Pontifícia Universidade Católica
do Rio de Janeiro

Mario Henrique Alves Souto Neto

Sparse Statistical Modelling with Applications to Renewable Energy and Signal Processing

DISSERTAÇÃO DE MESTRADO

Dissertation presented to the Programa de Pós-Graduação em Engenharia Elétrica of the Departamento de Engenharia Elétrica, PUC–Rio as partial fulfillment of the requirements for the degree of Mestre em Engenharia Elétrica.

Advisor: Prof. Álvaro de Lima Veiga Filho

Rio de Janeiro
July 2014

Pontifícia Universidade Católica
do Rio de Janeiro

Mario Henrique Alves Souto Neto

Sparse Statistical Modelling with Applications to Renewable Energy and Signal Processing

Dissertation presented to the Programa de Pós-Graduação em Engenharia Elétrica of the Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC–Rio, as partial fulfillment of the requirements for the degree of Mestre.

Prof. Álvaro de Lima Veiga Filho
Advisor
Departamento de Engenharia Elétrica — PUC–Rio

Prof. Matthias Kormaksson
IBM Research Brasil

Prof. Marcelo Cunha Medeiros
Departamento de Economia — PUC–Rio

Prof. Davi Michel Valladão
Departamento de Engenharia Industrial — PUC–Rio

Prof. José Eugenio Leal
Coordinator of the Centro Técnico Científico — PUC–Rio

Rio de Janeiro, July 30th, 2014

Mario Henrique Alves Souto Neto

Mario Henrique Alves Souto Neto graduated in Industrial Engineering from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil. During his MSc course, he has been actively participating on projects of the Laboratory of Applied Mathematical Programming and Statistics (LAMPS) at the Electrical Engineering Department of PUC-Rio.

Bibliographic data

Alves Souto Neto, Mario Henrique

Sparse Statistical Modelling with Applications to Renewable Energy and Signal Processing / Mario Henrique Alves Souto Neto ; advisor: Álvaro de Lima Veiga Filho — 2014.
74 f. : il. ; 30 cm

Dissertação (Mestrado em Engenharia Elétrica)-Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, Rio de Janeiro, 2014.
Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Estatística em alta dimensão;. 3. LASSO;. 4. Regularização;. 5. Processamento de sinais esparsos;. 6. Modelagem de energia renovável;. 7. Energia eólica;. 8. PCH;. 9. Monitoramento de fibras ópticas.. I. de Lima Veiga Filho, Álvaro. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Acknowledgments

I would like to dedicate this thesis to my grandmother Alda de Lourdes Alves Souto. Furthermore, I would like to thank everyone who directly or indirectly have contributed for this work.

A very special thanks goes out to my parents, Mario José Alves Souto and Léa Maria da Silva, and my sister Jéssica Silva Lustosa, for the support they provided me.

I would also like to thank Júlia Camargo, for being always supportive and an example for me.

I am grateful to the professors Álvaro Veiga, Cristiano Fernandez, Carlos Kubrusly, Eduardo Laber and Thomas Lewiner for their excellent technical assistance and inspiring lectures.

I must also acknowledge my friends Alexandre Moreira, Henrique Helfer and Joaquim Garcia, for the fundamental suggestions and for useful discussions.

I would like to thank the members of the LAMPS laboratory, Alexandre Street, Andrea Alzuguir, Ana Luiza Lopes, Aderson Passos, Arthur Brigatto, Bruno Fanzeres, Davi Michel Valladão, Gustavo Amaral, Gustavo Ayala, Lucas Freire and Sebastian Maier.

I want to express my gratitude to the Electrical Engineering Department, PUC-Rio and CNPq, for the oportunity and financial support which was essencial for the development of my research.

Abstract

Alves Souto Neto, Mario Henrique; de Lima Veiga Filho, Álvaro (Advisor). **Sparse Statistical Modelling with Applications to Renewable Energy and Signal Processing**. Rio de Janeiro, 2014. 74p. MSc Dissertation — Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Motivated by the challenges of processing the vast amount of available data, recent research on the flourishing field of high-dimensional statistics is bringing new techniques for modeling and drawing inferences over large amounts of data. Simultaneously, other fields like signal processing and optimization are also producing new methods to deal with large scale problems. More particularly, this work is focused on the theories and methods based on ℓ_1 -regularization.

After a comprehensive review of the ℓ_1-norm as tool for finding sparse solutions, we study more deeply the LASSO shrinkage method. In order to show how the LASSO can be used for a wide range of applications, we exhibit a case study on sparse signal processing. Based on this idea, we present the ℓ_1 level-slope filter. Experimental results are given for an application on the field of fiber optics communication.

For the final part of the thesis, a new estimation method is proposed for high-dimensional models with periodic variance. The main idea of this novel methodology is to combine sparsity, induced by the ℓ_1-regularization, with the maximum likelihood criteria. Additionally, this novel methodology is used for building a monthly stochastic model for wind and hydro inflow. Simulations and forecasting results for a real case study involving fifty Brazilian renewable power plants are presented.

Keywords

High-Dimensional Statistics; LASSO; ℓ_1 regularization; Sparse Signal Processing; Renewable Energy Stochastic Modelling; Wind Energy; Hydro Energy; Optical Fiber Monitoring; Big Data.

Resumo

Alves Souto Neto, Mario Henrique; de Lima Veiga Filho, Álvaro (Orientador). **Modelagem Estatística Esparsa com Aplicações em Energia Renovável e Processamento de Sinais**. Rio de Janeiro, 2014. 74p. Dissertação de Mestrado — Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Motivado pelos desafios de processar a grande quantidade de dados disponíveis, pesquisas recentes em estatística tem sugerido novas técnicas de modeloagem e inferência. Paralelamente, outros campos como processamento de sinais e otimização também estão produzindo métodos para lidar problemas em larga escala. Em particular, este trabalho é focado nas teorias e métodos baseados na regularização ℓ_1.

Após uma revisão compreensiva da norma ℓ_1 como uma ferramenta para defenir soluções esparsas, estudaremos mais a fundo o método LASSO. Para exemplificar como o LASSO possui uma ampla gama de aplicações, exibimos um estudo de caso em processamento de sinal esparso. Baseado nesta idea, apresentamos o ℓ_1 level-slope filter. Resultados experimentais sa õ apresentados para uma aplicação em transmissão de dados via fibra óptica.

Para a parte final da dissertação, um novo m étodo de estimação é proposto para modelos em alta dimensão com variância periódica. A principal ideia desta nova metodologia é combinar esparsidade, induzida pela regularização ℓ_1, com o método de máxima verossimilhança. Adicionalmente, esta metodoogia é utilizada para estimar os parâmetros de um modelo mensal estocástico de geração de energia eólica e hídrica. Simulações e resultados de previsão são apresentados para um estudo real envolvendo cinquenta geradores de energia renovável do sistema Brasileiro.

Palavras–chave

Estatística em alta dimensão; LASSO; Regularização; Processamento de sinais esparsos; Modelagem de energia renovável; Energia eólica; PCH; Monitoramento de fibras ópticas.

Contents

List of Figures

List of Tables

Where would you go
Where would you go with a lasso?

Phoenix, *Wolfgang Amadeus Phoenix.*

1
Introduction

In several businesses and industries data storage has grown exponentially over the last decade. Many factors are associated with this increasing trend of data collection. Firstly, the decrease of storage and hardware prices has allowed the presence of necessary infrastructure. Secondly, the up-rise of new tools specifically designed to handle massive volume of data has led to more efficient ways of managing such data bases. All of these different technologies have been brought together with the popular term *Big Data*.

On a white paper [1] dated from 2008, Google reported that it processes more than twenty-five Peta Bytes, i.e. 25×10^{15} bytes, of data. That is approximately twenty million novels per day of data, that goes from user searches to satellite images. Another example of an astonishing data volume are the experiments taken on the Large Hadron Collider (LHC) at CERN. With the support from 150 million sensors one Peta Byte of data are measured for each experiment [2].

At first sight, having such a large amount of available data might look like all good news. It is intuitive to believe that more information is better. However, such an abundance of data poses a compelling challenge for data analysis. The demand for extracting valuable information from these massive data sets takes the classical frameworks to their limits and generates a need for new tools and methods.

This work starts by exposing the limitations of the traditional methods when dealing with a high-dimensional framework. Next, we present a comprehensive review of ℓ_1-regularization and the LASSO. In the second half, new applications and methodologies are proposed as this work contribution.

Chapter 2 begins by introducing the concept of high-dimensional statistics based on the "$n < p$" concept. Then, we briefly summarize the theory of linear regression. In the following, we explain why the small n and large p problem poses a challenge to traditional statistical methods such as least squares. Finally, some alternative techniques, designed to tackle large scale data, are presented.

Chapter 3 focuses on sparse solutions as a primary tool to deal with

overdetermined systems. We start by showing how to model an optimization problem to search for a sparse solution to a noise-free system. Next, we describe conditions that guarantee the uniqueness of these solutions. Since searching for sparse solutions is computationally exhausting, we introduce the ℓ_1-norm as an alternative. The previous conditions are then extended to the relaxed problem. To finish the chapter, we show a geometrical interpretation for the ℓ_1-regularization in the presence of noise.

Chapter 4 is exclusively dedicated to the LASSO shrinkage method. After the definition, we show how sparsity can be controlled by properly choosing a tuning parameter. The second half of the chapter analyses three different algorithms for solving the LASSO.

In the last two chapters, the real contribution of this work is presented. Chapter 5 presents an application on sparse signal processing. Where, the ℓ_1-regularization is used to design a piece-wise linear filter, which benefits from sparsity. Next, we show how useful the filter can be for monitoring fiber-optic communication systems. More particularly, we draw an algorithm for detecting the location of a failure on fiber optic cables. Experimental results are exposed to validate the methodology.

In the Chapter 6, we propose a novel estimation method for multiresponse high-dimensional models under heteroskedasticity. Based on the LASSO, we suggest a maximum-likelihood estimation method for a VARX model with periodic variance. Additionally, we imply this technique for building a model for simulating and forecasting renewable energy supply. Relevant results are presented for a subset of the Brazilian power system.

2
High-Dimensional Statistics

In classical statistics, many results rely on the fact that the number of unknown parameters p is fixed and the sample size $n \to \infty$. With this framework the asymptotic properties of estimators are evaluated. However, in many large-scale problems both $n \to \infty$ and $p \to \infty$. On several cases, the number of unknown parameters might be even greater than the number of observations, i.e. $n \ll p$. The recent field of *High-Dimensional Statistics* seeks to extend classical approaches as well as propose new methods in order to enable inference when the number of unknown parameters is much greater than the number of observations.

Furthermore, given the large amount of data, every algorithm used for data analysis must be efficient with respect to the input size. Otherwise, running any survey would be impracticable for real-life applications. In this sense, using a polynomial time algorithm or the application of more suitable data structures could enable solving extremely large problems [3].

This thesis will focus on the *linear regression* problem under a high-dimensional framework. However, the results can be extended for *logistic regression* or *autoregressive models*.

2.1
Linear Regression

In several fields, linear regression is the most traditional tool for modeling the linear relationship between variables of interest. Linear regression was originally proposed by Legendre [4] and Gauss [5] for astronomy studies, e.g. establishing the motion of bodies around the sun. Over the years, it has expanded to a wide range of research areas like epidemiology, engineering and social sciences. Particularly, in the last decade, it has gained a great attention given the rise of the field of econometrics [6].

The traditional framework is expressed by a vector of observations $y \in \mathbb{R}^n$, often called endogenous variables, which are assumed to be explained by linear measurements on the design matrix $X \in \mathbb{R}^{n \times p}$. The matrix X is composed by a concatenation of column vectors denoted by $x_1, \cdots, x_p$. Each x_i

is an explanatory variable, also known as exogenous variable. The relationship bertween y and X is assumed to be linear (2-1).

The coefficients are represented by a vector $\beta \in \mathbb{R}^p$,where each element β_i of the vector is a weight given by the explanatory variable x_i. Furthermore, each coefficient β_i is the partial effect of the corresponding variable holding all other explanatory variables fixed. This concept is very useful and can be interpreted on how y will behave with respect to a one-unit change on x_i.

$$y = X\beta + \varepsilon \tag{2-1}$$

For a given sample of y and X, the *Ordinary Least Squares* (OLS) establish a criteria to determine the vector β that "best" fits the data. As a simple example, consider a simple linear regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \; \forall i = 1, \cdots, n$. By minimizing the sum of squared error $\sum_{i=1}^{n} \varepsilon_i^2$, the OLS selects the values for β_0 and β_1 which minimizes the vertical distances between the data and the responses predicted by the linear model Fig. [2.1].

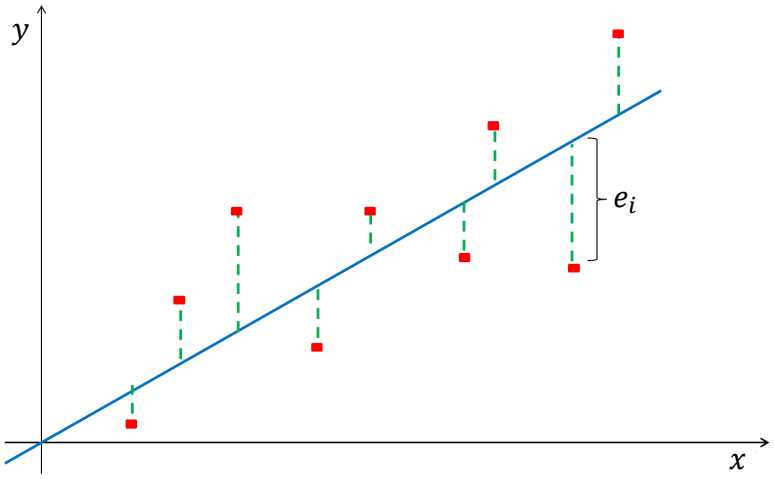

Figure 2.1: Leasts squares fit for simple regression $y_i = \beta_1 x_i + \varepsilon_i$.

$$\min_{\beta} \varepsilon^\top \varepsilon = \min_{\beta} (Y - X\beta)^\top (Y - X\beta) \tag{2-2}$$

The argument which minimizes (2-2), the OLS estimator $\hat{\beta}_{OLS}$, can be found by taking the first derivative with respect to β and setting to zero.

$$\frac{\partial (Y - X\beta)^\top (Y - X\beta)}{\partial \beta} = -2X^\top (Y - X\beta) = 0 \tag{2-3}$$

$$\implies \hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y \tag{2-4}$$

The OLS estimator (2-4) can be efficiently computed without necessary obtaining the inverse of $X^\top X$ via matrix decompositions techniques. The most common strategy is to do the QR factorization, using Householder reflections [7], of the design matrix X. Another attractive approach, is the Cholesky decomposition of $X^\top X$. This method is asymptotically more exhausting, but may be preferable under certain conditions.

The estimator is a function of random variables and therefore is a random variable itself. Thus, to establish metrics of comparison and evaluation for estimators it is necessary to assess some properties of this random variable. Most commonly this properties are based on the central moments and their asymptotic behavior. Before developing these properties some hypothesis must be considered for the noise ε. The expected value of it is equal to zero, i.e. $\mathbb{E}[\varepsilon_i] = 0 \, \forall i = 1, \cdots, n$. Different realizations of the noise are uncorrelated. In other words, the variance-covariance matrix of ε is diagonal with σ on the diagonal. Additionally, the variance must be finite $\sigma < \infty$.

First, it can be easily verified by (2-5) that the OLS estimator is *unbiased*. In other worlds, the difference between the expected value of $\hat{\beta}_{OLS}$ and its real value is equal to zero. The bias of an estimator can be interpreted as an accuracy metric on the capability of recovering the real parameter value. From classical statistical inference unbiased estimators are preferable to a biased estimator. However, as will be shown later, unbiased estimator may be useful under certain circumstances.

$$\begin{aligned}
&\mathbb{E}[\hat{\beta}_{OLS}] = \mathbb{E}[(X^\top X)^{-1} X^\top Y] \\
&= \mathbb{E}[(X^\top X)^{-1} X^\top (X\beta + \varepsilon)] \\
&= \beta + (X^\top X)^{-1} X^\top \mathbb{E}[\varepsilon] \\
&= \beta
\end{aligned} \tag{2-5}$$

Besides accuracy, another key property is the variability of an estimator (2-6). After obtaining the variance of the estimator $\Sigma_{\hat{\beta}_{OLS}}$, it is possible to infer on how precise the estimate will be. In this sense, the smaller the variance the "better" the estimator will perform.

$$\begin{aligned}
&Var[\hat{\beta}_{OLS}] = \Sigma_{\hat{\beta}_{OLS}} = \mathbb{E}[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)^\top] \\
&= \mathbb{E}[((X^\top X)^{-1} X^\top (X\beta + \varepsilon) - \varepsilon)((X^\top X)^{-1} X^\top (X\beta + \varepsilon) - \varepsilon)^\top] \\
&= (X^\top X)^{-1} X^\top \mathbb{E}[\varepsilon\varepsilon^\top] X (X^\top X)^{-1} \\
&= \sigma^2 (X^\top X)^{-1}
\end{aligned} \tag{2-6}$$

According to the *Gauss-Markov Theorem* the OLS estimator is the linear

unbiased estimator with the lowest variance among all unbiased estimators of β [8]. This result is very important and is the greatest argument for the use of OLS estimators.

There are several ways of comparing estimators based on their properties. The *Mean Square Error* (MSE) is the most commonly used metric to evaluate the efficiency of an estimator. Under general conditions, the estimator with the smallest MSE is preferable over any estimator. The main idea behind MSE is to measure the average performance of the estimator $\hat{\beta}$ under several repeated samplings of X. This measure is done by taken the expected value of the square difference between the estimator and the parameter's real value. As it is shown bellow, the MSE can be decomposed into a bias and a variance component. This decomposition will be very useful when developing an intuition for choosing unbiased estimators.

$$\begin{aligned}
&\text{MSE}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \beta)^2] \\
&= \mathbb{E}[((\hat{\beta} - \mathbb{E}[\hat{\beta}]) + (\mathbb{E}[\hat{\beta}] - \beta))^2] \\
&= Var[\hat{\beta}] + Bias^2[\hat{\beta}] + 2(\mathbb{E}[\hat{\beta}]^2 - \mathbb{E}[\hat{\beta}]^2 - \beta\mathbb{E}[\hat{\beta}] + \mathbb{E}[\hat{\beta}]\beta) \\
&= Var[\hat{\beta}] + Bias^2[\hat{\beta}] \qquad (2\text{-}7)
\end{aligned}$$

Since the $\hat{\beta}_{OLS}$ is unbiased (2-5) and has the lowest variance among the linear unbiased estimators, the OLS estimator is said to be BLUE (*Best Linear Unbiased Estimator*). In this sense, at first sight, there is no apparent reason for searching for any other linear estimator. However, as it is going to be explored on the next section, under a high-dimensional setting the OLS estimator is inappropriate.

