January 2013

Optimization in non-parametric survival analysis and climate change modeling

Iuliana Teodorescu
University of South Florida, alpinn@gmail.com

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the Mathematics Commons, and the Statistics and Probability Commons

Scholar Commons Citation
Teodorescu, Iuliana, "Optimization in non-parametric survival analysis and climate change modeling" (2013). Graduate Theses and Dissertations.
http://scholarcommons.usf.edu/etd/4782

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Optimization in Non-parametric Survival Analysis and Climate Change Modeling

by

Iuliana Teodorescu

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics & Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Chris Tsokos, Ph.D.
Chairman: Dmitry Goldgof, Ph.D
Marcus McWaters, Ph.D.
Les Skrzypek, Ph.D
Rebecca Wooten, Ph.D.

Date of Approval:
June 28, 2013

Keywords: Approximations, Non-parametric, Inference, Survival Analysis, Algorithmic Complexity

Copyright © 2013, Iuliana Teodorescu
Dedication

To my daughter Sonia, for her love and inspiration.
Acknowledgments

Professor Tsokos was more than just an adviser for this thesis: he motivated and inspired my research while also providing guidance in fundamental aspects of the path from the academic world to that of applied sciences and to inter-disciplinarity. I am grateful to him for all the help, advice, and assistance that he provided during my studies at USF.

I also wish to acknowledge professors Ladde, Wooten, and Ramachandran, for the constant support and advice during these four years, and for helping making my transfer from another university here completely seamless.

Professors McWaters and Skrzypek were extremely supportive as Ph.D. committee members, but they also helped me in many ways through the whole academic process.

While this doctoral research was finalized at USF, it represents the completion of a journey which I embarked upon as a graduate student at Northern Illinois University. Starting in a new field of study, in a new country, I benefited a lot from the personal example, support, and advice of professors Feltz and Polanski.

As a student at University of New Mexico, I was fortunate to learn directly from professor Christensen, an expert in linear models, accomplished teacher, and well-known author. I have many fond memories from his lectures and thoughtful commentaries.

This work could not have been accomplished without the support and understanding of my family: my daughter and my husband, my parents and my parents in-law. Their help and patience made all of this possible; without their faith in me I could not have persevered.
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Predicting large-scale climate perturbations in the Southern Pacific</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>The Dynamical Linear Model</td>
<td>38</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Theoretical considerations</td>
<td>38</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Data analysis</td>
<td>41</td>
</tr>
<tr>
<td>4.1.3</td>
<td>DLM formulation</td>
<td>41</td>
</tr>
<tr>
<td>4.1.4</td>
<td>The MCMC algorithm</td>
<td>43</td>
</tr>
<tr>
<td>4.2</td>
<td>Filtering, smoothing and forecasting</td>
<td>44</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Filtering recurrences and smoothing the data</td>
<td>44</td>
</tr>
<tr>
<td>4.2.2</td>
<td>$k$-step ahead forecasting</td>
<td>45</td>
</tr>
<tr>
<td>4.3</td>
<td>Interpretation of the results and conclusions</td>
<td>45</td>
</tr>
<tr>
<td>4.4</td>
<td>Using the generalized algorithm for prediction of La Niña</td>
<td>47</td>
</tr>
<tr>
<td>5</td>
<td>Estimating the atmospheric CO$_2$ levels - US data and model with interactions</td>
<td>48</td>
</tr>
<tr>
<td>5.1</td>
<td>Regression analysis and model building</td>
<td>48</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Second-order model: parameter determination and validation</td>
<td>48</td>
</tr>
<tr>
<td>5.2</td>
<td>Eigenvalue analysis of the second-order interactions matrix</td>
<td>50</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Canonical analysis of the quadratic model</td>
<td>52</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Confidence region shapes and conic sections</td>
<td>54</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Specific analysis for the atmospheric CO$_2$ quadratic model</td>
<td>56</td>
</tr>
<tr>
<td>5.3</td>
<td>Applications: predictions based on nonlinear analysis</td>
<td>57</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Order of magnitude analysis for the canonical variables</td>
<td>57</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Managing CO$_2$ emissions: accountability policies and metrics</td>
<td>58</td>
</tr>
<tr>
<td>5.4</td>
<td>Predicting large deviations in the CO$_2$ levels using Algorithm I</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>Estimating the atmospheric CO$_2$ levels - EU data and comparison to US model</td>
<td>61</td>
</tr>
<tr>
<td>6.1</td>
<td>Regression analysis and model building</td>
<td>63</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Validation of the fitted model</td>
<td>65</td>
</tr>
<tr>
<td>6.2</td>
<td>Comparing the US and EU models</td>
<td>65</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Specific features of CO$_2$ contributors for EU data</td>
<td>66</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Regionalization of atmospheric CO$_2$ analysis</td>
<td>68</td>
</tr>
<tr>
<td>7</td>
<td>Estimating the atmospheric CO$_2$ levels - EU model with interactions</td>
<td>69</td>
</tr>
<tr>
<td>7.1</td>
<td>The model, parameters, and descriptive quantities</td>
<td>69</td>
</tr>
<tr>
<td>7.1.1</td>
<td>Eigenvalue analysis of the second-order interactions matrix</td>
<td>70</td>
</tr>
<tr>
<td>7.1.2</td>
<td>Canonical analysis of the quadratic model</td>
<td>71</td>
</tr>
<tr>
<td>7.1.3</td>
<td>Confidence region shapes and conic sections</td>
<td>72</td>
</tr>
<tr>
<td>7.2</td>
<td>Conclusions and predictions based on nonlinear analysis</td>
<td>75</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Nonlinear analysis of contributing factors</td>
<td>75</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Relative importance of attributable variables</td>
<td>76</td>
</tr>
<tr>
<td>7.2.3</td>
<td>Comparing the US and EU models</td>
<td>77</td>
</tr>
<tr>
<td>8</td>
<td>Conclusions and perspectives</td>
<td>78</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>79</td>
</tr>
</tbody>
</table>
List of Tables

Table 1 Probability for sea-level pressure large deviations, 1998. ......................... 47
Table 2 Linear regression coefficients for attributable variables. ............................. 50
Table 3 Probability for atmospheric CO$_2$-level large deviations. ........................... 60
Table 4 Statistical evaluation criteria for model (6.1.1) ........................................... 64
Table 5 Ranking by statistical relevance for attributable variables and interactions. .... 65
Table 6 Comparison of statistical relevance for attributable variables and interactions, US vs. EU. .......................................................... 66
Table 7 Contributions by country to gas fuel emissions, for year 2008 (in thousand metric tons of carbon). ......................................................... 67
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Typical Kaplan-Meier estimate and time dependence.</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Sea level pressure, measured monthly (1882-1998, Darwin research station, Australia).</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>Sea level temperature, selected years between 1936 and 1981.</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>Small (bottom), large (middle), and extreme (top) deviations for sample means.</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>The atomic distribution as limit of a sequence of Gaussian distributions.</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>Comparing the Kaplan-Meier and large-deviations estimator to the exact formula.</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>Normal air and water circulation patterns in Equatorial Pacific.</td>
<td>38</td>
</tr>
<tr>
<td>8</td>
<td>Temperature maps for El Niño and La Niña episodes.</td>
<td>39</td>
</tr>
<tr>
<td>9</td>
<td>The 1400 sea level pressure data points.</td>
<td>42</td>
</tr>
<tr>
<td>10</td>
<td>Periodogram of pressure data.</td>
<td>43</td>
</tr>
<tr>
<td>11</td>
<td>Predicted values for 12 consecutive months, ten years into the future.</td>
<td>46</td>
</tr>
<tr>
<td>12</td>
<td>Atmospheric CO₂ data taken at the Mauna Loa station (first panel) and its corresponding $Q - Q$ plot (second panel). Data obtained from CDIAC-ORNL.</td>
<td>49</td>
</tr>
<tr>
<td>13</td>
<td>Confidence regions for the elliptical case.</td>
<td>54</td>
</tr>
<tr>
<td>14</td>
<td>Confidence regions for the hyperbolic case.</td>
<td>55</td>
</tr>
<tr>
<td>15</td>
<td>Confidence regions for the degenerate (flatness) case.</td>
<td>57</td>
</tr>
<tr>
<td>16</td>
<td>Schematic representation of the carbon cycle (DOE report [47] and supporting documentation).</td>
<td>62</td>
</tr>
<tr>
<td>17</td>
<td>EU CO₂ emissions, in thousands of metric tons [50].</td>
<td>62</td>
</tr>
</tbody>
</table>
Abstract

Many of the open problems of current interest in probability and statistics involve complicated data sets that do not satisfy the strong assumptions of being independent and identically distributed. Often, the samples are known only empirically, and making assumptions about underlying parametric distributions is not warranted by the insufficient information available. Under such circumstances, the usual Fisher or parametric Bayes approaches cannot be used to model the data or make predictions.

