:472–484.

Wang, Z., Yao, D.-Q., and Huang, P. (2007). A new location-inventory policy with reverse logistics applied to B2C e-markets of China. *International Journal of Production Economics*, 107(2):350–363.

Watson, P. K. and Teelucksingh, S. S. (2002). *A practical introduction to econometric methods: Classical and modern*. University of West Indies Press.

Weatherford, L. and Pfeifer, P. (1994). The economic value of using advance booking of orders. *Omega*, 22:105–111.

Weaver, J. R. and Church, R. L. (1986). A location model based on multiple metrics and multiple facility assignment. *Transportation Research Part B: Methodological*, 20(4):283–296.

Winner, L. (2016). Miscellaneous data sets. `http://www.stat.ufl.edu/~winner/datasets.html`.

Yang, Y. and Zou, H. (2015). Group Lasso Penalized Learning Using A Unified BMD Algorithm. `https://cran.r-project.org/web/packages/gglasso/gglasso.pdf`.

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.

Yue, J., Chen, B., and Wang, M. (2006). Expected value of distribution information for the newsvendor problem. *Operations Research*, 54(6):1128–1136.

Zhang, B. (2012). Multi-tier binary solution method for multi-product newsvendor problem with multiple constraints. *European Journal of Operational Research*, 218(2):426–434.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497.

Zhu, Z., Zhang, J., and Ye, Y. (2013). Newsvendor optimization with limited distribution information. *Optimization Methods and Software*, 28:640–667.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.