2.2 Consequences of High-Dimensionality

Considering an explicit high-dimensional case where $p > n$, i.e. the number of unknowns is greater then the number of equations, the OLS estimator will necessarily suffer from the curse of dimensionality. In other words, the model will be excessively complex, compromising the uniqueness of the estimator. Furthermore, the model will *over-fit* the sample data and probably will have poor out-of-sample predictive performance.

Theorem 2.1. *If $p > n$ no unique ordinary leasts squares solution exists.*

Proof. Since $X \in \mathbb{R}^{n \times p}$, the rank of the design matrix can be at most equals to n. Additionally, from linear algebra it is known that $\text{rank}(X^\top X) = \text{rank}(X)$. Considering that $(X^\top X) \in \mathbb{R}^{p \times p}$ the matrix $(X^\top X)$ will never have full

rank and consequently not have an inverse. Therefore, the OLS estimator $\hat{\beta}^{ols} = (X^\top X)^{-1} X^\top y$ will not exist. □

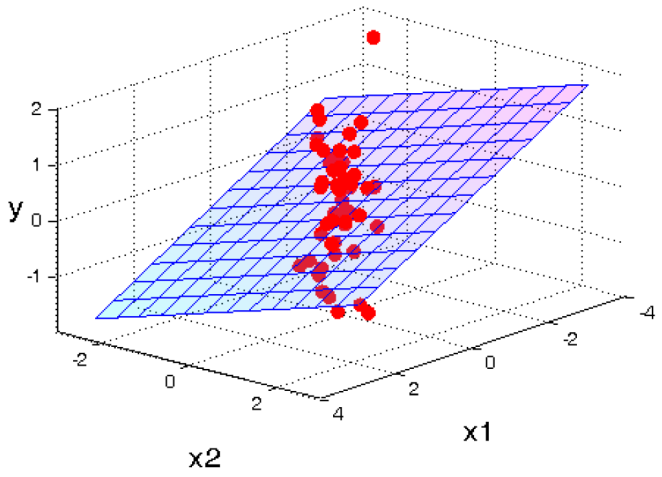

Figure 2.2: Ilustrative example of rank deficient regression.

The Fig[2.2] seeks to highlight the idea of a rank-deficient regression. Since a truly high-dimensional case can not be illustrated, consider the bivariate regression $y = X\beta + \varepsilon$ where the matrix $X = [x_1 \;\; x_2]$ has rank equals to one. As a consequence, the data is spread over a straight-line in the three-dimensional space and there are infinite ways of fitting a plane to the data.

A common alternative to overcome the fact that the matrix $(X^\top X)$ is rank-deficient is to compute the Moore-Penrose pseudo-inverse [9]. Unfortunately, this method is highly unstable with respect to the data, even small perturbations on the matrix X can lead to extremely different solutions. A more robust strategy consists of selecting a more parsimonious model to fit the data. Less complex models are often preferable since they offer some interpretability and present less-variance. Next sections describe three of the most popular techniques for model selection.

2.3 Subset Selection

Among the model selection methods, *subset selection* is the most intuitive of them all. The main idea relies on seeking a subgroup of k variables out of a set of p potential explanatory variables. Ideally, k should be considerably smaller then p, leading to a parsimonious model. The quantity k needs to be established by some information criteria [10] or goodness of fit [11][12].

In a high-dimensional framework, evaluating all possible subsets is computationally infeasible. To overcome this difficulty, most of the algorithms are

based on different heuristics like *greedy* search, e.g. *Backward-stepwise*, *Mallow's Cp* [12], *Forward-Stepwise regression* [13], *Forward-Stagewise regression* and *Autometrics* [14] [15]. The main idea behind those methods is to sequentially insert or remove a variable to the model via least squares.

These kinds of *hard-thresholding* techniques usually suffer from instability and are often extremely data-driven. Additionally, there is no guarantee that any of the aforementioned greedy methods will select the optimal subset.

2.4 Principal Components

The purpose of *Principal Components Analysis* (PCA)[16] is to perform a coordinate transformation on the data matrix X in a way that the data is orthogonal in this new system of coordinates. The result is a uncorrelated set $\{z_1, \cdots, z_k\}$ of linear combinations of the columns of X. This orthogonal set is called the principal components of X and the cardinality k is equal to or smaller then p. The principal components are usually obtained by *Singular Value Decomposition* [7].

This method is useful to reduce the dimensionality of a linear regression and is robust to multicollinearity. First of all, one can apply PCA on the design matrix X and preserve a representative subset containing only the first m principal components. Subsequently, the model is built using $\{z_1, \cdots, z_m\}$ as explanatory variables for the response y.

The major problem with this method, is that there is no special reason for the first m principal components to be the best regressors for y. The ideal setting corresponds to selecting the components with high-variance that properly explain the response. The *Supervised Principal Components* [17] is an algorithm based on this idea. Basically, the algorithm selects the first principal components that have a minimum correlation with the endogenous variable y. This technique is also used as a *pre-conditioning* [18] for shrinkage methods.

2.5 Shrinkage methods

Shrinkage methods are based on different forms of limiting the feasible region of the coefficients β. This restraint can be achieved by adding constraints to the estimation problem, e.g. by penalizing the objective function. In shrinkage methods, as opposed to subset selection, exogenous variables are gradually inserted on the model. This technique is also known as soft-thresholding on the signal processing literature [19].

The development of efficient algorithms and solvers, associated with the increasing computational power, led to the popularization of shrinkage based techniques. Additionally, the establishment of theoretical results has provided optimality conditions for most of the methods. For these reasons, new shrinkage techniques and extensions are constantly being proposed in the literature.

One of the pioneers on shrinkage methods was *Ridge regression* [20]. Then the *non-negative garrotte* [21] was introduced, which influenced the most popular of them, the LASSO [22]. In turn, the LASSO has inspired several derivatives and extensions, like the *Graphical* LASSO [23] and *SCAD* [24]. More recently, the *Dantzig Selector* [25] is gaining considerable attention. Every one of these methods has advantages and disadvantages. Depending on the application, one method may be more suitable than another. In this work, we will focus on applications and properties of the LASSO estimator.

3
ℓ_1 Regularization

From classical linear algebra it is known that if X has full-rank and $p \leq n$, then the system $y = X\beta$ is either determined or overdetermined. In the first case, the system has a unique solution and can be efficiently computed by several well-known algorithms, such as Gaussian elimination, matrix decompositions or Simplex. On the contrary, if $p > n$ the system is undetermined, i.e. there are more equations than unknowns, then the system has either an infinite number of solutions or no solution at all. In this context, to find a unique solution, additional information needs to be given. In this work, we will assume the hypothesis that the solution β is sparse.

Over the last twenty years, methods based on sparse solutions has caught great attention from the academia. Due to the impressive success of these methods, such as the LASSO [22] and Compressive Sensing [26], a great effort has been made to build interesting theoretical results. This chapter offers a small glimpse of the underlying theory, in order to explain why such methods work so well and under which conditions they are supposed to work. Additionally, we explore the geometric intuition intrinsic to the ℓ_1-norm and its interpretation. Most of the presented proofs and figures are based on [27][28][29][30].

To develop a basic intuition on sparse solutions, consider the classic twelve-coin problem. There is a set of twelve coins, where eleven of them are of equal weight. The goal is to identify which of the coins has a different weight and discover if it weights more or less then the others, using a balance the least number of times as possible. The naive solution would be to compare all possible pair of coins, i.e. $\binom{12}{2} = \frac{12!}{10!2!} = 66$. However, this problem has been published along with a solution stating that three measurements are sufficient [31]. Since three out of sixty-six measures are taken, the solution can be considered sparse. This puzzle points out how sparse measures might contain all the necessary information.

Additionally, in several applications it seems natural to assume that the true value of β is sparse. As in the case of image compression, such as the JPEG2000 protocol [32], or even on biostatistics, where thousands of genes

are potential explanatory variables for a disease [33]. As a matter of fact, there is a growing literature with several applications where the sparsity of the regressors is expected and quite intuitive.

3.1 Uniqueness of the sparsest solution

Suppose we want to recover a signal $\beta \in \mathbb{R}^p$ taking measures of the form $y = X\beta$. For a given vector of observations $y \in \mathbb{R}^n$ and a measurement matrix $X \in \mathbb{R}^{n \times p}$, where X has much fewer rows than columns ($n \ll p$). Additionally, X is normalized to unit ℓ_2-norm, i.e. all columns of X have variance equal to one. Since the system is extremely underdetermined, one possibility is to assume that the signal β is a sufficiently sparse vector β_0.

First of all, it is necessary to quantify the concept of sparsity. In the context of the linear system $y = X\beta$, it can be computed using the ℓ_0-norm, also known as counting norm, which is defined as the number of nonzero entries in the vector $||\beta||_0 = \#\{j : \beta_j \neq 0\}$. Despite being called a norm, the ℓ_0-norm is not even a pseudonorm since it is not *absolutely homogeneous* [34], still, it preserves the triangle inequality. One could obtain the sparsest solution for an underdetermined system of equations by solving the following problem.

$$\begin{aligned} (P_0) : \text{minimize } & ||\beta||_0 \\ \text{subject to:} & \\ X\beta &= y \end{aligned} \tag{3-1}$$

Under certain conditions on the design matrix X and the sparsity of the signal β, the recovery of β can be obtained by solving the convex relaxation (P_1). Such conditions ensure that the solution of (P_0) and (P_1) are the same and also that both are unique. There are several conditions, such as the *Restricted Isometry Property*[35][36] or the *Exact Recovery Condition* [37]. In this work we will focus on the matrix *Coherence* since this property can be easily verified. To begin with, we are going to investigate the necessary conditions for the uniqueness of a solution for (P_0).

Before establishing any conditions, it is necessary to introduct a few key concepts. Firstly, a solution $\beta \in \mathbb{R}^p$ is said to be *K-sparse* if it has at most K nonzero elements. Secondly, the Spark of a matrix X, i.e. $spark(X)$, is the smallest number of columns of X that columns form a linearly dependent family. In this sense, $spark(X)$ is an integer that belongs to the interval $[2, n+1]$.

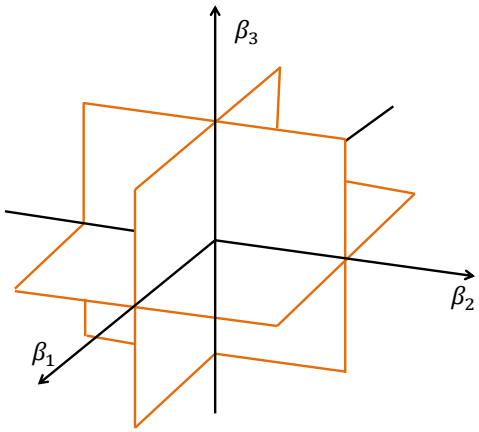

Figure 3.1: The set of 2-sparse solutions on a three-dimensional space.

Suppose there exist two distinct solutions to the system $y = X\beta$, in other words $\exists\ \beta_1 \neq \beta_0 : y = X\beta_1 = X\beta_0$. Since the difference $\beta_1 - \beta_0$ belongs to the kernel of X, i.e. $X(\beta_1 - \beta_0) = 0$, there is a subset of columns of X that are linearly dependent. Additionally, the number of nonzero elements in $(\beta_1 - \beta_0)$ indicates that there is an equal or less amount of linearly dependent columns in X. By the definition of $spark(X)$ it is straight-forward that $||\beta_1 - \beta_0||_0 \geq spark(X)$.

From the triangle inequality $||\beta_1 - \beta_0|| \leq ||\beta_1||_0 + ||\beta_0||_0$ and consequently,

$$||\beta_1||_0 + ||\beta_0||_0 \geq spark(X) \tag{3-2}$$

Theorem 3.1. If the under-determined system $y = X\beta$ admits a solution β_0 that obeys $||\beta_0||_0 < \frac{spark(X)}{2}$, then β_0 is the sparsest unique solution for (P_0).

Proof. If the solution β_0 obeys $||\beta_0||_0 < \frac{spark(X)}{2}$, any other solution β_1 to the system must obey $||\beta_1||_0 > \frac{spark(X)}{2}$ in order to preserve the inequality 3-2. In this sense, β_0 would be the sparsest solution for $y = X\beta$. Summarizing, any K-sparse signal can be uniquely recover by (P_0) if $K < \frac{spark(X)}{2}$. □

Since checking an inequality based on the spark of the design matrix is computationally exhaustive , it requires $O(2^p)$ computations, it is more reasonable to check for computable properties of X. In this regard, the most widely used metric is coherence. The concept is the same as that applied in physics to describe wave interferences. The coherence μ of X is defined as the maximal absolute inner product between any pair of columns of X. Another common way of formalizing the coherence is by means of the Gram matrix. Where the Gram matrix G can be obtained by $X^\top X$.

$$\mu(X) = \max_{i \neq j} |\langle x_i, x_j \rangle| = \max_{i \neq j} |\langle G_{i,j} \rangle| \tag{3-3}$$

A matrix is said to be incoherent if μ is small, i.e. if the largest off-diagonal element of the Gram matrix is small. The coherence of any matrix $n \times p$ is in the range $\left[\left(\frac{p-n}{n(p-1)}\right)^{1/2}, 1\right]$. The lower bound is known as *Welch Bound* on the telecommunication literature [38]. When the system is overdetermined, $n \ll p$, the lower bound tends asymptotically to $1/\sqrt{n}$.

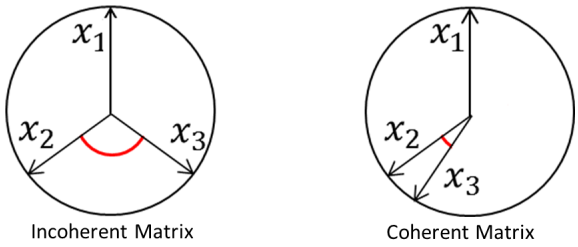

Figure 3.2: Angle between three columns of X defined using an inner product.

Considering that the columns of X are normalized to unit ℓ_2-norm, coherence is equivalent to the largest correlation between the variables that composes X. Another possible interpretation, Fig.[3.2], is that coherence is the cosine of the smallest angle between two columns of X. Thus, the presence of a coherent design matrix is known as multicollinearity on the field of statistics.

The ideal setting is to have the coherence $\mu(X)$ as close to $1/\sqrt{n}$ as possible. Donoho has shown in [27] that an upper bound can be stated using the coherence of X. This bound is going to be constructed by the following corollaries and theorems.

Corollary 3.1. *Every diagonally dominant matrix is nonsingular.*

Corollary 3.2. *Consider an arbitrary square matrix A and its corresponding Gram matrix $G = A^\top A$. If the Gram matrix G is nonsingular then all columns of A are linearly independent.*

Theorem 3.1. *Every matrix X obeys the inequality $spark(X) \geq 1 + \frac{1}{\mu(X)}$.*

Proof. Take an arbitrary set of m columns of X and denote by $G(m)$ the sub-Gram matrix correspondent to these columns. $G(m)$ is sad to be diagonally dominant if $\sum_{i \neq j} G_{i,j}(m) < G_{i,i}(m)\ \forall i = 1, \ldots, p$. Since $G_{i,i} = 1\ \forall\, i = 1, \ldots, p$ then $G(m)$ needs to obey $\sum_{i \neq j} G_{i,j}(m) < 1$.

By definition, the elements of the Gram matrix are limited by the coherence. In this sense, one can state that $\sum_{i \neq j} G_{i,j}(m) < (m-1)\mu(X)$. Thus, if $m < 1 + \frac{1}{\mu(X)}$ every sub-Gram matrix, made by m columns of X, will be diagonally dominant.

According to Corollary 3.1. if $G(m)$ is diagonally dominant then it is nonsingular. Additionally, by Corollary 3.2., if $G(m)$ is nonsingular the corresponding m columns of X are linearly independent.

In this sense, by the definition of spark one can assert that:

$$spark(X) \geq m + 1 > 1 + \frac{1}{\mu(X)}.$$

□

Theorem 3.2. *If the under-determined system $y = X\beta$ admits a solution β_0 that obeys $||\beta_0||_0 < \frac{1+1/\mu(X)}{2}$, then β_0 is the sparsest unique solution for (P_0).*

Proof. By the previous theorem $spark(X) \geq 1 + \frac{1}{\mu(X)}$ and $\frac{spark(X)}{2} \geq \frac{1}{2}(1 + \frac{1}{\mu(X)})$. Consequently if $||\beta_0||_0 < \frac{1+1/\mu(X)}{2}$ then $||\beta_0||_0 < \frac{spark(X)}{2}$ also holds true. From theorem 3.1 β_0 will be the sparsest unique solution for (P_0). □

There are some instances in which (P_0) can be solved by integer programming methods such as cutting planes or branch and bound [39] [40]. However, in general (P_0) is combinatorial and belongs to the class of NP-hard problems [41] [42]. This statement can be verified by reducing the *Exact cover by 3-sets problem* to (P_0), the proof of reducibility can be found at [43]. Since the Exact cover by 3-sets problem is a decision adaptation of one of Karp's 21 NP-complete problems [44], i.e. is a well-known classic NP-complete problem, then it implies that (P_0) is NP-hard. In other words, solving this problem requires searching for the best subset of regressors on a exponential number of potentials subsets. For this reason, this problem can not be solved in polynomial time.

3.2 Convex Relaxation

The computational intractability of the problem suggests a convex relaxation. In this sense, the relaxation must have a polyhedron as feasible region which contain all feasible points of the original problem. It can be seen graphically that replacing the ℓ_0 norm by the ℓ_1 norm is the convex hull of all feasible points of (P_0) as in Fig.[3.3]. As a result, the following problem (P_1) is the tightest convex relaxation and is contained in any other convex relaxation.

$$\begin{aligned} (P_1) : \text{minimize } & ||\beta||_1 \\ \text{subject to:} & \\ X\beta = y & \end{aligned} \tag{3-4}$$

In order to highlight the differences between norms, let's consider the bivariate case where $p = 2$. Assuming that the columns of X have unit ℓ_2 norm, i.e., all explanatory variables have standard deviation equal to one. The

following figure compares the feasible region $||\beta||_k \leq s$ for $k = 0, 0.5, 1, 2$ and ∞, where $s \in \mathbb{R}^+$.