However, this situation is quite often encountered in some of the main challenges facing statistical, data-driven studies of climate change, clinical studies, or financial markets, to name a few.

We propose a novel approach, based on large deviations theory, convex optimization, and recent results on surrogate loss functions for classifier-type problems, that can be used in order to estimate the probability of large deviations for complicated data. This may include, for instance, high-dimensional data, highly-correlated data, or very sparse data.

The thesis introduces the new approach, reviews the current known theoretical results, and then presents a number of numerical explorations meant to quantify how far the approximation of survival functions via large deviations principle can be taken, once we leave the limitations imposed by the existing theoretical results.

The explorations are encouraging, indicating that indeed the new approximation scheme may be very efficient and can be used under much more general conditions than those warranted by the current theoretical thresholds.

After applying the new methodology to two important contemporary problems (atmospheric CO₂ data and El Niño/La Niña phenomena), we conclude with a summary outline of possible further research.
1.1 Introduction

There is a wide consensus in the scientific community and beyond, that the 21st century will be the century of “big data” [1]: according to some estimates, in 2011 the amount of new data generated exceeded the total amount of knowledge accumulated throughout history; also, data generated yearly keeps increasing by about 40-50%. Yet, the real challenge lies not so much in the sheer amount of information being generated and manipulated, but in the task of quantifying and analyzing it. In particular, statistical analysis of complex, complicated data, is at the core of some of the main challenges of science today:

1) climate studies related either to climate change research, or to forecasting models for large spatial and temporal scales of atmospheric and ocean dynamics;

2) clinical studies in medicine, molecular biology, and related areas;

3) large-scale simulations of complex systems, from economic trade to social networks.

In each of the situations described above, there is a need to quantify, describe, analyze, and interpret the data using statistical methods. Specifically, one of the most pressing questions to address in all these cases is estimating the probability of a “large deviation”, whether it represents a rapid increase in the atmospheric temperature, a genetic mutation, or the collapse of the national power grid. Broadly defined, this problem belongs to survival analysis. Therefore, we can say that one of the main problems facing statisticians today is that of performing survival analysis in the context of complicated, incomplete, or large-dimensional data; the trait that these types of data have in common is that typical methods of statistical inference (based on Central Limit Theorem, or parametric Bayesian, for example) are not applicable. As a result, the challenge consists not only
in analyzing complicated data, but also in identifying new inference methods, both analytically and numerically efficient.

This thesis contains such a study, and consists of several distinct parts:

• in Chapter 2, we formulate the general problem of survival analysis for “real data”, i.e. for the cases in which textbook methods are not applicable; after reviewing the traditional methods and their variations, the basic results of Large Deviations Theory (LDT) are discussed, from the introduction of the Cramér functional to Varadhan’s Lemma. Several examples are discussed, and the chapter concludes with a general formulation of the analytical problem, along with the limitations of the exact methods;

• in the next chapter, a survey of main ideas and results from approximation theory and estimation of empirical data is presented; methods like kernel estimation for empirical p.d.f., measures of convergence of approximations, optimal choices for basis of approximating functions, and comparisons between these. The important property of convexity is singled out as most relevant for the central problem of this thesis, and with it, the main ideas behind convex optimization. We then implement recent results from adaptive learning, filtering, and optimization with surrogate functions, are discussed. Based on these advances, a specific implementation of the Block-Partitioning Algorithm (BPA) is proposed in the context of optimal approximation of the LDT rate function. The algorithm is then implemented, tested, and compared according to standard procedures for numerical approximation techniques;

• in the last four chapters, a series of important problems real data are investigated and addresses. The new methodology developed in the previous chapters is applied, along with other statistical procedures, such as:

  – modeling the atmospheric CO$_2$ from anthropogenic sources and estimating the probability of large yearly increases in the CO$_2$ level;

  – modeling the ocean-level temperature and pressure data in the Southern Pacific system (part of the Southern Oscillation/El Niño - La Niña problem).

• the final discussion touches on other possible applications of the new method and relevant aspects of numerical approximation schemes.
Chapter 2
Statistical inference and large deviations theory: from exact results to approximate algorithms

In this chapter, we formulate the general problem of survival analysis for “real data”, i.e. for the cases in which textbook methods are not applicable; after reviewing the traditional methods and their variations, the basic results of Large Deviations Theory (LDT) are discussed, from the introduction of the Cramér functional to the Gärtner-Ellis theorem. Several examples are discussed, and the chapter concludes with a general formulation of the analytical problem, along with the limitations of the exact methods.

2.1 The basic concepts of survival analysis

2.1.1 Survival function

In order to define the basic concepts of survival analysis, we consider a random variable $T$, which represents “lifetime” and can be taken to be real and positive, $T \geq 0$. If its probability density function (p.d.f.) and cumulative density function (c.d.f.) are $f_T(t)$ and $F_T(t)$, respectively, so that

$$f_T(t) = F_T'(t) = \frac{d}{dt} F_T(t)$$

and

$$F_T(t) = P(T \leq t),$$

then the basic goal of survival analysis is to find the survival function, conventionally denoted by $S(t)$ (with $t$ signifying “time”), which is defined as

$$S_T(t) := P(T > t),$$

and can interpreted as “the probability to survive at least until time $t$”.

3
The survival function can be also expressed in terms of distribution and density functions as:

\[ S_T(t) = P(T > t) = \int_t^\infty f_T(u) \, du = 1 - F_T(t). \]  

(2.1.1)

We also obtain the relation

\[ S'_T(t) = -F'_T(t) = -f_T(t). \]  

(2.1.2)

2.1.2 Hazard function and cumulative hazard function

The “hazard function”, \( \lambda \), represents the “event” rate at time \( t \), conditional on survival at least until time \( t \) (by “event” meaning a hazard which causes the lifetime \( T \) to end at \( t + dt \)):

\[ \lambda(t) \, dt = P(t \leq T < t + dt \mid T \geq t) = \frac{P(t \leq T \leq t + dt)}{P(T \geq t)} = \frac{f(t) \, dt}{S(t)}. \]

From Eq. (2.1.2), we have

\[ \lambda(t) = -\frac{S'(t)}{S(t)}. \]  

(2.1.3)

Integrating this equation we obtain

\[ S(t) = \exp \left[ -\int_0^t \lambda(u) \, du \right] = \exp(-\Lambda(t)), \]

where the “cumulative hazard function” \( \Lambda(t) = -\log S(t) \) represents the “accumulation” of the hazard over time:

\[ \Lambda(t) = \int_0^t \lambda(u) \, du. \]  

(2.1.4)

We conclude that knowing the cumulative hazard function solves the basic problem of survival analysis, and provides the answer in the form:

\[ S(t) = P(T \geq t) = \exp(-\Lambda(t)). \]  

(2.1.5)

Since from (2.1.4) it is clear that \( \Lambda(t) \geq 0 \) is an increasing function of time, Eq. (2.1.5) gives the survival function as an exponential, so that if the hazard function is a constant (\( \lambda = \text{constant “risk”} \)), then the probability to survive decays exponentially, etc. We consider this in the example below.

**Example 1** If \( T \sim \text{Exp}(k) \), then \( f_T(t) = \frac{1}{k} e^{-t/k} \), \( F_T(t) = 1 - e^{-t/k} \), \( S(t) = e^{-t/k} \), so indeed

\[ S(t) = e^{-\Lambda(t)}, \quad \Lambda(t) = \frac{t}{k}, \quad \lambda(t) = \frac{1}{k} = \text{const}. \]
2.1.3 Survival analysis for real data

If we know the exact distribution of lifetime $T$, we can then use the formulas above to compute, either exactly or approximately, the quantities defined. However, in reality we cannot assume that we are given this analytical information; instead, we must assume that we only have a finite sample of $n$ observations $T_1, T_2, \ldots, T_n$, and attempt to find quantities such as

$$P \left( \sum_{k=1}^{n} T_k \geq n \cdot t \right), \quad P(T_{n+1} \geq t),$$

in various circumstances, identified in Discussion 1. Before proceeding with the data types classification, we mention here the most widely used (especially by practitioners in medical sciences) survival function statistic, known as the Kaplan-Meier statistic, $\hat{S}(t)$. It is defined [2] under the assumptions that the data are i.i.d. r.v. with unknown distribution, as follows: for a given sequence of ordered observation times $t_1 < t_2 < \ldots < t_n$, compute the number of elements remaining in the sample $n_k$ (how many patients are still being observed at time $t_k$) and the number of deaths occurred up to time $t_k$, $d_k$. Then the Kaplan-Meier (K-M) estimate at time $t$ is the maximum-likelihood estimator

$$\hat{S}(t) = \prod_{t_k \leq t} \frac{n_k - d_k}{n_k}. \tag{2.1.6}$$

A typical plot of the estimate as a function of time is shown in Figure 1. The vertical marks (or “ticks”) seen on the horizontal segments indicate the time of an event (i.e. death, or patient lost to censoring) occurring during that interval.