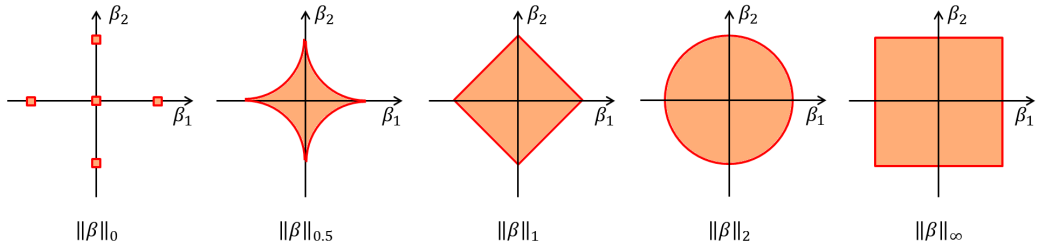

Figure 3.3: Unit spheres in $\mathbb{R}^2$ as feasible regions.

Since the ℓ_1 norm is a convex function in β, the problem can now be solved by convex programming techniques such as interior points or the simplex method. In doing so, any instance of this problem can be solved in polynomial time in the size of β. Additionally, modern off-the-shelf optimization solvers are prepared to deal with significant large instances of convex optimization problems.

The formulation of (P_1) can be translated into a Linear Programming problem [39] by expressing each coefficient as the sum of the positive and negative value it might assume $\beta_j = \beta_j^+ - \beta_j^-$ and replacing the absolute value by $\beta_j^+ + \beta_j^-$. This formulation (3-5) can be easily implemented on any optimization software.

$$\text{minimize} \sum_{j=1}^{p}(\beta_j^+ + \beta_j^-) \tag{3-5}$$

subject to:

$$\sum_{j=1}^{p} x_{ij}(\beta_j^+ - \beta_j^-) = y_i \quad \forall\, i = 1, \cdots, n$$

$$\beta_j^+, \beta_j^- \geq 0 \quad \forall\, j = 1, \cdots, p$$

As a result of these computational advantages, the ℓ_1 approximation has been used on several applications. In the field of statistics it has empowered the classic linear regression to deal with a high-dimensional framework. Further, it has been used to improve compressing techniques for several medias, such as images [45], audio [46] and video [47]. Finally, classification problems such as face recognition [48]. Specially the recent field of Compressed Sensing [26] [49], which for instance allowed a significant reduction on MRI (Magnetic Resonance Imaging) scan time.

All these range of applications and the tractability of ℓ_1 regularization techniques justify the term "ℓ_1-magic" used by Candès and Romberg [50]. The

following subsections investigate the signal recovery in a noise-free system and for a system in which measures are corrupted with some noise. Additionally, some conditions for the equivalence between (P_1) and (P_0) are given.

Since the original problem (P_0) is computationally intractable, our main focus is to describe conditions that guarantee that the convex relaxation (P_1) is going to recover the same solution as (P_0).

It can be easily seen that under a high level of coherence the formulation (P_1) will not necessarily recover the correct signal β. Consider that the signal is 2-sparse and the design matrix has the worst case of coherence, where two columns of X are completely correlated $\langle x_1, x_2 \rangle = 1$. In this case, the solution of (P_1) might be $(\beta_1,\ 0,\ \cdots,\ 0)^\top$, $(0,\ \beta_2,\ 0,\ \cdots,\ 0)^\top$ or any linear combination of those two vectors.

Theorem 3.3. *If the under-determined system $y = X\beta$ admits a solution that obeys*

$$||\beta||_0 < \frac{1 + 1/\mu(X)}{2} \tag{3-6}$$

, then this is the sparsest unique solution for both (P_0) and (P_1). Ensuring the equivalence between (P_0) and (P_1).

Proof. Let β_0 be the solution for (P_0) and $\mathcal{C}$ the set of all potential solutions for (P_1) that are different from β^*.

$$\mathcal{C} = \{\gamma : \gamma \neq \beta^*,\ ||\gamma||_1 \leq ||\beta^*||_1,\ ||\gamma||_0 > ||\beta^*||_0,\ X(\gamma - \beta^*) = 0\}.$$

In other words the set $\mathcal{C}$ contains solutions that (P_1) might prefer over the sparsest feasible solution β^*. By the Theorem 3.2. if $||\beta^*||_0 < \frac{1+1/\mu(X)}{2}$, β^* is the sparsest solution and the condition $||\gamma||_0 > ||\beta^*||_0$ is redundant. Letting $e = \gamma - \beta^*$ one can redefine $\mathcal{C}$ as:

$$\mathcal{C}_s = \{e : e \neq 0,\ ||e + \beta^*||_1 - ||\beta^*||_1 \leq 0,\ Xe = 0\}.$$

The main idea of the proof is to show that $\mathcal{C}_s$ is empty and therefore there is no alternative solution that (P_1) can pick. In fact, we are going to relax the conditions of $\mathcal{C}_s$ and show that even this larger set, that includes $\mathcal{C}_s$, is empty.

Without loss of generality, assume that the K nonzero elements of β^* are in the first entries of the vector.

$$||e + \beta^*||_1 - ||\beta^*||_1 = \sum_{j=1}^{K}(|e_j + \beta_j^*| - |\beta_j^*|) + \sum_{j>K}|e_j| \leq 0.$$

By the inequality $|e_j + \beta_j^*| - |\beta_j^*| \geq -|e_j| \, \forall j = 1, \ldots, p.$ one can relax the second condition by requiring only that $-\sum_{j=1}^{K} |e_j| + \sum_{j>K} |e_j| \leq 0$. Adding and subtracting $\sum_{j=1}^{K} |e_j|$ leads to $-2\sum_{j=1}^{K} |e_j| + (\sum_{j>K} |e_j| + \sum_{j=1}^{K} |e_j|) \leq 0$. Thus, we have the larger set $\mathcal{C}_s^1 \supseteq \mathcal{C}_s$ defined as:

$$\mathcal{C}_s^1 = \left\{ e : e \neq 0, \; \sum_{j=1}^{p} |e_j| - 2\sum_{j=1}^{K} |e_j| \leq 0, \; Xe = 0 \right\}.$$

Lets also relax the third condition $Xe = 0$, or equivalently $X^\top Xe = 0$. By adding and subtracting e we have $X^\top Xe - e = -e$. Taking the element-wise absolute value $|e| = |(X^\top X - I)e| \leq |(X^\top X - I)||e|$. Since the entries of $X^\top X$ are limited by the coherence, $|e| \leq |(X^\top X - I)||e| \leq \mu(X)(\mathbb{1}_{(p\times p)} - I)|e|$ holds true. This last inequality suggests a new set $\mathcal{C}_s^2 \supseteq \mathcal{C}_s^1$ defined by:

$$\mathcal{C}_s^2 = \left\{ e : e \neq 0, \; \sum_{j=1}^{p} |e_j| - 2\sum_{j=1}^{K} |e_j| \leq 0, \; |e| \leq \frac{\mu(X)}{1+\mu(X)} \mathbb{1}_{(p\times p)} ||e||_1 \right\}.$$

We can limit our analysis to values of e for which $||e||_1 = 1$. This restriction is not critic, since for any element $e \in \mathcal{C}_s^2$ all αe also belongs to $\mathcal{C}_s^2$. In this sense, we have the bounded set:

$$\mathcal{C}_l = \left\{ e : ||e||_1 = 1, \; 1 - 2\sum_{j=1}^{K} |e_j| \leq 0, \; |e| \leq \frac{\mu(X)}{1+\mu(X)} \mathbb{1}_{(p\times 1)} \right\}.$$

In order to obey the condition $1 - 2\sum_{j=1}^{K} |e_j| \leq 0$ the vector e needs to concentrate most of its energy on the first K entries. Using the maximum value allowed for each entry of e,

$$1 - 2K\frac{\mu(X)}{1+\mu(X)} \leq 0 \Rightarrow K \geq \frac{1 + 1/\mu(X)}{2}.$$

Contradicting the Theorem hypothesis. As a consequence, if $K \leq \frac{1+1/\mu(X)}{2}$, the original set $\mathcal{C}$ is empty and therefore there is no solution for (P_1) alternative to the solution of (P_0).

□

The inequality (3-6) allows a straightforward interpretation of the hardness of the incoherence hypothesis. As Fig.[3.4] indicates, even a small level of coherence restricts (P_1) to uniquely recover only extremely sparse signals. For instance, consider the case where the matrix X has coherence equal to 0.15, which is pretty much a soft assumption, then according to 3-6 (P_1) will be able to uniquely recover only 3-sparse signals.

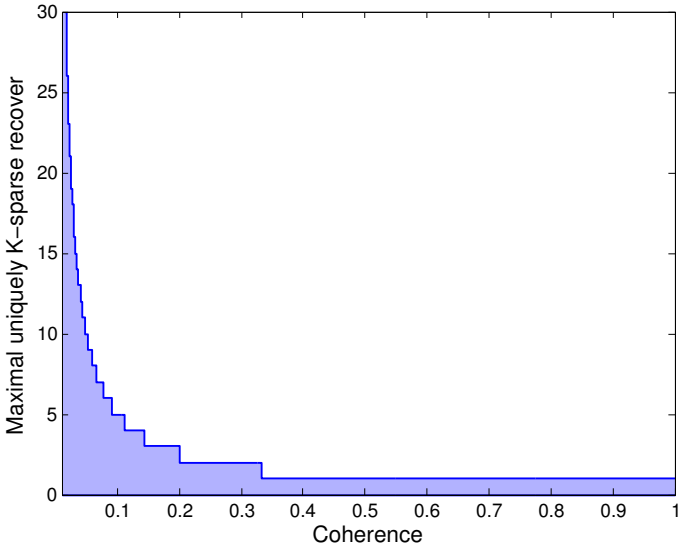

Figure 3.4: Maximal unique sparse recovery by (P_1) versus the coherence of X.

Observing Fig.[3.4] it is clear that the inequality (3-6) is very restrictive. At first sight, if one choses an arbitrary design matrix X it seems that (P_1) will hardly be able to uniquely recover the signal β unless the signal is 1-sparse, 2-sparse or at most 3-sparse. This statement often holds true for several applications. However, the analysis that establish the inequality 3-6 was based on the worst case. In practice, it is possible to recover the correct signal with a high probability even if (3-6) do not hold. Results in probabilistic analysis are explored using random matrices in [35] [51].

3.3 Adding Some Noise

In most practical applications it is reasonable to assume that the measurements are contaminated with some level of noise. This noise may arise for different reasons. In engineering applications, the sensor may take noisy measures or the output may come from a channel affected by some noise. In statistics modeling, the noise is often added to model in order to represent inputs to the system that are not known. In this work we will assume that the system is corrupted by an additive noise $\varepsilon \in \mathbb{R}^n$, leading to the system

$$y = X\beta + \varepsilon. \tag{3-7}$$

Additionally, we assume the hypothesis that each element ε_i of the vector ε is normally distributed. This hypothesis is motivated by the *Central Limit Theorem* [52]. The main idea is that the noise is actually a sum of a large number of independent random factors. In this regard, even if each random

factor is not normally distributed, accordingly to the Central Limit Theorem the sum is going to be normally distributed. Thus, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_p)^\top$ where $\varepsilon_i \sim \mathcal{N}(0, \Sigma)$ and iid (i.e. independent identically distributed) $\forall i = 1, \cdots, p$.

In order to minimize the noise variance, consider a loss-function $\mathcal{L}\colon \mathbb{R}^n \to \mathbb{R}$ that measures the disparity between the observations y and the prediction $\hat{y} = X\hat{\beta}$. The loss function is often represented by the residual sum of squares but it may assume other designs. Another common loss-function is the absolute sum of residuals used on quantile regression [53] and some applications on digital communications systems, such as channel decoding [35].

The requirements for convexity are that the loss-function must be convex in β and the feasible set must be a convex set [54]. In this sense, in a high-dimensional set-up, the estimator may be obtained by the constrained problem 3.3.

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \;\; \mathcal{L}(\beta, Y, X)$$

subject to:

$$\left(\sum_{j=1}^{p} |\beta_j|^q \right)^{1/q} \leq s$$

For $q = 1$, this problem can be viewed as a slight modification of (P_1), where the equality constraint $y = X\beta$ is relaxed and a tolerance of the type $||y - X\beta||_2 < \gamma$ is adopted.

There is a geometric motivation for claiming that $q = 1$ in (3.3) is appropriate for recovering a sparse signal. Using the ℓ_1-norm, instead of other norms, produces a spiky polyhedron as feasible region. The sharpness of the polyhedron along with the fact that the objective function is convex, makes it more likely that a sparse solution may be found. In this sense, it is possible to say that the ℓ_1-norm constraint induces a sparse solution.

This phenomena can be pictured in the $\mathbb{R}^2$ contour plots for three possible cases. The ellipses are the contour line for the loss-function $\mathcal{L}$ and the orange areas are the feasible regions for the ℓ_1-norm and the ℓ_2-norm respectively. In the first case, Fig. 3.5, the real signal β_0 is sparse with $\beta_2 = 0$. The formulation 3.3 with $q = 1$ can identify the sparse solution, whereas with $q = 2$ the solution has only nonzero elements.

In the second case, despite the real signal β_0 being non-sparse, the solution for $q = 1$ is sparse. This happens because the absolute value of β_2 is small enough to be shrunk to zero, Fig. 3.6. At first sight, improperly recovering a sparse signal may be considered as a drawback. However, in several applications, like image compression, it is desirable to obtain a solution with an incomplete amount of linear measurements as long as most of the information

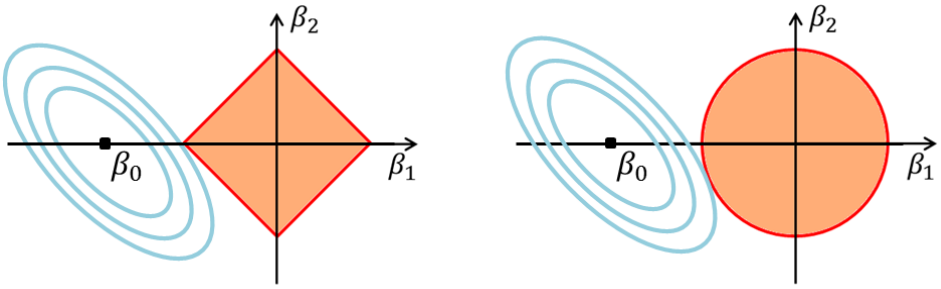

Figure 3.5: Case 1: Sparse solution for sparse signal

is preserved.

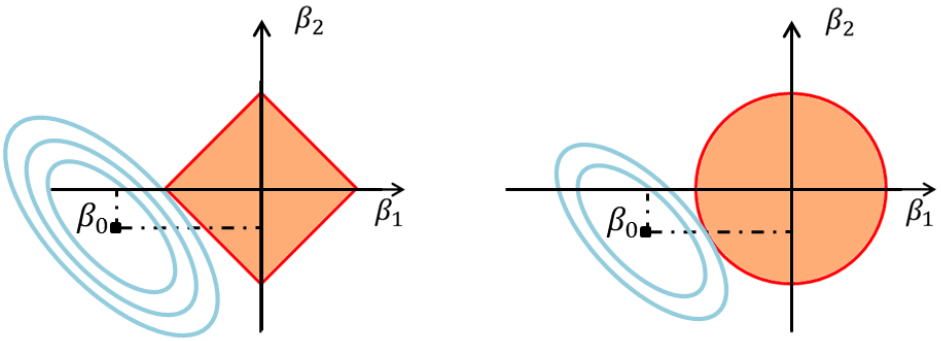

Figure 3.6: Case 2: Sparse solution for non-sparse signal

Finally, in the third case, Fig. 3.7, both the ℓ_1-norm and the ℓ_2-norm formulations find a non-sparse solution. Since neither the absolute value of β_1 and β_2 are sufficiently small, the coefficients are not shrunk to zero. Accordingly, solving (3.3) is equivalent to perform a hard thresholding on the signal β. Being sufficiently small depends on the loss-function and the value of s, which establishes how tight the constraint is. The greater the value of s is more likely it will be to obtain a sparse solution to (3.3).

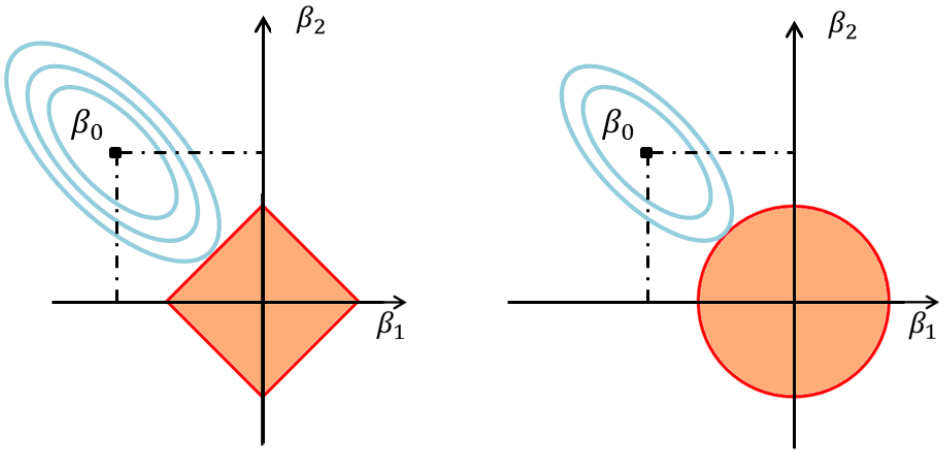

Figure 3.7: Case 3: Non-sparse solution for non-sparse signal

4
LASSO

Since Tibishirani [22] introduced the LASSO (*least absolute shrinkage and selection operator*) in the mid-nineties, it has become an extremely popular technique to tackle high-dimensional data. Ever since, several researches have proposed new results, algorithms and extensions for the LASSO. For instance the *Adaptative LASSO* [55] and the *Group LASSO* [56].

The computational advantage and the methods simplicity has made the LASSO an attractive method for modeling large data-sets. Several fields ranging from biostatistics [57] to social science has been using the LASSO to make sense out of huge amounts of data. An interesting retrospective can be found in [58], allowing an understanding of the evolution of the method over the years.

Two years after the original LASSO, Chen, Donoho and Saunders [59] also proposed a similar method known as *Basis Pursuit DeNoising* (BPDN) in the signal processing literature. This paper has inspired several engineering applications like audio denoising [60], image denoising [61] and signal compression.

$$\hat{\beta} \in \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \tfrac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1 \right\}$$

The LASSO 4 is comprised of a loss function, which is the classic sum of squares of residuals, and a penalization on the ℓ_1 norm of the regressors. In other words, the method is basically a plain multivariate linear regression with an ℓ_1-norm penalization term, which induces variable selection. As explained on the previous chapter, due to the ℓ_1 regularization, (4) efficiently does variable selection and estimation simultaneously.

For practical purposes, it is assumed that the predictors are standardized, i.e. $\sum_{i=1}^{n} x_{ij}/n = 0$ and $\sum_{i=1}^{n} x_{ij}^2/n = 1 \; \forall j = 1, \cdots, p$. This transformation allows the coefficients β_j to be comparable, since the columns of X are at the same scale. Additionally, to avoid problems with the standardization of the intercept, it is assumed that the response y is centered at zero . Since most of the algorithms and proofs are designed based on this assumptions, for now on they are going to be considered as the standard framework.

4.1 Regularization Parameter

The parameter λ is a tunning that controls the amount of regularization, i.e. it controls the trade-off between the errors and sparsity. The bigger the λ is, the more elements of β are set to zero. With $\lambda = 0$ the solution for 4 is the same as the *Ordinary Leasts Squares* solution. Figure 4.1 gives an example of a 4-sparse signal in order to highlight the difference between the solutions for the LASSO as λ varies. As the regularization decreases the coefficients are less shrinkage. On this particular case, with $\lambda = 223.1$ the signal is recovered with a little bias. Note that for $\lambda = 17.7$, the solution approaches the OLS estimate and consequently is not sparse.

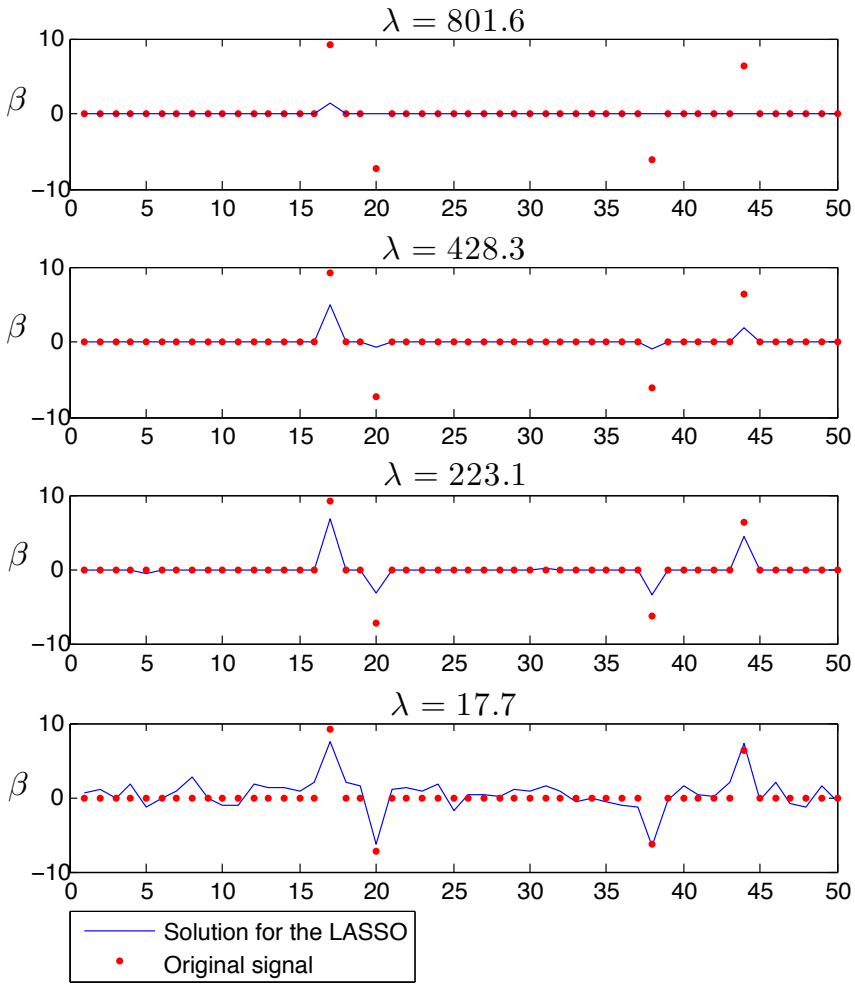

Figure 4.1: Distinct solutions for different regularization parameters.

Additionally, it is possible to obtain the minimum value of λ for which all coefficients of β are set to zero. Firstly, taking the Karush-Kuhn-Tucker (KKT) optimality conditions [62] for the LASSO, we have the following first-

order condition:

$$X^\top(y - X\hat{\beta}) + \lambda\delta = \mathbf{0}, \tag{4-1}$$

where $\mathbf{0}$ is a p-dimensional vector of zeros and $\delta \in \mathbb{R}^p$ is the subgradient of the ℓ_1-norm evaluated at $\hat{\beta}$. In order to satisfy 4-1 the subgradient needs to obey,

$$\delta_i \in \begin{cases} \{+1\}, & \text{if } \hat{\beta}_i < 0 \\ \{-1\}, & \text{if } \hat{\beta}_i > 0 \\ [-1, 1], & \text{if } \hat{\beta}_i = 0 \end{cases} \quad \forall\, i = 1, \cdots, p. \tag{4-2}$$

Since we are interested in the case where $\hat{\beta}_i = 0\, \forall\, i = 1, \cdots, p$, then the subgradient is restricted to $X^\top y = -\lambda\delta$ where $\delta_i \in [-1, 1]\, \forall\, i = 1, \cdots, p$. To guarantee that λ will be able to set to zero all coefficients it is necessary and sufficient that:

$$\lambda > \lambda_{max} = ||X^\top y||_\infty, \tag{4-3}$$

where the infinity norm returns the maximum element of the vector $X^\top y$.

To come up with a proper estimation of β one has to choose the "best" λ over a range of possibilities. The only prior information is that $\lambda \in [0, \lambda_{max}]$. The usual strategy to establish λ is to solve (4) over a grid of values in the aforementioned interval. However, since $\lambda = \lambda_{max}$ does not bring any information, we only consider values that are smaller than $0.99 \times \lambda_{max}$. Further, since in a high-dimensional framework $\lambda = 0$ will lead to an unstable estimator, the smallest value of λ conisdered is $0.01 \times \lambda_{max}$.

Accordingly, we are going to evaluate the solutions over a finite number of values. For convenience, the solution for the k^{th} element on the grid will be denoted by $\hat{\beta}(\lambda_k)$. For most part of the algorithms, it is better to do the search backwards, since the previous solution $\hat{\beta}(\lambda_{k-1})$ can be used as a *warm start*. Additionally, the grid is usually not equidistant but logarithmic. In this fashion, a greater number of smaller values of lambda are evaluated. This strategy makes sense as even with a smaller penalty, variable selection will be achieved and the solution will be less biased Fig.4.1.

Since we are working with an over-complete design matrix X, the task of modeling the response y is restricted by choosing the amount of regularization. There is not a universal criteria for establishing the correct value of λ. The chosen criteria depends primarily on the goal of the model. If one is building a model to perform predictions the usual strategy is to use a *Cross-Validation* method, like *K-fold* or *leave-one-out* [63] [64]. On the other hand, if the main purpose of the model is recovering the original signal β or if the objective is

drawing conclusions on how X explains the response y, then BIC (*Bayesian information criterion*) is thought to be more suitable [65]. This belief is due to the asymptotic consistency of BIC as a selection criteria [66][30].

Since the case studies presented in this work are not focused on prediction, BIC will be the standard measure to select λ. We can index the residuals as follows $\varepsilon(\lambda_k) = y - X\hat{\beta}(\lambda_k)$. Under the assumption that residuals follow a Gaussian distribution and are independent and identically distributed. Then the Bayesian criteria can be expressed as:

$$\mathrm{BIC}(\lambda_k) = n \ln \hat{\sigma}^2_{\varepsilon(\lambda_k)} + ||\hat{\beta}(\lambda_k)||_0 \ln n, \tag{4-4}$$

where $\hat{\sigma}^2_{\varepsilon(\lambda_k)}$ is the residual variance. In Fig.[4.2] a numerical experiment was made to give an intuition on the relationship between BIC and λ. Since BIC 4-4 is penalized by the number of nonzero coefficients, the solutions where λ is to close to zero are discarded. On the other hand, BIC seeks a model that can properly filter the residuals in a fashion that it resembles white noise. By choosing the λ which minimizes BIC, a parsimonious model will be selected based on this trade-off.

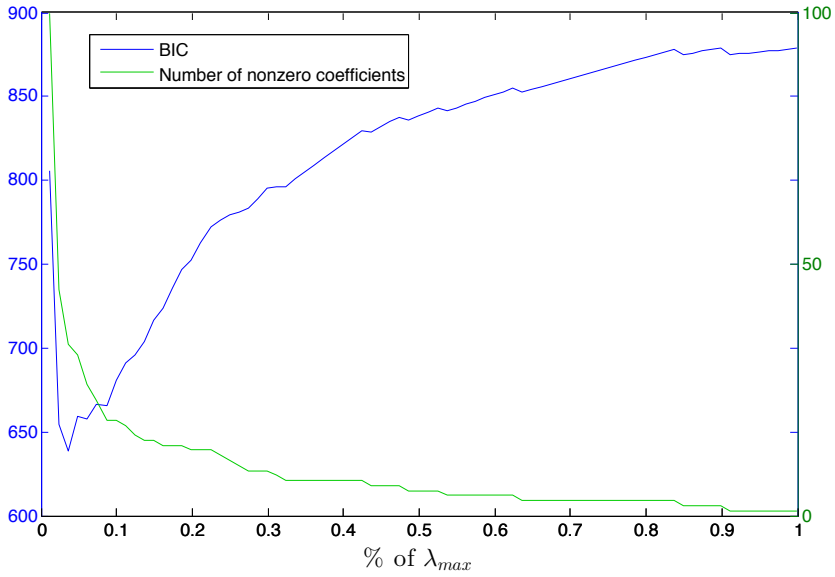

Figure 4.2: Distinct solutions for different regularization parameters.

4.2 Solution Methods and Algorithms

In this section the most well-known algorithms for solving the LASSO are exposed. Understanding such algorithms helps building an intuitive view

of the LASSO-path and how the shrinkage works. The methods are presented in the chronological order they were proposed on the literature.

Least-Angle Regression and *Pathwise Coordinate Descent* are available in the *glmnet* [67] software for Matlab© and R. Both methods are also available for Python on scikit-learn [68].

4.2.1 Quadratic Programming

Initialy, when the LASSO was proposed, the main idea was to use commercial optimization solvers. In this sense, the problem can be reformulated as a quadratic program of the form:

$$\begin{aligned} \text{minimize} \quad & c^\top x + \tfrac{1}{2} x^\top Q x \\ \text{subject to:} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{4-5}$$

Since the matrix Q equals the identity, it is readily verified that the matrix is positive definite. As a consequence, the problem 4-5 is convex and any feasible solution is a unique global minimizer [54].

There are two main methods for quadratic programming. The *Primal-dual path following algorithm*[39] and *Linear Complementary Programming*[62]. The first one, makes use of the *Interior Points Method* [69] and the second is basically a simplex method with a particular pivoting rule [70]. Both methods are based on solving the KKT optimality conditions. Since the problem is convex, the solution of KKT is the global optimal solution [62].

Since most of the commercial solvers work as black-boxes, solving the LASSO via quadratic programming was not considered attractive in the statistical community. One of the reasons for the popularization of the LASSO was the upcoming of new algorithms, which allowed an interpretation and provided an intuition about how the method works.

4.2.2 Least-Angle Regression

In 2009 Efron proposed a novel method for model selection known as *least angle regression* (LARS) [71], an accelerated version of *forward stepwise regression* [72]. Like forward stepwise regression, this method gradually selects explanatory variables to enter the model. However, LARS is less greedy since it does not fully enter the coefficients at once. In the same paper it is shown

that with a slight change to the original algorithm the LARS can compute the entire LASSO-path efficiently.

The main idea of the algorithm, is to first select the most correlated variable with the response and move its coefficient from zero towards its OLS estimator. If another variable is equally correlated with the residual, then the process is stopped and both coefficients are moved along towards their joint OLS estimator. This movement is again interrupted when a third variables has the same correlation with the residual as the previous two. Note that at each iteration the coefficients approach the OLS estimator, without necessarily achieving it. This process is repeated until the p variables enter the model.

The direction that the fitted model evolves during the algorithm has the interpretation of a *least angle* between the variables on the current active-set. Unlike from other algorithms there is no need to build a grid of λ. The continuos piecewise-linear path of the LARS algorithm is a exactly the LASSO-path [73]. Since variables may leave the active-set, the algorithm may take more then p steps to finish. The following pseudo-code describes the LARS modified for solving the LASSO.

Algorithm 1 LARS (lasso)

Initialize residual $r = y$ and the active-set $\mathcal{A} = \varnothing$
Set $\beta_1 = \cdots = \beta_p = 0$
Insert the most correlated variable X_j on $\mathcal{A}$
Compute $\delta = (X_{\mathcal{A}}^{\top} X_{\mathcal{A}})^{-1} X_{\mathcal{A}} r$
while $(\#\mathcal{A} \neq p)$ **do**
  $r = y - X_{\mathcal{A}} \beta_{\mathcal{A}}$
  Move $\beta_{\mathcal{A}}$ on the direction δ
  if (Any X_i has as much correlation with r as any variable of $\mathcal{A}$) **then**
    Add X_i to $\mathcal{A}$
  end if
  if (Any variable of the set $\mathcal{A}$ has a null coefficient) **then**
    Remove this variable from $\mathcal{A}$
  end if
  Compute $\delta = (X_{\mathcal{A}}^{\top} X_{\mathcal{A}})^{-1} X_{\mathcal{A}} r$
end while

This method has been commonly used for solving the LASSO given the idea behind it. The proof of correctness of the LARS algorithm can be found at [73].

4.2.3 Pathwise Coordinate Descent

The *Pathwise Coordinate Descent* has been previously suggested by [74], [75] and [76] and subsequently improved by [77] and [78]. This algorithm applies a general nonlinear programming technique known as *coordinate descent* to the LASSO problem. Due to the simplicity and efficiency of this method, Coordinate Desecent based algorithms are popular tools for large optimization problems.

The main idea of *Coordinate Descent* is to optimize a multivariate function by minimizing the objective function with respect to a single coordinate direction instead of the multivariate descent direction given by gradient vector [79][62]. This strategy might look like an heuristicat first sight. However, since the objective function is composed of a differentiable and convex component plus a non-differentiable but separable component then optimality is guaranteed. This result can be found in the nonlinear programming literature, particularly on the work of Paul Tseng [80] [81].

The LASSO loss-function is differentiable and convex and the ℓ_1-penalization is non-differentiable but it can be separated into the sum of the absolute value of each coordinate. For simplificity let us denote $f(\beta_1, \beta_2, \cdots, \beta_p) = \frac{1}{2}||Y - X\beta||_2^2 + \lambda||\beta||_1$. For a given value of λ, the i^{th} cycle of the algorithm can be described as:

$$
\begin{aligned}
&\beta_1^{(i)} \in \text{argminlimits}_{\beta_1} f(\beta_1, \beta_2^{i-1}, \beta_3^{i-1}, \cdots, \beta_p^{i-1}) \\
&\beta_2^{(i)} \in \text{argminlimits}_{\beta_2} f(\beta_1^{i}, \beta_2, \beta_3^{i-1}, \cdots, \beta_p^{i-1}) \\
&\vdots \\
&\beta_p^{(i)} \in \text{argminlimits}_{\beta_p} f(\beta_1^{i}, \beta_2^{i}, \beta_3^{i}, \cdots, \beta_p)
\end{aligned}
$$

This cyclic procedure is repeated until convergence is reached. As we shall see below, minimization trough one coordinate can be expressed as a shrinkage operator of the OLS estimate.

Initially, consider the LASSO for a simple linear regression $y_i = x_i\beta + \varepsilon_i \ \forall i = 1, \cdots, n$. The loss-function is defined as $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the prediction for the i^{th} observation obtained by $x_i\beta$. It is possible to recast the loss-function as $\sum_{i=1}^{n}(x_i\beta - x_x\hat{\beta})^2 = (\beta - \hat{\beta})^2 \sum_{i=1}^{n} x_i^2$. Since X is normalized, as descibed in ..., obtaining the estimator $\hat{\beta}^{lasso}$ is equivalent to solving the

following problem:

$$\min_{\beta} \tfrac{1}{2}(\beta - \hat{\beta}^{ols})^2 + \lambda|\beta| \tag{4-6}$$

where $\hat{\beta}^{ols}$ is the ordinary least squares estimator given by $\hat{\beta}^{ols} = \frac{\text{cov}[X,Y]}{V[X]} = \sum_{i=1}^{n} x_i y_i$.

In order to guarantee the optimality of 4-6, it necessary to assure the first-order condition from subdifferential calculus. In this sense, one may obtain the subsequent expression for the sub-gradient:

$$\partial\left(\tfrac{1}{2}(\beta - \hat{\beta}^{ols})^2 + \lambda|\beta|\right) = \beta - \hat{\beta}^{ols} + v = 0 \tag{4-7}$$

exists a $v \in \mathbb{R}$ such as:

$$\hat{\beta}^{lasso} = \hat{\beta}^{ols} - v, \qquad v = \begin{cases} -\lambda, & \text{if } \hat{\beta}^{ols} > 0 \text{ and } \lambda < |\hat{\beta}^{ols}| \\ +\lambda, & \text{if } \hat{\beta}^{ols} < 0 \text{ and } \lambda < |\hat{\beta}^{ols}| \\ \hat{\beta}^{ols}, & \text{if } \lambda \geq |\hat{\beta}^{ols}| \end{cases} \tag{4-8}$$

From (4-8) it is clear that the LASSO estimator is a shrunken version of the OLS estimator. Thus, $\hat{\beta}^{lasso}$ can be obtained as a result of a soft-thresholding operator on $\hat{\beta}^{ols}$ as illustrated in Fig.[4.3]. The dotted gray line represents the OLS estimate and the blue line is the corresponding LASSO estimate. If $\lambda \geq |\hat{\beta}^{ols}|$ then $\hat{\beta}^{ols}$ is shrunk to zero. The LASSO estimator can be written more neatly as:

$$\hat{\beta}^{lasso} = sign(\hat{\beta}^{ols})(|\hat{\beta}^{ols}| - \lambda), \tag{4-9}$$

where $sign(.)$ is a function that returns the sign of the scalar $\hat{\beta}^{ols}$. In this manner, obtaining the OLS estimator and applying 4-9 is equivalent to solving 4-6.