The usefulness and relevance of the Kaplan-Meier statistic is questioned by many scientists, both from a theoretical and practical point of view [3]-[6]. It is easy to notice that the estimate does not change if the original time sequence $t_1 < t_2 < \ldots < t_n$ is modified so that the “ticks” remain in the same order relative to the new times $t'_1 < t'_2 < \ldots < t'_n$. This makes little sense, however, since by bringing one of the new times $t'_k$ arbitrarily close to a subsequent event changes dramatically the survival time for that particular patient (whose own “survival time” becomes arbitrarily short), while the estimator $\hat{S}(t)$ remains the same. In particular, for sparse data (for which the number of events per interval is small, e.g. $d_k = k$), applying the time redefinition discussed here will give arbitrarily short individual survival times for each patient, which however will not be captured by $\hat{S}(t)$. 

5
These features were discussed in various proposals for improving the effectiveness of the estimator [3]-[6], including the attempt to define the “patient time” in a meaningful way. However, even the improved variants of the K-M estimate required a priori a sample of i.i.d. r.v. in order to have the property of being the maximum-likelihood estimator. Since this requirement is not satisfied in many situations, we find it necessary to discuss first a possible classification of data, from the point of view of its structure, size, and complexity.

**Discussion 1** The following is a classification of inference problems based on data features:

i) The sample is \( n \gg 1 \), \( T_1, T_2, \ldots, T_n \) are independent and identically distributed (i.i.d.) with the same distribution function, which is unknown (large sample inference problem [7]);

ii) The sample is not large, \( T_1, T_2, \ldots, T_n \) are independent and identically distributed (i.i.d.) with the same distribution function from a known parametric family \( f_T(t|\theta) \) (parametric Bayesian inference);

iii) the various assumptions from the previous two points are not valid, so that \( T_1, T_2, \ldots, T_n \) may be dependent, not identically distributed, and their distribution functions are not known.
For realistic situations, the case to consider is that of Discussion 1, (iii), for which almost no analytical methods exist. Making the simplifying assumptions of cases (i) and (ii) allows us to apply powerful methods of either Fisher or Bayes-type inferences, yet even in such situations there are significant complications. In order to illustrate these distinctions, we summarize in the next section the existing methods which can be used to perform survival analysis in the case of asymptotic theory, Discussion 1, (i), namely Generalized Extreme Value (GEV) theory. As we will see, these methods (together with the standard Central Limit Theorem-type of results) leave out an important class of questions, for which it is necessary to use another approach, given by the Large Deviations Principle (LDP). The latter is reviewed in the last section of this chapter, which concludes with the main problem studied in this thesis.

**EXAMPLE 2** To make this discussion very concrete, we choose the case of monthly measurements of air pressure and temperature, taken at sea level, between 1882 and 1998, at the Darwin research station, in Northern Territory, Australia. These measurements are an important part of a global monitoring system, which includes measuring pressure and air temperature across the Pacific, from Darwin in the West to Galapagos Islands in the East. The normal weather patterns in this area were studied and described systematically for the first time by Sir Gilbert Thomas Walker [51, 52]. The air and water patterns he discovered are called the **Walker circulation**. Anomalies (deviations from this pattern) include the phenomena known as El Niño and La Niña, which will be described in detail in Chapter 5. They are large-scale fluctuations, with global consequences, reaching all the way to continental USA. The pressure data is shown in Figure 2.

The temperature measurements are illustrated through a plot of selected data for the period from 1936 to 1981, in Figure 3. This type of plot is representative for similar measurements on climate performed at many other locations in the world. It shows the sea-level temperature, measured once per month, and represented as stacked yearly plots: one color shows measurements from January (J) to December (D) (averaged over five consecutive years), then the next year starts all over, in another color, etc. In this picture, a few select years are shown: from 1936 to 1981, every five years (so 1936, 1941, etc). The second column gives the years that will be compared to the ones shown (at 30 years difference): 1941 with 1970, 1946 with 1975, etc.

Even just by looking at the graphs in Figures 2, 3 it is rather easy to observe that the assumptions of Discussion 1, (i), do not apply. In fact, these observations are certainly not independent, because
Figure 2.: Sea level pressure, measured monthly (1882 -1998, Darwin research station, Australia).

there is strong seasonality in monthly measurements. They are also not necessarily identically distributed, because of a global trend which may be linked to overall climate changes. Finally, while \( n = 1400 \) may sometimes be considered a reasonably large sample, in this case it is not; this is because the global heating trend which is known to occur from independent experiments has typical time scales at the level of hundreds of years (perhaps even larger), with respect to which 100 years worth of data is a very short time, so the sample is actually quite small.

As we will see in Chapter 5, the assumptions of Discussion 1, (ii) (Bayesian-type inference), are more appropriate here, and the data can be studied with methods from Dynamical Linear Models (DLM). However, this approach has its own limitations (discussed in Chapter 5), so we are once again led to the need for alternative approaches (LDP).
In order to distinguish the limitations of large-sample inference for i.i.d. r.v.'s, we briefly review the available methods for describing large fluctuations, in the next section.

2.2 Distributions of extreme values in large samples of i.i.d.r.v.

Suppose we have an i.i.d. sequence of random variables, \(X_1, X_2, \ldots\), whose cumulative distribution function (c.d.f.) is \(F(x)\):

\[
F(x) = P(X_k \leq x), \quad k = 1, 2, \ldots
\]

Denote by \(\min(X_1, \ldots, X_n) = M_1 \leq M_2 \leq \ldots \leq M_n = \max(X_1, \ldots, X_n)\) the ordered sample, so that \(M_n\) is \(n^{th}\) sample maximum. Then from the usual formula for order statistics for the p.d.f. of \(M_k\) [7], we obtain for the c.d.f. of \(M_n\):

\[
P(M_n \leq x) = F^n(x) \quad \text{(2.2.1)}
\]
Formula (2.2.1) is not directly applicable in the large sample limit, since it simply says that for any value \( x \) for which \( F(x) < 1 \), we have \( P(M_n \leq x) \to 0 \). To obtain non-trivial limit results we must rescale in some sense: assume that we have two sequences \( \{a_n > 0\}_{n=1}^{\infty}, \{b_n\}_{n=1}^{\infty} \) such that:

\[
P\left[ M_n - \frac{b_n}{a_n} \leq x \right] = P(M_n \leq a_n x + b_n) = F^n(a_n x + b_n) \to H(x), \tag{2.2.2}
\]

where convergence is understood pointwise.

### 2.2.1 Generalized extreme value limit distributions

The form of limit distributions, originally stated in Ref. [8], and later derived rigorously in Ref. [9], asserts that if a non-degenerate (i.e. not a singleton) limit function \( H \) exists, then it must be one of the types nowadays called Gumbel, Fréchet, and Weibull. These types may be combined into a single Generalised Extreme Value (GEV) distribution:

\[
H(x) = \exp \left[ - \left( 1 + \frac{x - \mu}{\sigma} \right)^{-1/\nu} \right], \tag{2.2.3}
\]

where by definition \( (y)_+ = \max(y, 0) \), and \( \mu \) is the location parameter, \( \sigma > 0 \) is the scale parameter, and \( \nu \) is the shape parameter. Moreover, when considering the large sample limit for the conditional probability for exceeding some high threshold \( u \) (so that \( Y = X - u > 0 \)):

\[
F_u(y) = P(Y \leq y | Y > 0) = \frac{F(y + u) - F(u)}{1 - F(u)},
\]

we may find a limit (as \( n \to \infty, u \to \sup \{x : F(x) < 1\} \)):

\[
F_u(y) \to 1 - \left( 1 + \nu \frac{y}{\sigma u} \right)^{-1/\nu} \equiv G(y, \sigma u, \nu), \tag{2.2.4}
\]

where \( G(y, \sigma u, \nu) \) is the Generalized Pareto Distribution (GPD), c.f. Refs. [11, 12].