To extrapolate this result for a multivariate regression, consider the multivariate loss-function:

$$\mathcal{L}(\beta) = \tfrac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{p} x_{ik}\beta_k\right)^2. \tag{4-10}$$

Suppose that the value of every element of the vector β is known, except β_j, one can fix $\beta_1 = \tilde{\beta}_1, \beta_2 = \tilde{\beta}_2, \cdots, \beta_p = \tilde{\beta}_p$. This procedure leads to the following loss-function associated with the index j.

$$\mathcal{L}(\tilde{\beta}, \beta_j) = \tfrac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j\right)^2 \tag{4-11}$$

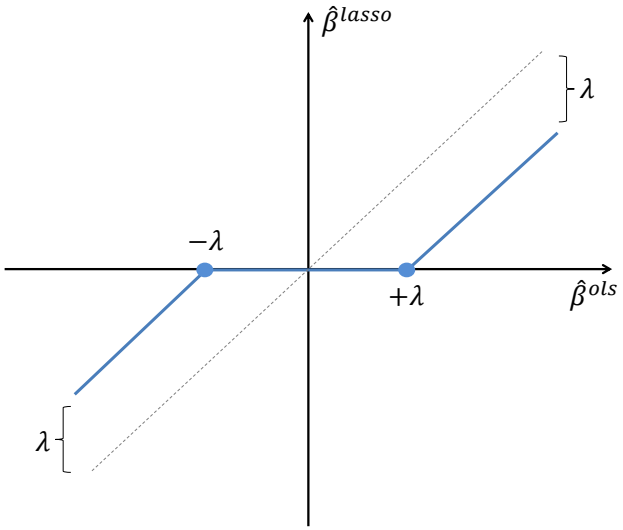

Figure 4.3: Soft-thresholding operator

The OLS estimator for β_j can be obtained by taking the derivative of 4-11 and equating to zero.

$$\frac{d\mathcal{L}(\tilde{\beta}, \beta_j)}{d\beta_j} = \sum_{i=1}^{n} \left(y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j \right) (-x_{ij}) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k \right) = \sum_{i=1}^{n} x_{ij}^2 \beta_j$$

Since $\sum_{i=1}^{n} x_{ij}^2 = 1$,

$$\hat{\beta}_j^{ols} = \sum_{i=1}^{n} x_{ij} \left(y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k \right) \qquad \forall j = 1, \cdots, p.$$

Then the soft-thresholding operator 4-9 can be applied to $\hat{\beta}_j^{ols}$ and then an iterative algorithm could be establish. For each penalization λ_i, the vector $\hat{\beta}^{lasso}(\lambda_i)$ is obtained by iteratively cycling through the p soft-thresholding operators applied to the OLS estimators given by $\hat{\beta}_j^{ols}$.

The main idea of the *Pathwise Coordinate Descent* algorithm is to successively apply coordinate descent for a different λ in the LASSO path. In order to speedup the algorithm, Friedman [77] proposes an efficient *warm starting* technique. The estimates are obtained for a decreasing grid $\lambda_1 > \cdots > \lambda_k > \cdots > \lambda_r$. Usually, λ_1 is set to $0.99 \times (\lambda_{max}/n)$ (4-3). The initialization for the k^{th} problem will be the solution for obtained with the penalization λ_{k-1}. In Fig.[4.4], the green arrows represent coordinate descent steps that were initialized with a previous solution.

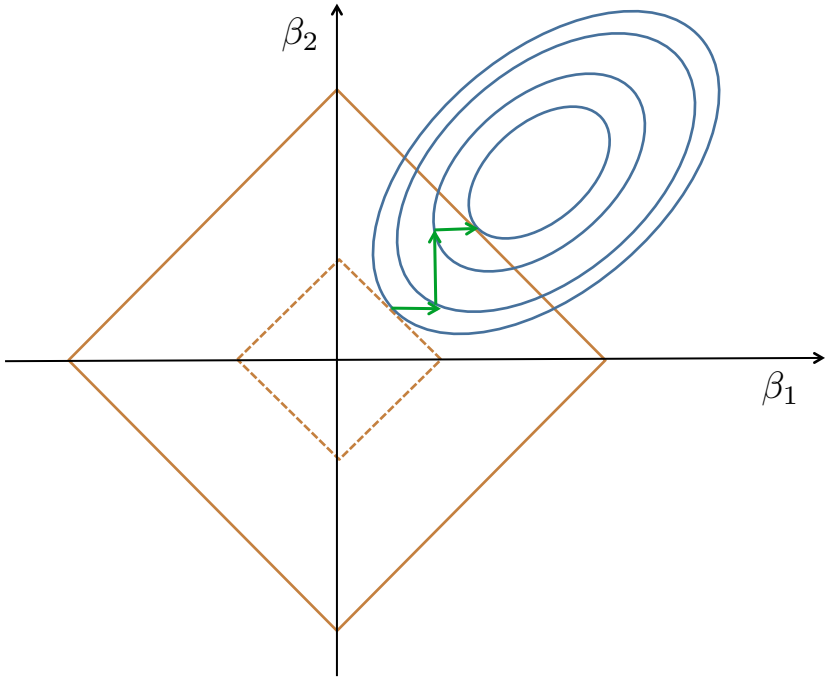

Figure 4.4: Warm starting for coordinate descent.

Considering that for several iterations it is highly probable that the coefficient that was set to zero will remain zero, it is less computationally demanding to redefine $\hat{\beta}_j^{ols}$ as the following. First, substitute $\sum_{k\neq j} x_{ik}\tilde{\beta}_k$ by $\hat{y}_i + x_{ij}\tilde{\beta}_j$. Where $\hat{y}_i$ denotes the most updated forecast for the i^{th} observation. Second, $y_i - \hat{y}_i$ can be replaced by the residual r_i. This modifications will lead to:

$$\hat{\beta}_j^{ols} = \sum_{i=1}^{n} x_{ij} r_i + \tilde{\beta}_j \quad \forall j = 1, \cdots, p. \tag{4-12}$$

This method is usually known as the *Naive* Coordinate Descent Algorithm. With this implementation, each iteration for an arbitrary β_j is done with $O(n)$ operations. A complete cycle over all coefficients can be done in $O(np)$.

Algorithm 2 *Naive* Coordinate Descent

Initialize $\tilde{\beta}_j = 0 \quad \forall\ j = 1, \cdots, p$
for all $\lambda_k \in \{\lambda_1, \cdots, \lambda_m\}$ **do**
  while not converge **do**
    for all $j \in \{1, \cdots, p\}$ **do**
      $r_i = y_i - \sum_{l=1}^{p} x_{il}\tilde{\beta}_j \quad \forall\ i = 1, \cdots, n$
      $\hat{\beta}_j^{ols} = \sum_{i=1}^{n} x_{ij} r_i + \tilde{\beta}_j$
      if $\lambda_k \geq |\hat{\beta}_j^{ols}|$ **then**
        $\tilde{\beta}_j = 0$
      else
        $\tilde{\beta}_j = sign(\hat{\beta}_j^{ols})(|\hat{\beta}_j^{ols}| - \lambda_k)$
      end if
    end for
  end while
  $\hat{\beta}_j^{lasso}(\lambda_k) = \tilde{\beta}_j \quad \forall\ j = 1, \cdots, p$
end for

It is possible to make this algorithm even more efficient by avoiding redundant computations. First of all, we introduce the set $\mathcal{A}$, called *active set*. The purpose of this set is to track when a coefficient $\tilde{\beta}_j$ is no longer zero for the first time. The term $\sum_{i=1}^{n} x_{ij} r_i$ computed for the OLS estimator can be replaced by $\sum_{i=1}^{n} x_{ij} y_i - \sum_{l \in \mathcal{A}} \left(\sum_{i=1}^{n} x_{il} x_{ij}\right) \tilde{\beta}_l$ which is the same as $\langle x_j, y \rangle - \sum_{l \in \mathcal{A}} \langle x_j, x_l \rangle \tilde{\beta}_l$. Consequently, a more effective strategy is to compute $\langle x_j, y \rangle \ \forall\, j = 1, \cdots, p$ in $O(np)$ at the beginning of the routine. Additionally, whenever an index j is added to the active set $\mathcal{A}$ it is necessary to obtain $\langle x_j, x_l \rangle \ \forall\, l = 1, \cdots, p$ in $O(np)$ as well. Convergence is obtained if the active set does not change after a complete cycle. This modifications leads to the *Covariance Updates* Coordinate Descent algorithm.

Algorithm 3 *Covariance Updates* Coordinate Descent

Initialize $\tilde{\beta}_j = 0 \quad \forall\ j = 1, \cdots, p$
Initialize active set $\mathcal{A} = \varnothing$
Compute and store $\langle x_j, y \rangle \quad \forall\ j = 1, \cdots, p$
for all $\lambda_k \in \{\lambda_1, \cdots, \lambda_m\}$ **do**
  while not converge **do**
    for all $j \in \{1, \cdots, p\}$ **do**
      $\hat{\beta}_j^{ols} = \frac{1}{n}\left[\langle x_j, y \rangle - \sum_{l \in \mathcal{A}} \langle x_j, x_l \rangle \tilde{\beta}_l\right] + \tilde{\beta}_j$
      if $\lambda_k \geq |\hat{\beta}_j^{ols}|$ **then**
        $\tilde{\beta}_j = 0$
      else
        $\tilde{\beta}_j = sign(\hat{\beta}_j^{ols})(|\hat{\beta}_j^{ols}| - \lambda_k)$
        if $j \notin \mathcal{A}$ **then**
          Compute and store $\langle x_j, x_l \rangle \quad \forall\ l = 1, \cdots, p$
          $\mathcal{A} = \mathcal{A} \cup \{j\}$
        end if
      end if
    end for
  end while
  $\hat{\beta}_j^{lasso}(\lambda_k) = \tilde{\beta}_j \quad \forall\ j = 1, \cdots, p$
end for

Figure [4.5] ilustrates an example of the *Covariance Updates* Coordinate Descent recovering a 15-sparse signal. The LASSO-path $\{\lambda_1, \cdots, \lambda_m\}$ on this example is a decreasing linear grid with 300 elements ranging from $0.99 \times \lambda_{max}$ to $0.01 \times \lambda_{max}$. As expected, as the regularization parameter decreases more variables enter the active set $\mathcal{A}$.

At the present moment, no theoretical global convergence rate has been establish for the pathwise coordinate descent methods. However, empirically it has been proven to be the fastest method for solving the LASSO, specially when the number of unknown parameters p is much greater than the number of observations n.

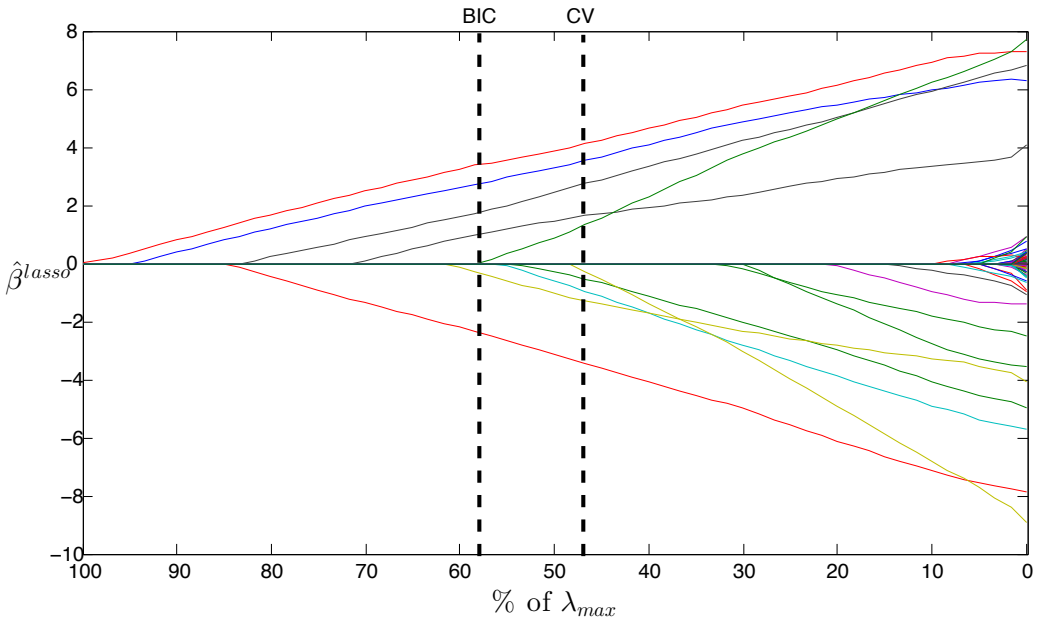

Figure 4.5: Solution path for a 15-sparse signal via Covariance Updates Coordinate Descent.

5
Case Study 1: Fiber Optics Failure Detection

In this chapter we start to evaluate the potential of the ℓ_1-regularization in solving real-world problems. As mentioned previously, the ℓ_1-norm has became a popular tool in the field of signal processing. In this work, we study one of those applications, more particularly in fiber-optic communication. Most of the experiments and techniques presented in this chapter are based on the work of [82], with additional improvements on the filter algorithm.

In a nutshell, optical fibers are an efficient media for data transmission. The optical communication is performed using an optical transmitter which modulates light - commonly originated from a Laser source - based on the data to be transmitted. The message is sent through the optical fiber and later demodulated by an optical receiver.

Due to the high effectiveness of optical fibers and their capacity for transmitting large amounts of data, they have been broadly used in place of other data transmission media such as copper cables. Several applications like cable TV and medical imaging are adopting this technology. More recently, home internet services are offering fast connectivity services based on fiber optics.

Despite their great success, optical fibers are made of glass implying mechanical fragility to the material which, under some conditions, may break. The ruptures usually result in a failure of data transmission which, in turn, may interrupt the link. Since optical fiber links frequently exceed the length of two kilometers, finding the location of the rupture with reasonable precision is essential.

5.1
Experiment description

One of the existing methods for detecting optical fiber breakpoints is the Photon Counting OTDR [83]. The main idea of the method is to stimulate the optical fiber with an optical pulse and count the number of photons that are backscattered, i.e. reflected to the direction from which they came, using a Single Photon Detector - a device capable of detecting single photons [ref].

In the presence of a rupture on the optical fiber, the number of backscattered photons after the break is reduced due to power coupling losses. The output of the method is a sequence representing the number of backscattered photons on each fiber position.

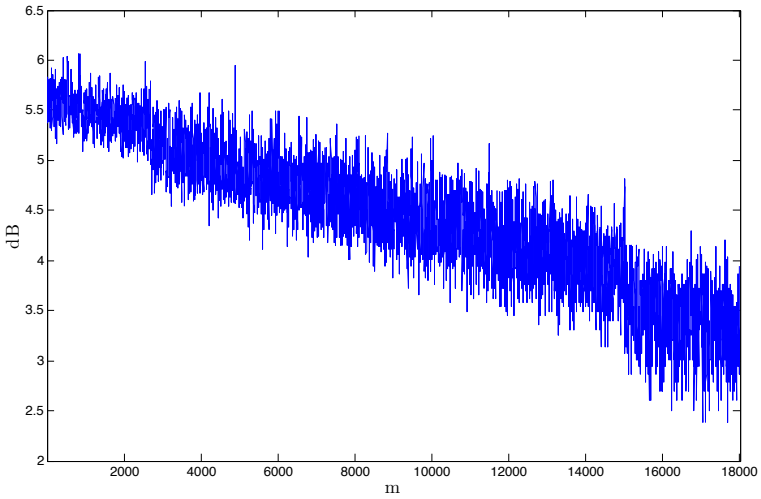

Figure 5.1: Output of the experiment.

As can be seen in Fig.5.1, the amount of counted photons has a linear decay due to an intrinsic power loss of the fiber. Additionally, two level shifts can be seen at approximately 2500 and 15000 meters.

In the following, we propose a methodology to detect the exact position of the breakpoints and obtain the relative power loss at the level shifts.

5.2 ℓ_1 Level-Slope Filter

Deeply inspired by the work of Boyd et al. [84], we develop a filter where the key ingredient is the ℓ_1-norm. The main idea is to fit to the measurements a piece-wise linear curve composed by a slope and occasional level shifts. The ℓ_1-regularization is in charge of selecting the location of level shifts in a parsimonious manner.

The sequence of measurements are denoted by $\{y_s\}_{s=1}^n$, where the index s corresponds to the location in meters on the fiber cable and n is the total of measurements. Consider also a slope coefficient $\alpha \in \mathbb{R}$ and a component of level shifts $w \in \mathbb{R}^n$. The ℓ_1 *Level-Slope Filter* can be defined as:

$$\min_{(\alpha,\{w_i\}_{i=1}^n)} \sum_{i=1}^{n}(y_i - \alpha i - w_i)^2 + \lambda\big(|\alpha| + \sum_{i=2}^{n}|w_i - w_{i-1}|\big). \qquad (5\text{-}1)$$

Analogously to the theory presented on the previous chapter, the parameter $\lambda \in \mathbb{R}_+$ controls the amount of regularization on the slope and level shifts. For a sufficiently large value of λ, the filter will select where to place the level shifts and determine their magnitudes, in order to minimize (5-1).

In order to directly apply the results and algorithms of the forth chapter, we can express the filter in the same form as the LASSO. This can be done by properly organizing the design matrix X as following:

$$X = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 2 \\ 1 & 1 & 1 & \cdots & 0 & 3 \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & n \end{bmatrix} \in \mathbb{R}^{n \times n+1}. \tag{5-2}$$

Thus, the matrix X is a concatenation of a $n \times n$ lower triangular matrix filled with ones and a sequential column vector. By doing $y = \tilde{X}\beta$, each coefficient $\beta_j \, \forall \, j = 2, \ldots, n$ corresponds to a level-shift in the filtered signal. In other words, they represent the level shifts at the position j. The slope coefficient is expressed by $\beta_{n+1} = \alpha$.

Additionally, the system is corrupted by an additive noise. In this sense, we can express the measurements by $y = X\beta + \varepsilon$.

$$\hat{\beta} \in \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \tfrac{1}{2} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right\} \tag{5-3}$$

Since the main goal of the filter is to perform noise reduction and properly recover the signal β, the chosen criteria for selecting a proper value of λ was the BIC. Additionally, given the fact that the LASSO estimator is usually biased we are going to use the *post*-LASSO technique. Considering that the number of slopes is probably very low, the vector β will be very sparse and consequently the obtaining the OLS estimator does not represent an extra challenge.

For this particular problem the design matrix is extremely coherent. As a result uniqueness is definitely not guaranteed. For most of the cases, the method can not tell the difference between choosing adjacent columns like x_j and x_{j+1} or any linear combination of them. To overcome this problem it is necessary to do a post-processing procedure with the objective of avoiding selecting consecutive coefficients of β.

Here we propose a heuristic based algorithm for avoiding this phenomenon. After solving 5-3 for the whole LASSO-path and selecting the best $\hat{\beta}(\lambda)$ according to the BIC, it is necessary to do an inspection on the solution

vector. When consecutive nonzero entries of $\hat{\beta}(\lambda)$ are found, it is necessary to choose only one of them to be nonzero. At this point, we have tried several strategies like selecting the first coefficient of each sequence or selecting the one on the middle. Empirically, selecting the last coefficient of each consecutive sequence has led to best results. After applying this strategy, we obtain the post-LASSO estimators using only the remaining nonzero coefficients.

5.3 Results

In order to validate the ℓ_1 Level-Slope Filter, two distinct experiments are made based on the aforementioned experiment. Since the experiments are made in laboratory, the exact location of the breakpoints are known. In this way, it possible to evaluate the precision of the filter in detecting a failure on the optical fiber.

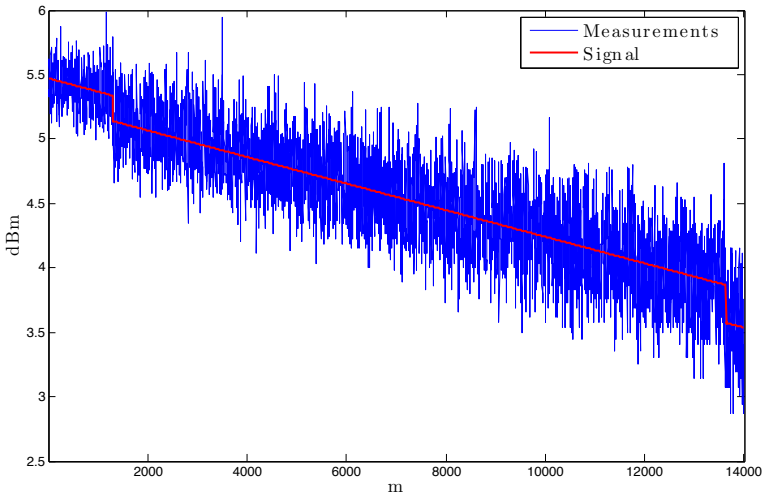

Figure 5.2: Signal recover for the first example.

The first example is done by taking seven thousand measurements on the range of 14,000 meters. As can be seen in Fig.[5.2], the filter correctly identifies the presence of two breakpoints. The precision on the level shift location can be evaluated at table [5.1]. The first breakpoint location has a difference of sixty-two meters to the location estimate by the filter. This may look very

Table 5.1: Breakpoint positions in meters for example 1.

Real positions	Estimate positions	Δ
1,370	1,308	62
13,664	13,644	20

Table 5.2: Breakpoint positions in meters for example 2.

Real positions	Estimate positions	Δ
2,770	2,740	30
15,064	15,046	18
17,200	17,172	28

imprecise, but it is actually a good approximation compared to the range of fourteen kilometers evaluated by the methodology.

In example 2 we have made 18,000 measurements. In this case, it is also possible to detect the end of the cable. The last breakpoint, at 17,200 meters, corresponds to end of the optic fiber cable. As can be seen at table [5.1], results for this example are slightly better, probably because more data is used to recover the filtered signal.

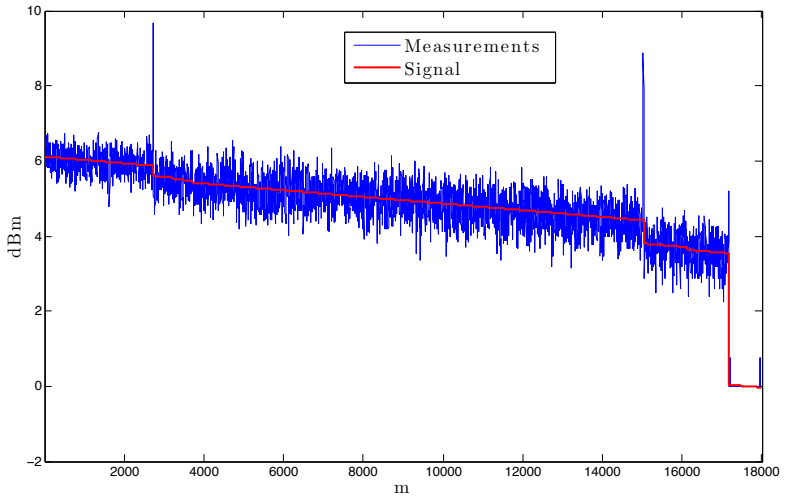

Figure 5.3: Signal recover for the second example.

5.4 Conclusion

In the previous examples, the ℓ_1 Level-Slope Filter has identified the position of the breakpoints with a reasonable precision. This result can be used to motivate a monitoring service for several types of fiber optics applications. An important detail is that, both ℓ_1 Level-Slope Filter and the Photon Counting OTDR can be employed without interrupting existing links on the fiber.

Several applications can benefit from this type of piece-wise linear filters. In macroeconomics, for instance, it can be used as an alternative to the Hodrick-Prescott filter [85]. Another potential application may arise in the

field of finance, where the ℓ_1 Level-Slope Filter may be used to quickly identify price shifts in stock prices.

In short, this application also suggests that the LASSO can be used to several purposes besides high-dimensional linear regression.

6
Case Study 2: Renewable Energy Stochastic Model

This chapter explores the advantages of using ℓ_1-regularization to build statistical models in a high-dimensional framework. More precisely, an application using the LASSO for estimation and selection of explanatory variables will be shown. A wide range of applications can be found in the statistical literature, for example in the field of biostatistics, where large DNA microarray [86][87], data-sets are analysed. In this work, we explore the usage of LASSO in the context of time-series analysis with a less frequent application.

Electrical grids are usually complex networks of thousands of nodes and thousands of lines [88]. As a consequence, large amounts of data are collected everyday by system operators, electric-distribution companies and generations companies. Additionally, increased modernization of electricity networks are bringing real-time data to reality. Considering that smart grid sensors can monitor several aspects of the system, e.g. nodal injection and weather conditions, making sense of this huge amount of data will be the future challenge for any system operator. More precisely, in this work, we focus on the challenge of modeling renewable energy supply.

Technological development associated with environmental concern has led to significant drop in renewable energy (RE) production costs, enabling a significant increase on RE shares, as compared to traditional fossil fuel generation, in the worldwide power system's matrix. However, the seasonality and variability intrinsic to these types of sources has become a challenge for power system agents. This uncertainty has a great impact on unit commitment and system planning. In a hydro-based power system, the capacity to anticipate variations in natural water inflow is of major importance for operation. This ability to properly forecast the increase of river's inflow can result in increased electric energy production due to enhanced flexibility in stored water management. Balancing demand and supply, without causing operational cost hikes, relies on the capacity of forecasting the renewable supplies, for instance, wind production and hydro inflow. For energy market participants, predicting the joint variability of these series is of most importance in their bidding and contracting strategies. Therefore, forecasting/simulation of RE supply has

become a valuable tool for power system agents both in the operation and commercialization sector.

Several studies using time-series modeling on short-term wind power are available [89–91]. Additionally, some work has been done using machine learning techniques [92, 93]. The majority of the works focused on modeling generated wind power per hour, mainly because most power systems perform the dispatch at every hour. Some work [94] [95] on a monthly basis has also been done. However, there is still a gap to be filled when it comes to the joint study of wind and hydro power, particularly in the case of a high-dimensional framework. This work addresses this problem, highlighting the particularities of modeling a large amount of renewable power plants.

In power systems with high penetration of renewable resources, the traditional linear modeling framework fails to estimate the unknown parameters. The high dimensionality curse arises due to the large number of generators, and the number of unknown parameters can easily exceed the number of observations. In order to overcome such problems, the estimation of the unknown parameters can be done under the assumption of sparsity of the regressors, i.e., the number of relevant regressors is much smaller than the number of potential explanatory variables.

6.1 Proposed model

In this section, we propose a model for wind power and hydro inflow monthly data using a multivariate time series approach. Given the physical relationship between renewable resources it is expected that many power plants affect each other intertemporally. Hence, we adopt a Vector Auto-Regressive with eXogenous variables (VARX) [96]. This class of auto-regressive models has been used in signal processing and econometrics, for stationary time-varying processes. Due to the multivariate structure, the model is able to capture linear interdependencies among several time series.

The exogenous variables can be composed by operative outputs of the power system dispatch. In Brazil, the most important operative variables are the reservoir inflows [97]. Given the Brazilian hydro based power system, the amount of reservoir usage is able to summarize the state of the system, which is related to energy prices. In view of the large size of the Brazilian system, the system is divided in four nodal areas (North, South, Southeast and Northeast subsystems). Each subsystem has its own storage capacity monitored by the reservoir inflow. On a different system, other operative variables may arise such as thermal power plant dispatch or energy spot prices.

Wind power and hydro inflow at time t are represented by the vector $Y_t \in \mathbb{R}^k$, where k is the sum of wind farms and hydro plants. The operative variables are denoted by $X_t \in \mathbb{R}^r$, where r represents the number of exogenous variables, and $\xi_t \in \mathbb{R}^k$ is a vector of Gaussian noise. The unknown parameters are $c \in \mathbb{R}^k$ a vector of intercepts, the Φ_i's are $k \times k$ coefficients matrices, the Θ_j's are $k \times r$ coefficient matrices and Σ is a $k \times k$ covariance matrix.

$$Y_t = c + \sum_{i=1}^{p} \Phi_i Y_{t-i} + \sum_{j=0}^{q} \Theta_j X_{t-j} + \xi_t \tag{6-1}$$

$$\xi_t \sim \mathcal{N}(0, \Sigma) \qquad \forall t = 1, ..., T \tag{6-2}$$

Furthermore, the analysis of renewable energy historical data suggests that variances and covariances also exhibit seasonal behavior assuming different values across months. This particular case of heteroskedasticity is commonly observed on physical phenomena. For example, during dry periods river inflow tends to decrease as well as inflow variability. In order to model these dynamics, it is necessary to impose a structure on the residual covariance matrix. Basically, it is necessary to relax the hypotheses of homoscedasticity and allows a periodic variance behavior. For instance, on a monthly basis twelve covariance matrices would describe the variance dynamics over a year. In this sense, the covariance matrix can be expressed by $\Sigma_{m(t)} = A_{m(t)} A_{m(t)}^{\top}$, where $m(t) \in \{1, 2, 3, ..., 12\}$ maps the temporal index $t = 1, ..., T$ into the respective months and $A_{m(t)}$'s are $k \times k$ matrices.

$$\xi_t = A_{m(t)} \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, I) \tag{6-3}$$

$$Y_t = c + \sum_{i=1}^{p} \Phi_i Y_{t-i} + \sum_{j=0}^{q} \Theta_j X_{t-j} + A_{m(t)} \varepsilon_t \qquad \forall \ t = 1, ..., T \tag{6-4}$$

Forecasting and simulations are done jointly for all elements of Y_t. It is expected that simulations, generated by such a model, are able to replicate the seasonal pattern and preserve the observed correlation of the renewable resources. For instance, complementarity between wind power and natural water inflow has to be present in the majority of all generated scenarios.

6.2 Estimation Algorithm

The proposed model requires the estimation of a set of unknown parameters $\psi = \{c, \{\Phi_i\}_{i=1}^{p}, \{\Theta_j\}_{j=1}^{q}, \{A_m\}_{m=1}^{12}\}$. In order to simplify the notation the model (6-4) is redefined as follows. Firstly, β is a $(12k^2 + k + krq) \times 1$ vector such that $\beta = vec([c\ \Phi_1\ \Phi_2\ ...\ \Phi_{12}\ \Theta_1\ \ ...\ \Theta_q])$. Secondly, W_{t-1} is

a sparse matrix with dimension $k \times (12k^2 + k + krq)$ given by $W_{t-1} = [1^\top \; Y_{t-1}^\top \; Y_{t-2}^\top \; ... \; Y_{t-12}^\top \; X_t^\top \; ... \; X_{t+1-q}^\top] \otimes I_{k\times k}$, where $\otimes$ denotes the Kronecker product.

$$Y_t = W_{t-1}\beta + A_{m(t)}\varepsilon_t \tag{6-5}$$

Given the aforementioned heteroskedasticity, the estimation of unknown parameters is not straightforward. As a consequence, traditional estimation methods lead to unreliable results. In this context, the literature offers some well-known methods for estimation under heteroskedasticity such as Weighted Least Squares [98] or GARCH [99]. Since, in this particular case, the variance has a periodic behavior we can design an *ad hoc* method based on the maximum likelihood criteria. Initially, It is necessary to maximize the log-likelihood function:

$$\ell(\psi) = \frac{Tk}{2}\ln(2\pi) + \frac{1}{2}\sum_{t=13}^{T}\ln(|\Sigma_{m(t)}^{-1}|) \tag{6-6}$$
$$-\frac{1}{2}\sum_{t=13}^{T}(Y_t - W_{t-1}\beta)^\top \Sigma_{m(t)}^{-1}(Y_t - W_{t-1}\beta)$$

Obtaining the derivative of (6-6) with respect to β and $\{\Sigma_i\}_{i=1}^{12}$ leads to the following nonlinear system:

$$\hat{\Sigma}_m = \sum_{t=13|m(t)=m}^{T} \frac{(Y_t - W_{t-1}\beta)(Y_t - W_{t-1}\beta)^\top}{n} \tag{6-7}$$
$$\forall \;\; m = 1, 2, ..., 12$$
$$\hat{\beta} = \left[\sum_{t=13}^{T} W_{t-1}^\top \Sigma_{m(t)}^{-1} W_{t-1}\right]^{-1}\left[\sum_{t=13}^{T} W_{t-1}^\top \Sigma_{m(t)}^{-1} Y_t\right] \tag{6-8}$$

Unlike in the homoscedastic case, such a system cannot be solved by substitution. Several numerical methods for solving nonlinear systems are available in the literature. The following algorithm is a newton based method that computes fixed-points iteratively, i.e. a point of the function's domain. The initial condition for $\hat{\Sigma}_m^0$ is set to the $k \times k$ identity matrix, then the solution of (6-7) is computed on (6-8) and the solution of (6-8) is computed on (6-7). This procedure is repeated recursively until a solution that satisfy both (6-7) and (6-8) is fond.

At first sight, the fixed-point algorithm 4 can handle the estimation for the model 6-4. However, for most of the instances the number of unknown parameters is considerably large and may exceed the number of observations.

Algorithm 4 Fixed-Point Method

iter $\leftarrow 0$
$\hat{\Sigma}_m^0 \leftarrow I_{k\times k} \quad \forall m = 1, 2, ..., 12$
Obtain $\hat{\beta}^0$ from (6-8) using $\hat{\Sigma}_m^0$
while $||\hat{\beta}^{iter} - \hat{\beta}^{iter-1}||_2 > tolerance$ **do**
  iter $\leftarrow$ iter + 1
  Obtain $\hat{\Sigma}_m^{iter}$ from (6-7) using $\hat{\beta}^{iter-1}$ $\forall m = 1, 2, ..., 12$
  Obtain $\hat{\beta}^{iter}$ from (6-8) using $\hat{\Sigma}_m^{iter}$
end while

For example, if six renewable power plants are being modeled, i.e. $k = 6$, the vector β is 486-dimensional which is clearly high-dimensional. Thus, for the reasons discussed in chapter 2, the fixed-point algorithm needs to be refined.

Firstly, given a high-dimensional framework the matrices Σ_m are usually singular. Thus, the inverse matrices Σ_m^{-1} are obtained by the Moore-Penrose pseudo-inverse [9]. At each iteration of the algorithm the pseudo-inverse is calculated via Singular Value Decomposition. The method used to obtain all eigenvalues and eigenvectors was the Jacobi eigenvalue algorithm [7], which takes advantage of the symmetry of covariance matrices.

Secondly, it is necessary to introduce an ℓ_1-regularization in order to induce sparsity on the estimation of β. In this fashion, just as the LASSO, the estimation procedure can select which coefficients are more important to explain the behavior of Y_t over time. For convenience the Lagrangian formulation $\ell^*(\beta, \{\Sigma_m\}_{m=1}^{12}) = \ell(\beta, \{\Sigma_m\}_{m=1}^{12}) - \lambda||\beta||_1$ is used.

Accordingly, we now need to solve multiple fixed-point algorithms. In other words, there will be a LASSO-like path of fixed-point algorithms for different values of λ. In each iteration of the fixed-point, the problem $\ell(\beta, \{\Sigma_m\}_{m=1}^{12}) - \lambda||\beta||_1$ will be solved by quadratic programming techniques as described on the fourth chapter.