### 2.2.2 Meaning of inherited parameters and analytical properties

The three parameters \( \mu, \sigma, \nu \) determine the analytical properties of the limiting distribution (both for the GEV and GPD functions) in a number of ways discussed below.

- Analytical properties controlled by the shape parameter:
i) $\nu > 0$ is the “long-tailed” case for which $1 - H(x) \sim x^{-1/\nu}$ as $x \to \infty$ (and corresponds to the Fréchet distribution with $\alpha = 1/\nu$);

ii) $\nu = 0$ is the “medium-tailed” case for which $1 - H(x)$ decreases exponentially for $x \to \infty$ (and corresponds to the Gumbel distribution);

iii) $\nu < 0$ is the “short-tailed” case, in which the distribution has a finite endpoint (the minimum value of $x$ for which $H(x) = 1$ at $x = \mu - \sigma / \nu$); corresponds to the Weibull distribution.

- Analytical properties controlled by the scale and location parameters:

1) **Support** The domain of definition for the GEV variable with parameters $(\mu, \sigma, \nu)$ is:
   
   - $x \in [\mu - \sigma / \nu, \infty)$ for $\nu > 0$,
   - $x \in (-\infty, \infty)$ for $\nu = 0$,
   - $x \in (-\infty, \mu - \sigma / \nu]$ for $\nu < 0$

2) **P.D.F.**
   
   $$f_{\mu,\sigma,\nu}(x) = \frac{t^{\nu+1}(x)}{\sigma} e^{-t(x)},$$
   
   where
   
   $$t(x) = 1 + \nu \frac{x - \mu^{-1/\nu}}{\sigma} \quad \text{if } \nu \neq 0, \quad t(x) = e^{-(x-\mu)/\sigma} \quad \text{if } \nu = 0.$$

3) **Mean**
   
   $$\mathbb{E}(X) = \mu - \frac{\sigma}{\nu} + \frac{\sigma}{\nu} \Gamma(1 - \nu)$$

4) **Variance**
   
   $$V(X) = \frac{\sigma^2}{\nu^2} \left[ \Gamma(1 - 2\nu) - \Gamma^2(1 - \nu) \right],$$
   
   where $\Gamma(t)$ is Euler’s Gamma function, $\Gamma(t) \equiv \int_0^\infty z^{t+1} e^{-z} dz$.

As indicated in Section 2.2.2, we can specialize the parameters which govern the GEV and GPD distributions and obtain the (classical) cases known as Gumbel, Fréchet, Weibull, and Pareto. Specifically, they correspond to:
Gumbel distribution (type I extreme values distribution)

Setting $\nu = 0$, we obtain the Gumbel distribution with support on $\mathbb{R}$ and c.d.f.

$$F_I(x; \mu, \sigma) = \exp[-e^{-(x-\mu)/\sigma}] \quad (2.2.5)$$

This is because the limit

$$\lim_{\nu \to 0} \left(1 + \nu \frac{x - \mu}{\sigma}\right)^{-1/\nu} = \lim_{\nu \to -\infty} \left(1 + \frac{1}{\nu x - \mu/\sigma}\right)^\nu = \frac{1}{e^{(x-\mu)/\sigma}} = e^{(x-\mu)/\sigma}$$

follows from Euler’s formula $(1 + a/n)^n \to e^a, n \to \infty$. Using this formula in (2.2.3) leads us to (2.2.5).

Fréchet distribution (type II extreme values distribution)

For $\nu > 0$, (2.2.3) gives the Fréchet class of distributions, for which (setting $\alpha = 1/\nu$), we have the c.d.f. with support $x \in [\tilde{\mu}, \infty)$:

$$F_{II}(x; \tilde{\mu}, \sigma, \alpha) = \exp \left[ - \left( \frac{x - \tilde{\mu}}{\sigma} \right)^{-\alpha} \right], \quad (2.2.6)$$

where $\tilde{\mu} = \mu - \alpha \sigma$ is the new location parameter. Eq. (2.2.6) is simply (2.2.3) with respect to the new location parameter. Clearly,

$$\lim_{x \to \tilde{\mu}^+} F_{II}(x; \tilde{\mu}, \sigma, \alpha) = \lim_{\epsilon \to 0^+} e^{-\epsilon^{-\alpha}} = e^{-\infty} = 0,$$

so $F_{II}$ is continuous and identically zero to the left of the domain of definition.

Weibull distribution (type III extreme values distribution)

For $\nu < 0$, we set again $0 < \alpha = -1/\nu$ and $\tilde{\mu} = \mu - \sigma/\nu$ and obtain the (reverse) Weibull distribution with support $x \in (-\infty, \tilde{\mu}]$ and c.d.f.

$$F_{III}(x; \tilde{\mu}, \sigma, \alpha) = \exp \left[ - \left( \frac{x - \tilde{\mu}}{\sigma} \right)^\alpha \right] \quad (2.2.7)$$

As in the previous case,

$$\lim_{x \to \tilde{\mu}^-} F_{III}(x; \tilde{\mu}, \sigma, \alpha) = \lim_{\epsilon \to 0^+} e^{-\epsilon^{\alpha}} = e^0 = 1,$$

so $F_{III}$ is continuous and equal to 1 to the right of the domain of definition.
Pareto distribution (power-law extreme values)

For the case of GDP, we can again consider that $\alpha = 1/\nu > 0$, in which case we obtain the usual Pareto (power-)law (with the shifted location parameter):

$$F_{IV}(x; \tilde{\mu}, \sigma, \alpha) = 1 - c_\alpha \left( \frac{x - \tilde{\mu}}{\sigma} \right)^{-\alpha}, \quad x \in [\tilde{\mu}, \infty) \quad (2.2.8)$$

and $c_\alpha$ a normalization factor. The case $\alpha \to \infty (\nu \to 0^+)$ degenerates into the usual (scaled and shifted) exponential distribution, by application of Euler’s formula:

$$\lim_{\nu \to 0^+} 1 - \left( 1 + \nu \frac{x - \mu}{\sigma} \right)^{-1/\nu} = 1 - \exp \left( -\frac{x - \mu}{\sigma} \right),$$

which can, of course, also be obtained as a limiting case of the Weibull family of distributions (by taking $\alpha = 1$ in (2.2.7) and changing the sign of the argument of the functions).

Discussion 2

The multi-parameter family of GEV covered by Section 2.2 may appear to be rich enough in order to model efficiently any type of large fluctuations considered in survival analysis. Indeed, the GEV range from the power-law decay of Pareto distributions (slowly-decaying, “long tails”) to the super-exponential decay (faster than Gaussian, extremely “short tails”) of the Weibull distributions. This versatility makes GEV very useful when modeling truly extreme values. However, this approach is not applicable in many cases, when the object of interest is a large deviations which is not extreme yet. This situation arises in the context of a large sample $n \gg 1$ of i.i.d.r.v., when estimating the probability that the sample mean $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$ deviates from the average $\mu = \mathbb{E}(X_k)$ by an amount of order $O(1)$, as $n \to \infty$, i.e. finding $P(\overline{X} > \mu + x)$. This case is clearly not covered by Central Limit Theorem, which provides estimates only for small deviations from $\mu$, by

$$P \left( \overline{X} > \mu + z_\alpha \frac{\sigma}{\sqrt{n}} \right) \to \alpha, \quad n \to \infty.$$ 

The GEV approach is also not applicable in this case, since the probability of large deviation $P(\overline{X} > \mu + x)$ is dominated by the behavior of the bulk of the sample $\{X_k\}$, and not by its extreme values. A specific example for this fact is given below.

Example 3 For $X_k \sim \text{Exp} \ (1), k = 1, 2, \ldots, n$, with p.d.f. and c.d.f. given by

$$f_X(x) = e^{-x}, \quad F_X(x) = 1 - e^{-x}, \quad \mathbb{E}(X) = 1, \quad V(X) = 1,$$

we have the following types of results in the large sample-limit ($n \gg 1$):
• Small deviations around the mean:

\[ \sqrt{n}(\bar{X} - 1) \overset{d}{\to} N(0, 1) \Rightarrow P \left( \frac{\bar{X} > 1 + \frac{z\alpha}{\sqrt{n}}}{\sqrt{n}} \right) \to \alpha, \quad n \to \infty. \] (2.2.9)

• Extreme value distribution:

\[ P(M_n \leq x) = F_n(x) \Rightarrow P(M_n > x) = 1 - (1 - e^{-x})^n \simeq ne^{-x}. \] (2.2.10)

• Large deviations of the sample mean (see Example 5 and Eq. (2.3.1) in the next section):

\[ P(\bar{X} > x) \sim x^n e^{n(1-x)}. \] (2.2.11)

Eqs. (2.2.9)-(2.2.11) already indicate that these three different regimes (small, extreme, and large deviations) are significantly different. To emphasize this, we plot the three estimates for \( n = 50 \) and \( x \in [1.1, 1.5] \) (i.e. for \( x \in [1.1\mu, 1.5\mu] \)) in Figure 4. The different behavior of the three types of “tails” for these regimes is clearly displayed.