Similarly to the LASSO, we can establish the minimum λ for which all elements of β are zero (6-9). The Karush-Kuhn-Tucker conditions state that on the optimal solution:

$$\mathbf{0} \in \{2\sum_{t=13}^{T} W_{t-1}^{\top}\Sigma_{m(t)}^{-1}(Y_t - W_{t-1}\beta) - \lambda \text{sign}(\beta) : \beta \in \mathbb{R}^k\}$$

Where:

$$sign(\beta_j) \in \begin{cases} \{+1\} & ,if\ \beta_j > 0 \\ [-1, +1] & ,if\ \beta_j = 0 \\ \{-1\} & ,if\ \beta_j < 0 \end{cases}$$

Then for $\beta = \mathbf{0}$ to be the optimal solution,

$$2\sum_{t=13}^{T} W_{t-1}^{\top}\Sigma_{m(t)}^{-1}Y_t \in [-\lambda, +\lambda]$$

This will happen for every

$$\lambda \geq \lambda^* = \left|\left| 2\left|\sum_{t=13}^{T} W_{t-1}^{\top}\Sigma_{m(t)}^{-1}Y_t\right|\right|\right|_{\infty} \tag{6-9}$$

Combining the aforementioned ideas we have the following estimation algorithm:

Algorithm 5 Fixed-Point & ℓ^1-regularization

$\hat{\Sigma}_m^0 \leftarrow I_{k\times k} \quad \forall m = 1, 2, ..., 12$
for $\lambda = 0$ to λ^* **do**
  iter $\leftarrow 0$
  $\hat{\beta}^0 = argmax\{\ell^*(\beta, \{\hat{\Sigma}_m^0\}_{m=1}^{12})\}$
  while $||\hat{\beta}^{iter} - \hat{\beta}^{iter-1}||_2 > tolerance$ **do**
    iter $\leftarrow$ iter + 1
    Obtain $\hat{\Sigma}_m^{iter}$ from (6-7) using $\hat{\beta}^{iter-1}$ $\forall m = 1, ..., 12$
    Calculate the pseudo-inverse Σ_m^{-1} via SVD $\forall m = 1, ..., 12$
    $\hat{\beta}^{iter} = argmax\{\ell^*(\beta, \{\hat{\Sigma}_m^{iter}\}_{m=1}^{12})\}$
  end while
  Obtain $BIC(\lambda)$ using $\hat{\beta}^{iter}$ and $\{\hat{\Sigma}_m^{iter}\}_{m=1}^{12}$
end for

Finally, we select the parameters $\hat{\beta}$ and $\{\hat{\Sigma}_m\}_{m=1}^{12}$ and λ which minimize $BIC(\lambda)$. After the estimation of unknown parameters it is possible to generate consistent scenarios for wind production and hydro power via Monte-Carlo [100] or Bootstrap [101]. Furthermore, the model can be used for forecasting the potential renewable energy for the forthcoming months.

6.3 Brazilian Power System

The novel methodology was tested for a subset of the Brazilian power system comprising 16 wind farms and 34 hydro plants. The time-series are composed of monthly data from January 1981 to December 2011 totalizing 372 observations.

Wind energy is represented by the capacity factor (%), i.e. the percentage of installed capacity (MW) that was generated (avgMW). Sixteen wind farms were analyzed, most of them located in the South and the Northeast of Brazil. It is well known that these regions are wind-rich areas. Furthermore, recent

prospections have shown that the whole production capacity is greater then the actual installed capacity of the entire Brazilian system. Opposed to short-term observations, the wind power is stationary on a monthly basis.

The data for hydro power is composed of monthly mean river inflow (m^3/s) where hydro plants are situated. Available data from the Operator of the National Electricity System (ONS) consisting of thirty-four rivers was used. In the case of the Brazilian power system, hydro power is highly representative. In 2009 according to ONS more than 80% of the generated energy came from hydro power. Due to environmental concerns, the most recent hydro plants are run-off river plants. These facts reinforce the interest in predicting river inflow since run-off river plants have considerably smaller reservoirs.

The explanatory variables are represented by the reservoir inflows of the main four Brazilian subsystems. Besides the inter-temporal observation, the lagged observation, i.e. the former month observed of reservoir inflow is used.

To begin with, it is necessary to estimate the unknown parameters of the VARX model. For the fifty power plants these parameters are β, a vector $30,450 \times 1$, and twelve 50×50 covariance matrices. The estimation was done by applying Algorithm 3 and choosing the best λ according to the BIC criteria which was 16.7. The whole procedure has taken sixteen minutes on an Intel®Core(TM) i7-3960C with a CPU of 3.3 GHz and 64 GB of RAM, using Xpress-MP 7.5 under Mosel computer. An amount of 29,582 elements of β were shrunk to zero. Which represents 97.15 % of the potential explanatory variables.

The next step consists of generating scenarios for a long time horizon. In this particular study the model will be used to project possible future outcomes for the forthcoming four years. An amount of two thousand scenarios were generated by re-injecting boostrapped residuals back into the model. Techniques for simulating scenarios via bootstrap are described in detail on [102] and [103].

Fig.6.1 and Fig.6.2 display the scenarios to a particular wind power plant and one river water inflow. The purple shaded area correspond to the thirty-one years of stacked historical data. Thereafter, the two thousand scenarios for a four years horizons are stacked. Each sequence of colored point corresponds to a different scenario over time. It is possible to notice that the model is able to reproduce the seasonal pattern observed on the historical data. In addition, the seasonal complementarity can be easily noticed through the four years of simulated data.

More accurate graphs are presented to measure the quality of generated scenarios. The following figures Fig.6.3 and Fig.6.4 compare, on a monthly

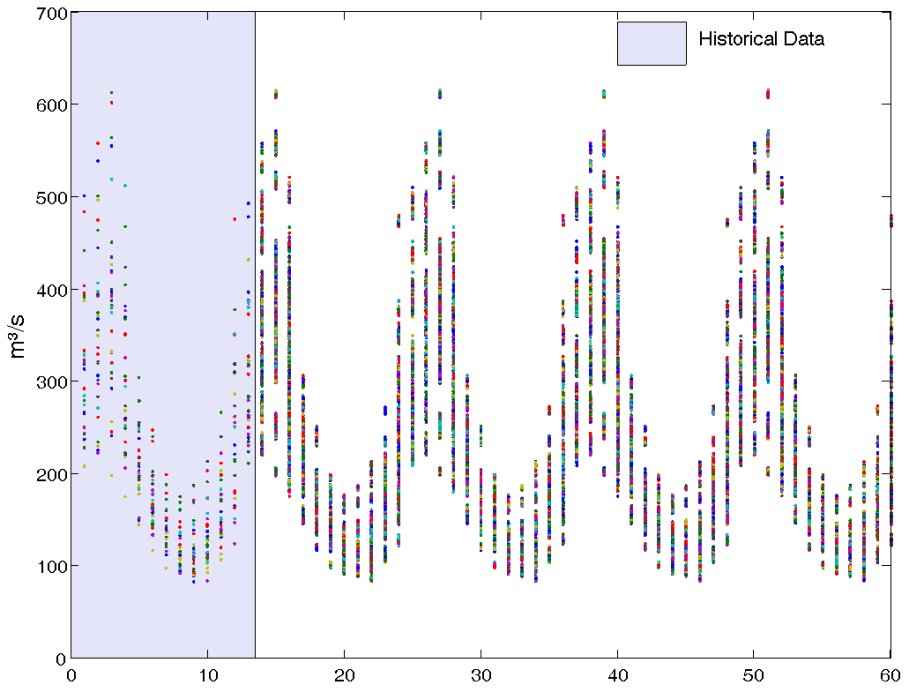

Figure 6.1: Simulated scenarios for water inflow at Barra dos Coqueiros.

basis, historical data to the quantiles of the two thousand scenarios. Fig.6.3 refers to Icaraizinho's wind farm, located in the Northeastern area of Brazil, and Fig.6.4 refers to the natural water inflow of the hydro plant of Tocantins, located on the North region of Brazil. The stacked points represent the historical data for each month, the black line represents the median, the blue line represents the upper 95% quantile and the lower 5% quantile and the red line represents the maximum and minimum for each month of the generated scenarios.

Firstly, most of the historical data is centered on the median. Secondly, one can verify that most of the monthly observed data are within the blue line band which represents a good fit for the generated scenarios.

Forecasting results are presented in TABLE 6.1 and on TABLE 6.2 using three well-known measures of accuracy. To begin with, the R^2 is obtained for each time-series. Also known as coefficient of determination, the R^2 ranges from 0 to 1 and corresponds to the relative amount of total variation that can be explained by the model. Next, it is measured the Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). Both these metrics are based on the absolute deviation between the observed data and the forecast.

The analyses of TABLE 6.1 and TABLE 6.2 suggest that the model

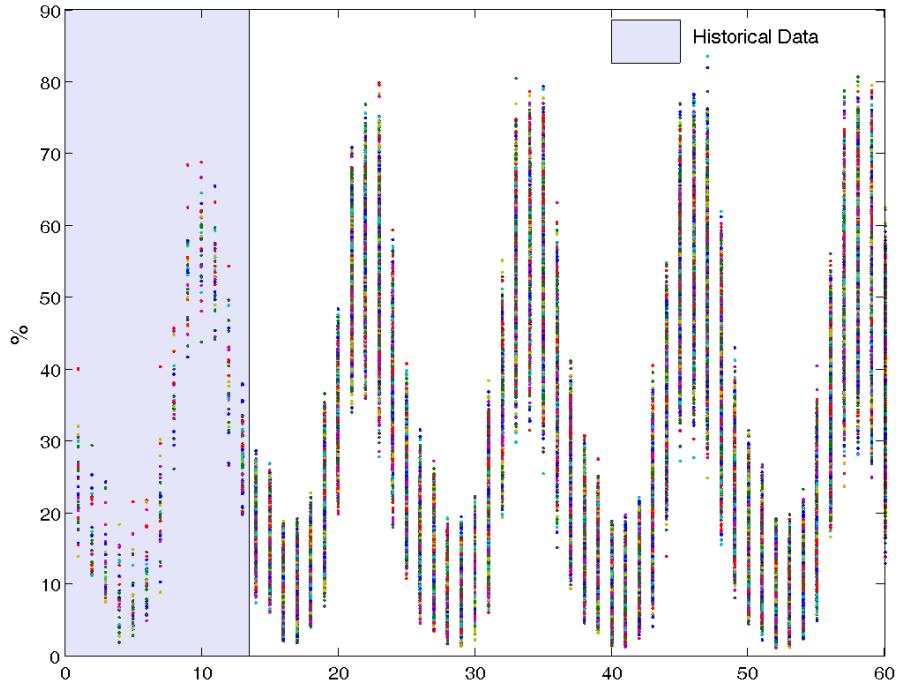

Figure 6.2: Simulated scenarios for Praia do Morgado's wind farm.

has shown a reasonable forecasting accuracy. However a few time series were particularly poorly predicted, such as Itaipu and Peixe Angical. The time-series of Itaipu corresponds to the Itaipu Dam, one of the largest hydroelectric power station in the world with 14,000 MW installed capacity. Due to these considerable dimensions it is very likely that it's underlying stochastic process differs from most part of the others natural river inflow.

These predictions could be used to formulate hedging techniques to reduce the generator exposure, e.g. a wind farm that has a minor forecasted generation for the next month. In order to avoid the volatility of the spot markets, one could define a hedging plan by means of financial instruments such as options purchasing or swaps. An alternative approach consists of energy storage methods, like pumped storage reservoirs or compressed air facilities. A comparison of such strategies has been made by [104] regarding the uncertainty of wind power generation.

Furthermore, it could be useful to enforce an efficient management of thermal energy on unit commitment and economic dispatch. Considering that most thermal units only support a gradual change of temperature, if the model forecasts lower renewable energy supply then thermal units could be brought on-line in advance. The influence of wind power on thermal system operation

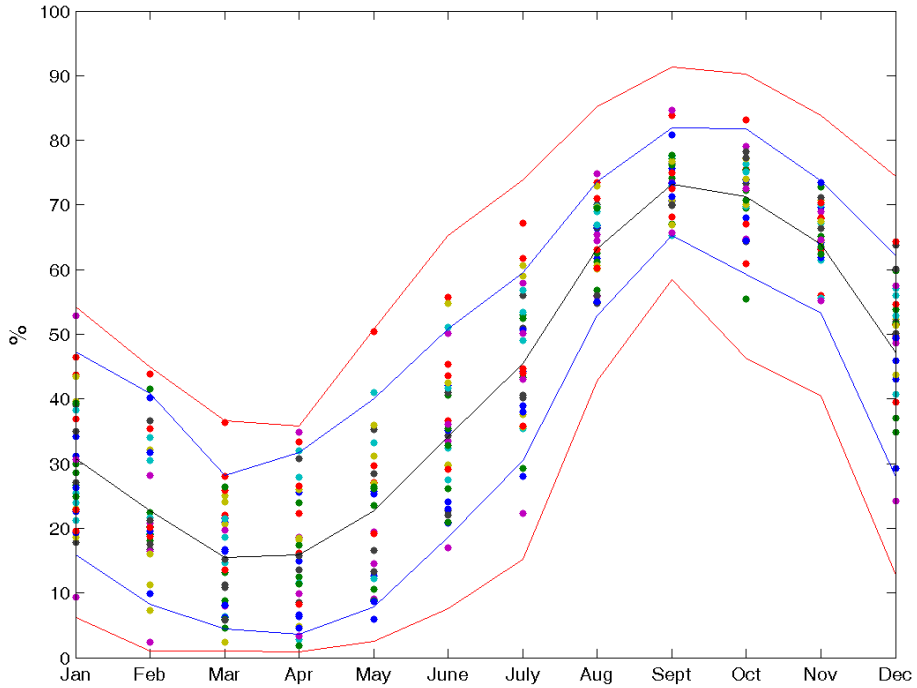

Figure 6.3: Quantiles of Icaraizinho.

is explored by [105] regarding the Dutch power system.

Particularly on the Brazilian power system, the generated scenarios can be useful for the system expansion and economic dispatch. The suggested model could be adapted to NEWAVE [97], the current dispatch methodology adopted in Brazil. In this fashion, the aforementioned simulations could be used as an input to the multistage stochastic optimization problem [106]. As a consequence, it is expected that the energy price will respond more precisely to fluctuations on the renewable generation.

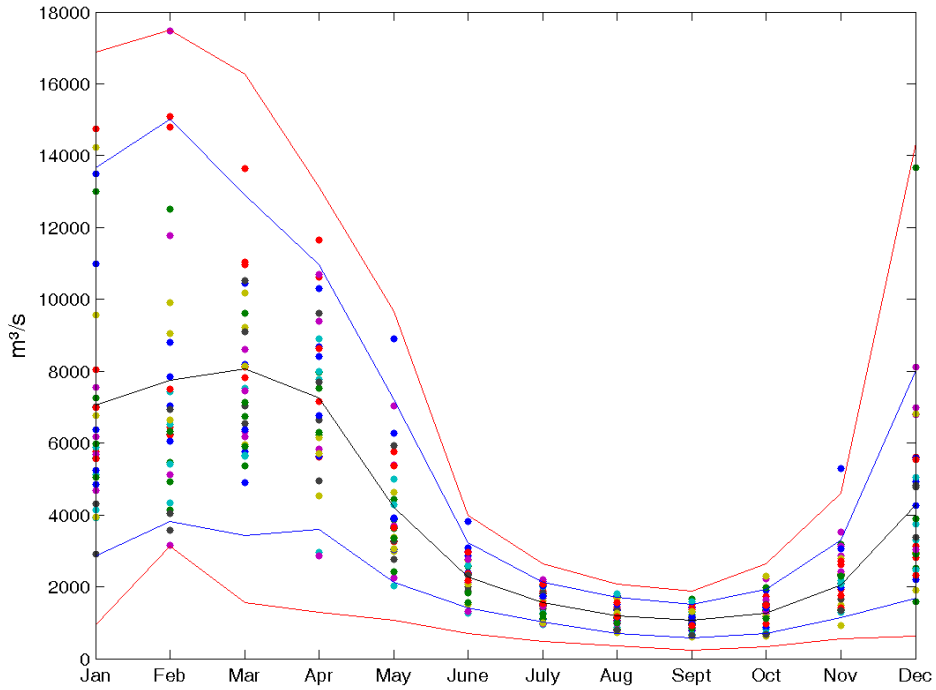

Figure 6.4: Quantiles of Tocantins.

Table 6.1: Forecasting accuracy measures

Wind Farms	MAE	MAPE	R^2
Alegria 1	3.76	21.59	0.75
Alegria 2	7.14	18.33	0.63
Bons Ventos	6.45	21.49	0.76
Canoa Quebrada	8.21	40.64	0.81
Cerro Chato	4.04	11.08	0.51
Cerro Chato 2	4.14	11.16	0.51
Icaraizinho	7.3	26.67	0.88
Mangue Seco 1	4.6	14.36	0.74
Mangue Seco 2	3.26	9.01	0.51
Mangue Seco 3	3.03	8.39	0.55
Mangue Seco 4	3.77	14.28	0.79
Praia do Morgado	4.18	21.46	0.93
Praia Formosa	4.59	20.47	0.86
Rio do Fogo	5.12	18.38	0.68
Sangradouro	5.43	16.59	0.64
Volta do Rio	5.82	22.61	0.76

Table 6.2: Forecasting accuracy measures

Natural Water Inflow	MAE	MAPE	R^2
Salto Verdinho	39.69	14.99	0.69
Vigario	9.4	7.18	0.64
Itaguaçu	45.56	16.52	0.75
Pereira Passos	22.44	18.37	0.61
Teles Pires	202.17	9.07	0.88
Santana	11.83	10.83	0.62
Ferreira Gomes	105.33	13.18	0.85
Ilha dos Pombos	221.27	36.11	0.59
Santa Cecília	16.37	15.33	0.57
Belo Monte	1191.5	14.95	0.85
Dardanelos	58.16	28.61	0.91
Salto	34.97	15.26	0.66
Santo Antonio do Jari	238.19	15.96	0.83
Tocos	5.15	50.86	0.53
Olho D'Água	16.04	16.21	0.58
Jupia	2192.4	22.76	0.72
Coaracy Nunes	196.21	20.72	0.85
Manso	34.34	17.59	0.72
Ponte de Pedra	6.77	9.26	0.77
Samuel	79.32	65.65	0.82
Santa Isabel	755.25	17.15	0.92
Balbina	110.36	19.5	0.68
Estreito Tocantins	1190.2	21.2	0.67
Lajeado	716.42	26.2	0.48
Tucurui	1420.9	12.06	0.9
Jirau	2258.9	25.03	0.91
Foz do Rio Claro	51.86	18.42	0.79
Guilman Amorim	25.46	30.88	0.54
Itaipu	2732.7	18.54	0.45
Itiquira	11.97	11.74	0.75
Peixe Angical	669.52	35.42	0.43
Porto Estrela	77.42	52.27	0.49
Barra dos Coqueiros	54.18	24.74	0.65
Cacu	45.06	18.58	0.79

6.4 Conclusion

Firstly, the proposed methodology, based on the maximum likelihood and the ℓ_1-regularization, was capable of estimating parameters for high-dimensional models within a reasonable time. The proposed model was able to generate scenarios that reproduces the observed dynamics, such as seasonality and complementarity between hydro and wind power.

Further research also suggest evaluating the model performance on different power systems. Specially on power systems with considerable participation of wind power. given the growing of solar energy installations, it is relevant to include solar plants to the proposed model. Positive results could be used to foster installation of solar energy.