As illustrated in Example 3, in order to find an efficient way to approximate the probability for large deviations of the sample mean, we cannot rely directly on CLT or GEV approaches. Moreover, when addressing realistic problems dealing with samples of variables that may be dependent and not identically distributed, neither of these two methods offer a solid starting point for numerical approximations. It is therefore necessary to consider specific theoretic results for dealing with large deviations, for the sample mean and beyond. This is the purpose of the specialized literature survey presented in Section 2.3. It summarizes the known analytical results in this area, starting from the simplest situation (samples of i.i.d. r.v. with known distribution functions) to the most complicated (LDP for samples of non-i.i.d.). The section closes with the formulation of the main problem investigated in this thesis.

### 2.3 The Large Deviations Principle

#### 2.3.1 Large Deviations Principle for i.i.d.r.v.

**Definition** 2.3.1 The Cramér functional (or “rate function”) for large deviations of the sample mean for a random variable \( X \) is defined as

\[ I(x) := \max_{t > 0} \left\{ tx - \ln(m_X(t)) \right\}, \]
Figure 4.: Small (bottom), large (middle), and extreme (top) deviations for sample means.

where $m_X(t)$ is the moment-generating function of $X$,

$$m_X(t) = \mathbb{E}(e^{tX}).$$

The rate function $I(x)$ is also called the Legendre-Fenchel transform of $\ln m_X(t)$.

**Example 4** For $X \sim N(\mu, 1)$, we have

$$m_X(t) = e^{t\mu + t^2/2},$$

so

$$I(x) = \max_{t>0} [tx - t\mu - t^2/2]$$

Find the maximum in the exponent by taking the first derivative:

$$\frac{d}{dt} [tx - t\mu - t^2/2] = 0 \Rightarrow x - \mu - t = 0 \Rightarrow t_m = x - \mu$$
Therefore,
\[ I(x) = t_m x - t_m \mu - t_m^2 / 2 = (x - \mu)^2 / 2. \]

**Example 5** We can similarly compute the Cramér functional for some common distributions:

- \( X \sim \text{Exp}(k) \), then
  \[ I(x) = \frac{x}{k} - 1 - \ln \frac{x}{k}, \quad x \geq 0 \]  \hspace{1cm} (2.3.1)

- \( X \sim \text{Poisson}(k) \), then
  \[ I(x) = k - x + x \log \frac{x}{k}, \quad x \geq 0. \]  \hspace{1cm} (2.3.2)

- \( X \sim \mathcal{N}(\mu, \sigma^2) \), then
  \[ I(x) = \frac{(x - \mu)^2}{2\sigma^2}, \quad x \in \mathbb{R} \]  \hspace{1cm} (2.3.3)

- \( X \sim \text{Bernoulli}(k) \), then
  \[ I(x) = x \ln \frac{x}{k} + (1 - x) \ln \frac{1 - x}{1 - k}, \quad x \in [0, 1]. \]  \hspace{1cm} (2.3.4)

**Theorem 2.1** (Large Deviations Principle) *The Large Deviations Principle (due to Cramér, Sanov, Gärtner, Ellis, Varadhan) states that the probability of “large deviations”*

\[ P(S_n \geq x) := P \left( \frac{1}{n} \sum_{k=1}^{n} X_k \geq x \right) \sim e^{-nI(x)}, \]

*where \( X_1, X_2, \ldots, X_n \) is a sample of i.i.d. r.v., and \( I(x) \) is the rate function defined earlier, and “\( \sim \)” means that the probability is determined only up to an overall normalization factor.*

**Remark 1** It is useful to note a few observations regarding Theorem 2.1 and the related results which will follow:

i) The result is known as a “principle” in the sense that it may hold under more general conditions than those indicated here, namely a similar estimate for large deviations exists even in the case of non-independent, not-identically distributed variables. We will review those generalizations in the next subsections.

ii) The original result is due to Cramér, who developed this method while working as an applied statistician in the insurance industry. From that point of view, the goal of the insurer is to minimize (and estimate) the risk of having a large number of insured submitting claims at the same time; similarly, this is the risk of an investment group to have a large percentage of shareholders sell their shares.
2.3.2 Hazard function from large deviations

By comparing Theorem 2.1 and Eq. (2.1.5), we conclude that

\[ \Lambda_n(t) \rightarrow nI(t), \quad n \rightarrow \infty, \]  

(2.3.5)

where \( \Lambda_n(t) \) is the cumulative hazard function for the average of a sample of size \( n \). Therefore, we can apply large deviations principle to survival analysis and obtain precise estimates for the cumulative hazard function, whenever we have an efficient method for computing the rate function.

Discussion 3 It is important to note that a direct application of large deviations theory to survival analysis (as discussed in this section) appears to be severely limited by the fact that, in order to compute the rate function from (2.3.1), we need the m.g.f. \( m_X(t) \), i.e. we need to know the distribution of \( X_k \)'s, which is precisely the information we don’t have in realistic situations. Luckily, the applicability turns out to be much more robust, and is preserved even under much more general (relaxed) conditions. The usefulness of the large deviations theory is due to the fact that we can make use of (2.3.5) even when the rate function \( I(x) \) is not the exact one (which would require knowing the p.d.f. \( f_X(x) \)), but an optimal approximation, for a given sample. Therefore, we can already formulate the main goal of this study as: finding an efficient approximation method for estimating the rate function from a given sample (of variables which are not necessarily i.i.d.), and using it to compute the probability of large deviations, survival, etc., from (2.3.5). Before moving on to specific approximation methods for this goal (in the next chapter), it is helpful to conclude this review of LDP with the most general theoretical results, which cover the cases of samples of non-i.i.d. variables.

2.3.3 The Gärtner-Ellis theorem

Theorem 2.1 expresses the LDP for sample means of i.i.d. r.v. It turns out that, in fact, it can generalized to cover the case where the sequence \( S_n \) is an arbitrary r.v. sample mean, not necessarily an i.i.d. sample mean. This is known as the Gärtner-Ellis theorem (GE), and takes the form

\[ \lim_{n \to \infty} \frac{1}{n} \ln P_{S_n}(S_n \geq x) = -I(x), \]  

(2.3.6)

where the rate function \( I(x) \) is given by

\[ I(x) = \sup_{t \in \mathbb{R}} \{ tx - \lambda(t) \}, \]  

(2.3.7)
and

\[ \lambda(t) = \lim_{n \to \infty} \frac{1}{n} \ln \mathbb{E}[e^{tS_n}], \quad (2.3.8) \]
called the scaled cumulant generating function (SCGF, a generalization of the notion of moment-generating function).

### 2.3.4 Sanov’s theorem (i.i.d. sample means)

Consider again the sample mean

\[ S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad (2.3.9) \]
corresponding to \( n \) real i.i.d. variables \( X_1, \ldots, X_n \) with pdf \( p(x) \). We follow the GE theorem and calculate the SCGF \( \lambda(t) \) defined in Eq. (2.3.8). Because of the i.i.d. property,

\[ \lambda(t) = \ln \mathbb{E}[e^{tX}]. \quad (2.3.10) \]

We apply this to find the rate function of a special sample mean defined by

\[ L_{n,j} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i,j} \quad (2.3.11) \]
for a sequence \( X_1, \ldots, X_n \) of \( n \) discrete i.i.d. r.v.s with finite state space \( \mathcal{X} = \{1, 2, \ldots, q\} \). This sample mean is called the empirical distribution of the \( X_i \)'s for a given realization of \( X_1, \ldots, X_n \). This number is normalized by the total number \( n \) of r.v.s, so what we have is the empirical frequency for the appearance of the symbol \( j \) in realizations of \( X_1, \ldots, X_n \). The numbers \( L_{n,j} \) for all \( j \in \mathcal{X} \) make up a vector \( L_n \) called the empirical vector.

To find the rate function associated with the random vector \( L_n \), we apply the GE Theorem but adapt it to the case of random vectors and obtain the rate function

\[ I(\mu) = \sum_{j \in \mathcal{X}} \mu_j \ln \frac{\mu_j}{P_j}. \quad (2.3.12) \]

This rate function is called the relative entropy or Kullback-Leibler divergence [20]. The full LDP for \( L_n \) is referred to as Sanov’s theorem [37].
2.3.5 Markov chains

**Definition 2.3.2** Let \( n \) and \( d \) be positive integers, \( n \geq 1 \) and \( d \geq 1 \), and define \( \Omega = \{1, \ldots, d\} \).

Consider a sequence of r.v.s \( \{X_1, X_2, \ldots, X_n\} \) such that

\[
P_{ij} = P(X_{k+1} = j | X_k = i)
\]

(2.3.13)

is independent of \( k \) for all \( i \) and \( j \) in \( \Omega \). Then the sequence \( \{X_1, X_2, \ldots, X_n\} \) is a Markov chain with state space \( \Omega \) and transition probabilities \( P_{ij} \) for \( i \) and \( j \) in \( \Omega \).

It follows from this definition that a Markov chain with known probability distribution of the initial state is completely characterized by a \( d \times d \) matrix containing the transition probabilities \( P_{ij} \),

\[
P = \begin{bmatrix}
P_{11} & P_{12} & \cdots & P_{1d} \\
P_{21} & P_{22} & \cdots & P_{2d} \\
\vdots & \vdots & \ddots & \vdots \\
P_{d1} & P_{d2} & \cdots & P_{dd}
\end{bmatrix}.
\]

This matrix is called the **transition probability matrix**. For a Markov chain with known transition probability matrix, the most likely state as \( n \to \infty \) can be calculated as follows. Define a vector \( V_k \) so that the \( i \)th element of \( V_k \) is the unconditional probability that the Markov chain is in state \( i \) at time \( k \). Hence, \( (V_k)_i = P(X_k = i) \), where \( V'_k = [(V_k)_1, \ldots, (V_k)_d] \).

The limiting, or steady state, probabilities, if they exist, are then given by

\[
\Pi^{(i)} = \lim_{n \to \infty} [(P')^n] \cdot V^{(i)}_1.
\]

(2.3.14)

Since \( [\Pi^{(i)}]_j = \sum_{k=1}^{d} \lim_{n \to \infty} [(P')^n]_{jk} \delta_{ik} = \lim_{n \to \infty} [(P')^n]_{ji} \), it follows that \( [\Pi^{(i)}]' = [\Pi^{(i)}_1, \ldots, \Pi^{(i)}_d] \) is the \( i \)th row of \( P_\pi = \lim_{n \to \infty} P^n \).

Under certain conditions [14], the limit will exist and the rows of \( P_\pi \) will be identical. We will denote one of these rows as \( \Pi \). The elements of \( \Pi \) correspond to the long-range probabilities that the Markov chain is in each of the states. Such Markov chains are called **ergodic**.

The GE Theorem can still be applied in this case, for the variable \( S_n = \sum_{k=1}^{n} X_k \). If the Markov chain is ergodic [33], the SCGF of \( S_n = \sum_{k=1}^{n} X_k \) is given by

\[
\lambda(k) = \ln \pi(\tilde{P}_k),
\]

(2.3.15)
where \( \pi(\tilde{P}_k) \) is the dominant eigenvalue (or Perron-Frobenius eigenvalue) of the matrix \( \tilde{P}_k \) whose elements \( \tilde{P}_k(x, x') \) are defined by \( \tilde{P}_k(x, x') = P(x'|x)e^{kx'} \). It is called the tilted matrix associated with \( S_n \). We then obtain that \( S_n \) has an LDP with rate function
\[
I(x) = \sup_{k \in \mathbb{R}} \{ kx - \ln \pi(\tilde{P}_k) \}.
\] (2.3.16)

2.4 Conclusions

In this chapter, we have established the connection between large deviations theory and survival analysis, and we have formulated the main problem studied in the thesis: finding an efficient approximation for the scaled cumulant generating function of a given sample. Specifically, the most notable results discussed in this chapter are:

1) the classification of inference problems based on the degree of complexity of the data, and the relevant methods applicable for each case (Discussion 1);

2) the various formulations of the Large Deviations Principle, from the Cramér functional for samples of i.i.d. with known distribution, to the Gärtner-Ellis theorem for non-i.i.d. samples;

3) the link between survival analysis and probability of large deviations, specifically, between cumulative hazard and rate functions (2.3.5).

As the goal of this work is to use an approximation of the rate function (for a given sample), in order to compute estimates for large deviations probabilities and survival functions, it becomes necessary to identify the best methods available for finding such approximations. As it turns out, estimating rate functions is particularly convenient (when compared to other estimation problems, such as empirical p.d.f. estimation), due to a special property: convexity. In the next chapter, we introduce the basic concepts of approximation theory, as they apply to the case of random variables, and conclude with a class of optimal algorithms which were recently developed in the context of convex optimization. In turn, this will eventually bring us to the specific algorithm proposed here for optimal estimation of rate functions.
Chapter 3

Estimation of large deviations probabilities for empirical distributions

3.1 Goals and outline

In the previous chapter we have arrived at the conclusion that even in the case of complicated data samples (or rather especially in that case), it can be very convenient to compute a generalized rate function $I(x)$, to be used for predicting the probability of large deviations for any given sub-sample average. However, we would want to be able to do this without making assumptions on the structure of the data, i.e. without first applying a model (large sample of i.i.d., Bayes, DLM, Markov chains, etc) and then extracting the rate function using the formulas discussed in the preceding chapter.

We do not address the question of finding such an estimate for the rate function from the purely theoretical point of view. Rather, using recent results obtained in the field of convex optimization, we present an algorithm that can efficiently provide the numerical estimate, given a sample. The reasons (and heuristic evidence) behind the proposed numerical method are discussed as well.

The chapter is structured as follows:

- we introduce the basic concepts of estimation with empirical distributions, and discuss the main concepts used to classify and compare different approximations (space of functions, measures of convergence);

- specifically for the case of approximation of empirical p.d.f. using kernel functions, we indicate the different choices previously used, and the optimal case as it discussed in [13];

- after analyzing the role of convexity in optimal approximations of empirical data, the algorithm is introduced and explained;

- a discussion regarding the specific recent theoretical results supporting this numerical approach concludes the chapter.
3.2 The non-parametric inference approach

Parametric analysis assumes that the distributions of interest belong to certain classes, and therefore the data, models, etc. can be fitted accordingly, using exact analytical expressions.

By contrast, nonparametric analysis makes no assumptions and works exclusively with the empirical c.d.f. and derived quantities. We give here a short description of the basic concepts:

Let $X$ be a random variable with distribution function $F$ and let $X = (x_1, \ldots, x_n)'$ be an observed sample from $F$. Suppose $R(X, F)$ is a statistical quantity that depends in general on both the unknown distribution $F$ and on the sample $X$. For example, $R(X, F)$ could be an estimator of an unknown parameter. If $F$ is unknown, then the exact distribution of the random variable $R(X, F)$ is generally unknown.

A method to nonparametrically estimate the distribution of $R(X, F)$ consists of the following steps:

(i) From the observed sample $X$, use the empirical distribution function, $\hat{F}_n$, as an estimate of the probability function $F$. The empirical distribution function is defined by $\hat{F}_n(x) = \frac{n(x)}{n}$, where $n(x)$ is the number of values $x_i$ in $X$ that are less than or equal to $x$.

(ii) Draw $B$ samples of size $n$ from $\hat{F}_n$ conditional on $X$. Denote these as $X^*_j$, for $j = 1, \ldots, B$.

(iii) For each sample $X^*_j$, compute $R^*_j = R(X^*_j, \hat{F}_n)$ and approximate the distribution of $R(X, F)$ with the empirical distribution of $R^*_1, \ldots, R^*_B$.

The empirical distribution function can also be computed, based on the sample available. Denote this function by $\hat{F}$. A $(1 - \alpha)100\%$ confidence interval based on the percentile method of Efron [27] is given by $[\hat{F}^{-1}(\alpha), \hat{F}^{-1}(1 - \alpha)]$. Here, $x_L = [\hat{F}]^{-1}(\alpha)$ is the largest value of $x$ such that the number of elements in the sample that are less than $x$ is smaller than $\alpha n$. Likewise, $x_U = [\hat{F}]^{-1}(1 - \alpha)$ is the smallest value of $x$ such that the number of elements in the sample that are smaller than $x$ is larger than $(1 - \alpha)n$. Specifically,

$$x_L = \max \left\{ x : \hat{F}_n(x) \leq \alpha \right\}, \quad x_U = \min \left\{ x : \hat{F}_n(x) \geq 1 - \alpha \right\}.$$
### 3.3 Elements of approximation theory

In the following section, we define the main concepts and results of approximation theory for distribution functions, necessary in order to compare and classify various methods for estimation.