Since the model is originally designed to tackle high-dimensional data, an application to smart grids seems natural. Considering that smart grid sensors can monitor nodal injection and weather conditions, a large amount of real-time data could be contemplated by the model. In this way, scenario simulations would provide the grid with relevant information to consider renewable energy uncertainty on the power management.

7
Discussions

We applied the LASSO shrinkage method to two problems from different fields. In both cases, this technique was successful tool for developing solutions to complex models. This result suggests that a proper usage of ℓ_1-norm can be the key ingredient for solving a wide variety of large-scale problems. In the following, we point out two important aspects that can foster future research regarding ℓ_1-regularization.

7.1
Tailor-made solution methods

Since the applications of ℓ_1-regularization usually are large-scale problems, the algorithm efficiency is extremely critical. In this regard, certain researches are proposing better solution methods. In several situations, the algorithms presented at the forth chapter may benefit from a particular structure of the problem.

Some variations of traditional methods is available at the optimization literature. Works like [107] and [84] suggests custom interior-points methods. On the other hand, the work of [108] suggests a block coordinate descent strategy.

On the ℓ_1 Level-Slope Filter, given the special structure of the design matrix, it is very likely that the method would benefit from a custom coordinate descent algorithm. For the VARX with the balancing inequality, an efficient algorithm, like a custom interior-points method, is extremely important since quadratic programming methods are usually slow for large scale constrained problems.

7.2
Multiresponse regularization

Despite the fact that the proposed methodology for the VARX has led to reasonable results, there is a clear imbalance between the predict power among the endogenous variables. Note for instance the differences between the R^2 and MAPE for *Santa Isabel* and *Porto Estrela*. Some level of divergence is expected

considering that some variables are more well-behaved than others. However, a model with a more homogeneous prediction accuracy would be preferable. In this sense, there is a need to develop a regularization method that can take into account that the model is multiresponse. This would allow some kind of control over the shrinkage in between the different responses.

Bibliography

[1] J. Dean and S. Ghemawat, *MapReduce: simplified data processing on large clusters*, **Communications of the ACM**, vol. 51, no. 1, pp. 107–113, 2008.

[2] D. Düllmann, *Petabyte databases* in **ACM SIGMOD Record**, vol. 28, p. 506, ACM, 1999.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *et al.*, **Introduction to algorithms**, vol. 2. MIT press Cambridge, 2001.

[4] A. M. Legendre, **Nouvelles méthodes pour la détermination des orbites des cometes**. F. Didot, 1805.

[5] K. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientum (Hamburg: Perthes und Besser) tr., Davis, CH 1857. Theory of the motion of the heavenly bodies moving about the sun in conic sections (Boston: Little, Brown & Company). Original reprinting, 1981*, **Werke, Bd**, vol. 7, pp. 1–280.

[6] J. Wooldridge, **Introductory econometrics: A modern approach**. Cengage Learning, 2012.

[7] G. H. Golub and C. F. Van Loan, **Matrix computations**, vol. 4. JHU Press, 2012.

[8] R. L. Plackett, *Some theorems in least squares*, **Biometrika**, vol. 37, no. 1-2, pp. 149–157, 1950.

[9] R. Penrose, *A generalized inverse for matrices* in **Proc. Cambridge Philos. Soc**, vol. 51, pp. 406–413, Cambridge Univ Press, 1955.

[10] S. Konishi and G. Kitagawa, *Generalised information criteria in model selection*, **Biometrika**, vol. 83, no. 4, pp. 875–890, 1996.

[11] H. Riedwyl, *Goodness of fit*, **Journal of the American Statistical Association**, vol. 62, no. 318, pp. 390–398, 1967.

[12] R. R. Hocking, *A Biometrics invited paper. The analysis and selection of variables in linear regression*, **Biometrics**, pp. 1–49, 1976.

[13] R. B. Bendel and A. A. Afifi, *Comparison of stopping rules in forward stepwise regression*, **Journal of the American Statistical Association**, vol. 72, no. 357, pp. 46–53, 1977.

[14] J. A. Doornik, *Autometrics* in **in Honour of David F. Hendry**, Citeseer, 2009.

[15] C. Epprecht, D. Guegan, Á. Veiga, *et al.*, *Comparing variable selection techniques for linear regression: LASSO and Autometrics*, 2013.

[16] H. Hotelling, *Analysis of a complex of statistical variables into principal components.*, **Journal of educational psychology**, vol. 24, no. 6, p. 417, 1933.

[17] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, *Prediction by supervised principal components*, **Journal of the American Statistical Association**, vol. 101, no. 473, 2006.

[18] D. Paul, E. Bair, T. Hastie, and R. Tibshirani, *" Preconditioning" for feature selection and regression in high-dimensional problems*, **The Annals of Statistics**, pp. 1595–1618, 2008.

[19] D. L. Donoho, *De-noising by soft-thresholding*, **Information Theory, IEEE Transactions on**, vol. 41, no. 3, pp. 613–627, 1995.

[20] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, **Technometrics**, vol. 12, no. 1, pp. 55–67, 1970.

[21] L. Breiman, *Better subset regression using the nonnegative garrote*, **Technometrics**, vol. 37, no. 4, pp. 373–384, 1995.

[22] R. Tibshirani, *Regression shrinkage and selection via the lasso*, **Journal of the Royal Statistical Society. Series B (Methodological)**, pp. 267–288, 1996.

[23] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, **Biostatistics**, vol. 9, no. 3, pp. 432–441, 2008.

[24] Y. Kim, H. Choi, and H.-S. Oh, *Smoothly clipped absolute deviation on high dimensions*, **Journal of the American Statistical Association**, vol. 103, no. 484, pp. 1665–1673, 2008.

[25] E. Candes and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n*, **The Annals of Statistics**, pp. 2313–2351, 2007.

[26] D. L. Donoho, *Compressed sensing*, **Information Theory, IEEE Transactions on**, vol. 52, no. 4, pp. 1289–1306, 2006.

[27] D. L. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ1 minimization*, **Proceedings of the National Academy of Sciences**, vol. 100, no. 5, pp. 2197–2202, 2003.

[28] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, **SIAM review**, vol. 51, no. 1, pp. 34–81, 2009.

[29] J. A. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, **Information Theory, IEEE Transactions on**, vol. 52, no. 3, pp. 1030–1051, 2006.

[30] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, **The elements of statistical learning**, vol. 2. Springer, 2009.

[31] H. D. Grossman, *The twelve-coin problem*, **Scripta Mathematica**, vol. 11, pp. 360–363, 1945.

[32] D. T. Lee, *JPEG 2000: retrospective and new developments*, **Proceedings of the IEEE**, vol. 93, no. 1, pp. 32–41, 2005.

[33] S. Zheng and W. Liu, *An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification*, **Computers in biology and medicine**, vol. 41, no. 11, pp. 1033–1040, 2011.

[34] C. S. Kubrusly, **The elements of operator theory**. Springer, 2011.

[35] E. J. Candes and T. Tao, *Decoding by linear programming*, **Information Theory, IEEE Transactions on**, vol. 51, no. 12, pp. 4203–4215, 2005.

[36] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, **Information Theory, IEEE Transactions on**, vol. 52, no. 2, pp. 489–509, 2006.

[37] J. A. Tropp, *Greed is good: Algorithmic results for sparse approximation*, **Information Theory, IEEE Transactions on**, vol. 50, no. 10, pp. 2231–2242, 2004.

[38] L. Welch, *Lower bounds on the maximum cross correlation of signals (Corresp.)*, **Information Theory, IEEE Transactions on**, vol. 20, no. 3, pp. 397–399, 1974.

[39] D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*, 1997.

[40] L. A. Wolsey, **Integer programming**, vol. 42. Wiley New York, 1998.

[41] B. K. Natarajan, *Sparse approximate solutions to linear systems*, **SIAM journal on computing**, vol. 24, no. 2, pp. 227–234, 1995.

[42] C. H. Papadimitriou and K. Steiglitz, **Combinatorial optimization: algorithms and complexity**. Courier Dover Publications, 1998.

[43] R. C. Thompson, *System Identification Via Basis Pursuit*. PhD thesis, Arizona State University, 2012.

[44] R. M. Karp, **Reducibility among combinatorial problems**. Springer, 1972.

[45] P. Frossard, P. Vandergheynst, R. M. Figueras i Ventura, and M. Kunt, *A posteriori quantization of progressive matching pursuit streams*, **Signal Processing, IEEE Transactions on**, vol. 52, no. 2, pp. 525–535, 2004.

[46] R. Gribonval and E. Bacry, *Harmonic decomposition of audio signals with matching pursuit*, **Signal Processing, IEEE Transactions on**, vol. 51, no. 1, pp. 101–111, 2003.

[47] T. Nguyen and A. Zakhor, *Matching pursuits based multiple description video coding for lossy environments* in **Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on**, vol. 1, pp. I–57, IEEE, 2003.

[48] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, *Robust face recognition via sparse representation*, **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, vol. 31, no. 2, pp. 210–227, 2009.

[49] E. J. Candès and M. B. Wakin, *An introduction to compressive sampling*, **Signal Processing Magazine, IEEE**, vol. 25, no. 2, pp. 21–30, 2008.

[50] E. Candes and J. Romberg, *l1-magic: Recovery of sparse signals via convex programming*, **URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf**, vol. 4, 2005.

[51] M. Elad, **Sparse and redundant representations: from theory to applications in signal and image processing**. Springer, 2010.

[52] K. B. Athreya and S. N. Lahiri, **Measure theory and probability theory**. Springer, 2006.

[53] R. Koenker, **Quantile regression**. No. 38, Cambridge university press, 2005.

[54] S. P. Boyd and L. Vandenberghe, **Convex optimization**. Cambridge university press, 2004.

[55] H. Zou, *The adaptive lasso and its oracle properties*, **Journal of the American statistical association**, vol. 101, no. 476, pp. 1418–1429, 2006.

[56] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, vol. 68, no. 1, pp. 49–67, 2006.

[57] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, *Genome-wide association analysis by lasso penalized logistic regression*, **Bioinformatics**, vol. 25, no. 6, pp. 714–721, 2009.

[58] R. Tibshirani, *Regression shrinkage and selection via the lasso: a retrospective*, **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, vol. 73, no. 3, pp. 273–282, 2011.

[59] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, **SIAM journal on scientific computing**, vol. 20, no. 1, pp. 33–61, 1998.

[60] K. Siedenburg and M. Dörfler, *Audio denoising by generalized time-frequency thresholding* in **Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio**, Audio Engineering Society, 2012.

[61] M. Elad and M. Aharon, *Image denoising via learned dictionaries and sparse representation* in **Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on**, vol. 1, pp. 895–900, IEEE, 2006.

[62] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, **Nonlinear programming: theory and algorithms**. John Wiley & Sons, 2013.

[63] R. R. Picard and R. D. Cook, *Cross-validation of regression models*, **Journal of the American Statistical Association**, vol. 79, no. 387, pp. 575–583, 1984.

[64] J. Shao, *Linear model selection by cross-validation*, **Journal of the American statistical Association**, vol. 88, no. 422, pp. 486–494, 1993.

[65] S. Chand, *On tuning parameter selection of lasso-type methods-A Monte Carlo study* in **Applied Sciences and Technology (IBCAST), 2012 9th International Bhurban Conference on**, pp. 120–129, IEEE, 2012.

[66] G. Schwarz, *Estimating the dimension of a model*, **The annals of statistics**, vol. 6, no. 2, pp. 461–464, 1978.

[67] Friedman, Jerome and Hastie, Trevor and Tibshirani, Robert, *GLMnet for Matlab* -

[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, *Scikit-learn: Machine learning in Python*, **The Journal of Machine Learning Research**, vol. 12, pp. 2825–2830, 2011.

[69] I. Adler, M. G. Resende, G. Veiga, and N. Karmarkar, *An implementation of Karmarkar's algorithm for linear programming*, **Mathematical programming**, vol. 44, no. 1-3, pp. 297–335, 1989.

[70] P. Wolfe, *The simplex method for quadratic programming*, **Econometrica: Journal of the Econometric Society**, pp. 382–398, 1959.

[71] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *et al.*, *Least angle regression*, **The Annals of statistics**, vol. 32, no. 2, pp. 407–499, 2004.

[72] S. Weisberg, **Applied linear regression**, vol. 528. John Wiley & Sons, 2005.

[73] R. J. Tibshirani, *The lasso problem and uniqueness*, **Electronic Journal of Statistics**, vol. 7, pp. 1456–1490, 2013.

[74] W. J. Fu, *Penalized regressions: the bridge versus the lasso*, **Journal of computational and graphical statistics**, vol. 7, no. 3, pp. 397–416, 1998.

[75] S. K. Shevade and S. S. Keerthi, *A simple and efficient algorithm for gene selection using sparse logistic regression*, **Bioinformatics**, vol. 19, no. 17, pp. 2246–2253, 2003.

[76] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, **Communications on pure and applied mathematics**, vol. 57, no. 11, pp. 1413–1457, 2004.

[77] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, *et al.*, *Pathwise coordinate optimization*, **The Annals of Applied Statistics**, vol. 1, no. 2, pp. 302–332, 2007.

[78] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, **Journal of statistical software**, vol. 33, no. 1, p. 1, 2010.

[79] D. P. Bertsekas, *Nonlinear programming*, 1999.

[80] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, **Mathematical Programming**, vol. 117, no. 1-2, pp. 387–423, 2009.

[81] P. Tseng *et al.*, **Coordinate ascent for maximizing nondifferentiable concave functions**. Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 1988.

[82] G. Amaral, *FPGA Applications on Photon Detection Systems* Master's thesis, Pontifical Catholic University, Rio de Janeiro, Brazil, 2014.

[83] P. Eraerds, M. Legré, J. Zhang, H. Zbinden, and N. Gisin, *Photon counting OTDR: advantages and limitations*, **Journal of Lightwave Technology**, vol. 28, no. 6, pp. 952–964, 2010.

[84] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, *\ell_1 Trend Filtering*, **Siam Review**, vol. 51, no. 2, pp. 339–360, 2009.

[85] R. J. Hodrick and E. C. Prescott, *Postwar US business cycles: an empirical investigation*, **Journal of Money, credit, and Banking**, pp. 1–16, 1997.

[86] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, *Class prediction by nearest shrunken centroids, with applications to DNA microarrays*, **Statistical Science**, pp. 104–117, 2003.

[87] D. Amaratunga, J. Cabrera, and Z. Shkedy, **Exploration and Analysis of DNA Microarray and Other High-Dimensional Data**. John Wiley & Sons, 2014.

[88] G. A. Pagani and M. Aiello, *The power grid as a complex network: a survey*, **arXiv preprint arXiv:1105.3338**, 2011.

[89] B. G. Brown, R. W. Katz, and A. H. Murphy, *Time series models to simulate and forecast wind speed and wind power*, **Journal of climate and applied meteorology**, vol. 23, pp. 1184–1195, 1984.

[90] J. L. Torres, A. Garcia, M. De Blas, and A. De Francisco, *Forecast of hourly average wind speed with ARMA models in Navarre (Spain)*, **Solar Energy**, vol. 79, no. 1, pp. 65–77, 2005.

[91] Z. Huang and Z. Chalabi, *Use of time-series analysis to model and forecast wind speed*, **Journal of Wind Engineering and Industrial Aerodynamics**, vol. 56, no. 2, pp. 311–322, 1995.

[92] I. G. Damousis, M. C. Alexiadis, J. B. Theocharis, and P. S. Dokopoulos, *A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation*, **Energy Conversion, IEEE Transactions on**, vol. 19, no. 2, pp. 352–361, 2004.

[93] G. Kariniotakis, G. Stavrakakis, and E. Nogaret, *Wind power forecasting using advanced neural networks models*, **Energy conversion, ieee transactions on**, vol. 11, no. 4, pp. 762–767, 1996.

[94] A. N. Celik, *Energy output estimation for small-scale wind power generators using Weibull-representative wind data*, **Journal of Wind Engineering and Industrial Aerodynamics**, vol. 91, no. 5, pp. 693–707, 2003.

[95] V. Quintana and A. Chikhani, *A stochastic model for mid-term operation planning of hydro-thermal systems with random reservoir inflows*, **Power Apparatus and Systems, IEEE Transactions on**, no. 3, pp. 1119–1127, 1981.

[96] H. Lütkepohl, *New introduction to multiple time series analysis*, 2005.

[97] M. V. Pereira, L. A. Barroso, and J. Rosenblatt, *Supply adequacy in the Brazilian power market* in **Power Engineering Society General Meeting, 2004. IEEE**, pp. 1016–1021, IEEE, 2004.

[98] J. D. Hamilton, **Time series analysis**, vol. 2. Cambridge Univ Press, 1994.

[99] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, **Journal of econometrics**, vol. 31, no. 3, pp. 307–327, 1986.

[100] C. P. Robert and G. Casella, **Monte Carlo statistical methods**, vol. 319. Citeseer, 2004.

[101] B. Efron, *Bootstrap methods: another look at the jackknife*, **The annals of Statistics**, pp. 1–26, 1979.

[102] D. N. Politis and J. P. Romano, *The stationary bootstrap*, **Journal of the American Statistical Association**, vol. 89, no. 428, pp. 1303–1313, 1994.

[103] H. R. Kunsch, *The jackknife and the bootstrap for general stationary observations*, **The Annals of Statistics**, vol. 17, no. 3, pp. 1217–1241, 1989.

[104] K. W. Hedman and G. B. Sheblé, *Comparing hedging methods for wind power: Using pumped storage hydro units vs. options purchasing* in **Probabilistic Methods Applied to Power Systems, 2006. PMAPS 2006. International Conference on**, pp. 1–6, IEEE, 2006.

[105] B. C. Ummels, M. Gibescu, E. Pelgrum, W. L. Kling, and A. J. Brand, *Impacts of wind power on thermal generation unit commitment and dispatch*, **Energy Conversion, IEEE Transactions on**, vol. 22, no. 1, pp. 44–51, 2007.

[106] M. Pereira and L. M. Pinto, *Multi-stage stochastic optimization applied to energy planning*, **Mathematical Programming**, vol. 52, no. 1-3, pp. 359–375, 1991.

[107] K. Koh, S.-J. Kim, and S. P. Boyd, *An interior-point method for large-scale l1-regularized logistic regression.*, **Journal of Machine learning research**, vol. 8, no. 8, pp. 1519–1555, 2007.

[108] S. Yun, P. Tseng, and K.-C. Toh, *A block coordinate gradient descent method for regularized convex separable optimization and covariance selection*, **Mathematical programming**, vol. 129, no. 2, pp. 331–355, 2011.