**Remark 2** In this section, symbols $x$, $X$ and $\vec{x}$ (and similar for the Greek letters) are assumed to be elements of $\mathbb{R}^n$. The transposed vector is denoted by a prime.

#### 3.3.1 Fourier analysis, distributions

**Definition 3.3.1** The $L^p$-norm ($p \geq 1$) on the space of functions $f : M \subseteq \mathbb{R}^n \rightarrow \mathbb{C}$, with weight function $\rho : M \rightarrow [0, \infty)$, is defined through

$$||f||_{L^p_{(\mathbb{R}^n, \rho dx)}} \equiv \left[ \int_M |f(\vec{x})|^p \rho(\vec{x}) d\vec{x} \right]^\frac{1}{p}$$

**Remark 3** When $\rho = 1$, we will use the simplified notation $||f||_p$.

**Theorem 3.1** The space of functions with finite $L^p$ norm ($p \geq 1$) is a vector space over $\mathbb{C}$. We will refer to them as $L^p$-spaces.

For applications, the following property is extremely important:

**Definition 3.3.2** A vector space $V$ with norm $||.||$ is complete if for any sequence of elements $\{v_k\}_{k=1}^{\infty} \subset V$, absolute convergence with respect to the norm implies that the limit of the sequence is in $V$:

$$\sum_{k=1}^{\infty} ||v_k|| < \infty \Rightarrow \lim_{k \rightarrow \infty} v_k \in V.$$ 

**Remark 4** A complete normed vector space is also called a Banach space.

In fact, a richer structure can be obtained for spaces with a scalar product:

**Definition 3.3.3** A pre-Hilbert space $V$ with scalar product $(,)$ is complete if for any sequence of elements $\{v_k\}_{k=1}^{\infty} \subset V$, absolute convergence with respect to the norm induced by the scalar product implies that the limit of the sequence is in $V$:

$$\sum_{k=1}^{\infty} ||v_k|| < \infty \Rightarrow \lim_{k \rightarrow \infty} v_k \in V.$$
A complete pre-Hilbert space is also called a Hilbert space.

**Theorem 3.2** The dual of the $L^p$–space ($p > 1$) is isomorphic to the $L^q$–space over the same domain $D$, with $q = p/(p - 1) > 1$, for the scalar product

$$(f, g) \equiv \int_D f(X)\overline{g(X)}dX.$$

**Theorem 3.3** The spaces $L^2(D)$ are self-dual and complete (hence, they are Hilbert spaces).

Correspondingly, we can define the following functional spaces:

**Definition 3.3.4** Let $f : \mathbb{R}^n \to \mathbb{C}$, and $n \in \mathbb{N}$. A multi-index power $X^\alpha$, $\alpha = (n_1, n_2, \ldots, n_k)$, $k \leq n$, $\sum_{i=1}^k n_i = n$, is the power function $x_1^{n_1}x_2^{n_2}\ldots x_k^{n_k}$; a multi-index derivative $D^\alpha$, $\alpha = (n_1, n_2, \ldots, n_k)$, $k \leq n$, $\sum_{i=1}^k n_i = n$, is the differential symbol $\partial_{x_1}^{n_1}\partial_{x_2}^{n_2}\ldots\partial_{x_k}^{n_k}$.

**Definition 3.3.5** (Spaces of test functions)

1) Let $U \subset \mathbb{R}^n$ be a compact domain (closed and bounded). The space of $C^\infty$ functions with support in $U$, $f(x) = 0$ for $x \in \mathbb{R}^n \setminus U$ is denoted by $D(U)$;

2) The Schwartz space $\mathcal{S}(\mathbb{R}^n)$ consists of functions $f \in C^\infty(\mathbb{R}^n)$ for which $\sup_{(\alpha, \beta)} |X^\alpha D^\beta f(X)| < \infty$, $(\forall)$ multi-indices $(\alpha, \beta)$;

3) The space of periodic functions $\mathcal{P}[0, T]$ consists of real-valued functions $f(x) = f(x+T)$, $(\forall) x \in \mathbb{R}$, where $T$ is the period.

**Theorem 3.4** For any open set $U \subset \mathbb{R}^n$, and any $p \geq 1$, the following relationships are true:

$$D(U) \subset \mathcal{S}(\mathbb{R}^n) \subset L^p(\mathbb{R}^n), \quad [L^p(\mathbb{R}^n)]^* \subset \mathcal{S}(\mathbb{R}^n)^* \subset D(U)^*.$$

**Remark 6** The dual spaces $\mathcal{S}(\mathbb{R}^n)^*$, $D(U)^*$ are called distribution spaces. Their elements $\lambda$ are generalized functions on which one can operate with differential symbols through the natural scalar product $(\lambda, f) \equiv \int_{\mathbb{R}^n} \lambda(X)f(X)dX$ in the distributional sense:

$$(D^\alpha \lambda, f) = (-1)^{|\alpha|}(\lambda, D^\alpha f),$$

where $(-1)^{|\alpha|} \equiv (-1)^{n(\text{mod} \ 2)}$. 

24
3.3.2 Atomic distributions as limits of smooth functions

For periodic functions, the following representation due to J. Fourier holds:

**THEOREM 3.5** \((\forall) f \in \mathcal{P}[0,T], (\exists) a_n \in \mathbb{R} \) and such that \(a_0,\{a_k,b_k\}_{k=1}^{\infty} \in \mathbb{R} \) and

\[
\lim_{n \to \infty} ||f_n - f||_2 = 0.
\]

To find the coefficients of the Fourier series approximation \(f_n\) of \(f\), we use the orthonormality of the sine and cosine functions, and project on each component:

**THEOREM 3.6** Under the conditions of the previous theorem,

\[
a_0 = \frac{1}{T} \int_0^T f(x)dx,
\]

\[
a_n = \frac{2}{T} \int_0^T f(x) \cos(2\pi nx/T)dx,
\]

\[
b_n = \frac{2}{T} \int_0^T f(x) \sin(2\pi nx/T)dx.
\]

Note that it is straightforward to rescale the variable \(x\) such that \(T = 2\pi\).

An alternative representation is obtained by using Euler’s formula

\[
e^{inx} = \cos(nx) + \sin(nx),
\]

so that the Fourier series takes the complex form

\[
f_n(x) \equiv a_0 + \sum_{k=-n}^{n} c_k e^{ikx}, \quad c_k \in \mathbb{C},
\]

where \(c_k = (a_k - ib_k)/2, \quad c_{-k} = c_k\).

The values \(\omega_n = 2\pi n/T\) are called “modes” or “harmonics” of the Fourier series and correspond to eigenvalues of the translation operator with period \(T\). The set \(\{\omega_n\}_{n \in \mathbb{N}}\) represents the spectrum of the translation operator.

In the limit \(T \to \infty\), the spectrum covers densely the real axis, and we arrive at the notion of Fourier transform:
**Definition 3.3.6** The Fourier transform can be defined for any $L^1$-integrable function, as

$$\tilde{f}(\vec{k}) = \frac{1}{(\sqrt{2\pi})^n} \int_{\mathbb{R}^n} e^{-i\vec{k} \cdot \vec{x}} f(\vec{x}) d\vec{x}$$

Then the inverse transform also exists, and is given by

$$f(\vec{x}) = \frac{1}{(\sqrt{2\pi})^n} \int_{\mathbb{R}^n} e^{i\vec{k} \cdot \vec{x}} \tilde{f}(\vec{k}) d\vec{k}.$$  

In the case where the function $f(\vec{x})$ is a probability distribution function, its Fourier transform is known as the characteristic function, and is widely used to prove convergence in distribution of sequences of random variables.

**Remark 7** The formalism of Fourier integral representations allows us to introduce the notion of atomic measure (or singleton, Dirac distribution) as a limit of smooth functions.

**Definition 3.3.7** Let $\Sigma_{n \times n}$ be a positively-defined quadratic form, and $X, \mu \in \mathbb{R}^n$. The multivariate Gaussian distribution $N(\mu, \Sigma)$ is defined through the probability density function

$$f(X, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\det(\Sigma)|}} \exp \left[ -\frac{1}{2} (X - \mu)' \Sigma^{-1} (X - \mu) \right].$$

**Theorem 3.7** The Fourier transform of $N(\mu, \Sigma)$ is $e^{ik'\mu} |\det(\Sigma)|^{-1/2} N(0, \Sigma^{-1})$, that is

$$\int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n |\det(\Sigma)|}} e^{-\frac{1}{2} (X-\mu)' \Sigma^{-1} (X-\mu) + ik'X} dX = \exp \left[ -\frac{1}{2} k' \Sigma k + ik'\mu \right].$$

In the following, we will use the simplified “diagonal” multivariate distributions

$$f_{\sigma}(X, \mu) \equiv \frac{1}{\sqrt{2\pi\sigma^n}} \exp \left[ -\frac{1}{2\sigma^2} ||X - \mu||^2 \right].$$

**Theorem 3.8** These functions satisfy

$$||f_{\sigma}||_1 = 1, \quad ||f_{\sigma}||_2 = \frac{1}{(\sqrt{\pi\sigma})^{n/2}}.$$  

**Definition 3.3.8** The (weak) limit

$$\lim_{\sigma \to 0} f_{\sigma}(X, \mu) \equiv \delta^{(n)}(X - \mu)$$

defines the $n$–dimensional Dirac generalized function, in distribution sense.
**Remark 8** Since $||\delta||_2 = \lim_{\sigma \to 0} ||f_\sigma||_2 = \infty$, it is not an element of the Schwarz space, but only of its dual.

Since the Fourier transform $\tilde{f}_\sigma(k, \mu) \to e^{ik \cdot \mu}/(2\pi)^n$, it follows from the definition of the inverse Fourier transform that

$$f(\vec{x}) = \frac{1}{(\sqrt{2\pi})^n} \int_{\mathbb{R}^n} e^{ik \cdot \vec{x}} \tilde{f}(\vec{k}) d\vec{k} = \lim_{\sigma \to 0} \int_{\mathbb{R}^n} \tilde{f}_\sigma(k, \mu) \tilde{f}(\vec{k}) d\vec{k} =$$

$$= \lim_{\sigma \to 0} \frac{1}{(\sqrt{2\pi})^n} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \tilde{f}_\sigma(k, \mu) e^{-ik \cdot \vec{x}} f(\vec{x}) d\vec{x} d\vec{k} = \lim_{\sigma \to 0} \int_{\mathbb{R}^n} f_\sigma(\vec{x}, \mu) f(\vec{x}) d\vec{x},$$

and therefore we arrive at the following integral representation.

**Theorem 3.9** The Dirac distribution has the spectral representation

$$\delta^{(n)}(X, \mu) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{ik \cdot (X-\mu)} d\vec{k}.$$

From the spectral representation, the following properties follow:

**Theorem 3.10** For any function $f \in L^2(\mathbb{R}^n, dx)$, we have

1) Reproducing kernel identity:

$$\int_{\mathbb{R}^n} \delta^{(n)}(\vec{x} - \vec{y}) \cdot f(\vec{y}) d\vec{y} = f(\vec{x}).$$

2) The Fourier transform of the Dirac distribution is the constant function equal to 1;

3) Parseval’s lemma:

$$||f(\vec{x})||_2 = ||\tilde{f}(\vec{k})||_2.$$

Let us illustrate what the limit in Definition 3.3.8 signifies, in terms of weak convergence and approximation theory. We choose to represent several Gaussian distributions, with increasingly sharper concentration of measure (i.e. smaller and smaller standard deviation, $\sigma$), in Figure 5.

This direct illustration serves as basic example for the concept of kernel approximation of empirical distributions, because the atomic measure is found (by a corresponding limit process) in the dual space of every distribution space of interest c.f. Theorem 3.4. More precisely, the kernel-based approximation method consists of the following steps:
• choose the space of smooth functions in which the desired approximation is to be found (e.g. the standard spaces defined in Definition 3.3.5);

• for the given space, consider a sequence of appropriate kernel functions (which provide a map between the space and its dual), such that, in the corresponding norm (Definition 3.3.2), the sequence converges to the atomic distribution;

• finally, for the given measure of convergence (norm) and family of kernel functions, solve the optimization problem of finding the best choice for the parameters of the family of approximating functions.

We further discuss this general procedure for the case of kernel estimates of empirical p.d.f., in the next section.

3.4 Point-wise approximations of empirical p.d.f. and comparison with empirical m.g.f.

Assume that if $X_1, \ldots, X_n$ are i.i.d. random variables having a common p.d.f. $f(x)$, then the kernel estimate of $f(x)$ is defined by

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$
where $h$ is the bandwidth and $K(u)$ is the kernel function. The kernel estimate of the cumulative distribution function $\hat{F}(x)$ is then

$$\hat{F}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{x} K \left( \frac{y - X_i}{h} \right) dy,$$

and correspondingly the estimate of survival function becomes

$$\hat{S}_n(x) = 1 - \hat{F}_n(x)$$

It is usually assumed that $K(x, y)$ is a symmetric function, which can be taken to be normalized to 1, centered (zero first moment), and positive-definite (positive second centered moment).

Properties of the kernel function $K(u)$ partially determine the properties of the kernel density estimates, such as smoothness and integrability. For example, if $K(u)$ is a proper density function (a positive function of $L^1$-norm one), then the kernel density estimate is also a proper density function. If $K(u)$ is $k-$times differentiable, so is $\hat{f}_n(x)$.

The optimal bandwidth is determined by minimizing the mean square error (MSE) for the estimate,

$$MSE(\hat{f}) = \mathbb{E}((\hat{f} - f)^2) = \text{Bias}(\hat{f})^2 + \text{Var}(\hat{f})$$

Denote $\|K\|^2 = \int K^2(t) dt$, $M_2 = \int t^2 K(t) dt$, then if $h \to 0$, $nh \to \infty$, and the underlying density $f$ is a sufficiently smooth $L^2$ function, it is known that $\text{Bias} \sim h^2 M_2$, and $\text{Var} \sim \|K\|^2/(nh)$.

Therefore, if the bandwidth decreases, then the bias of the kernel estimate also decreases; however, the variance increases, yielding an overspread, inaccurate estimate of the kernel density. Conversely, if the bandwidth increases, the variance of the kernel estimate decreases while the bias increases. Thus there is significant smoothing of the underlying characteristics of the probability density. Taking both of these results into account, we arrive a better estimate of the global accuracy of $\hat{f}(x)$, the asymptotic significant mean integrated square error (AMISE), obtained by integrating the MSE over the entire real line: We then minimize AMISE by choosing the appropriate kernel and the bandwidth. Perhaps not surprisingly, the optimal kernel function was found by Epanecnikov in 1969 to be

$$K(u) = \frac{3}{4}(1 - u^2)\mathbb{I}(|u| \leq 1),$$

that is to say a kernel with compact support (corresponding to the space $D(U)$ in Theorem 3.4). Since $D(U)$ is included in all the other distribution spaces, its dual is the largest of the duals, so it affords the most precise kernel estimation.
3.5 Approximating empirical rate functions: compactness, convexity, and optimality

Consider again the sample of i.i.d. r.v.’s \( \{x_k\}^n_{k=1} \), with unknown distribution function, and the problem of estimating the corresponding rate function for LDP. From the defining relation (2.3.1), at a given value of the argument \( x \), we have

\[
I(x) = \sup_t [tx - \log m(t)], \quad m(t) = \mathbb{E}(e^{tX}).
\]

Therefore, for distributions with compact support (or for any finite sample), we can write

\[
-I(x) = \inf_t [\log m(t) - tx] \Rightarrow e^{-I(x)} = \inf_t m(t)e^{-tx},
\]

using the monotonicity of the exponential function. Therefore, in order to estimate the large-deviations probability \( e^{-I(x)} \), we must first minimize over \( t \) the function \( m(t)e^{-tx} \). We do this using at first the empirical distribution of the sample \( \{x_k\} \), i.e. by replacing

\[
m(t) \rightarrow \hat{m}(t) := \frac{1}{n} \sum_{k=1}^{n} e^{tx_k}.
\]

A direct calculation of the first-order derivative gives the extremum condition as

\[
e^{-tx} \sum_k x_k n_k e^{tx_k} = e^{-tx}x \sum_k n_k e^{tx_k},
\]

that results into

\[
x = \frac{\sum_k x_k n_k e^{tx_k}}{\sum_k n_k e^{tx_k}} = \mathbb{E}_{P_t}(X),
\]

where \( n_k \) represents the multiplicity of the observation \( x_k \) in the sample, and

\[
P_t(x_k) := \frac{n_k e^{tx_k}}{\sum_j n_j e^{tx_j}} \quad (3.5.1)
\]

is a Gibbs-type discrete distribution over \( \{x_k\} \), with parameter \( t \).

Differentiating \( \hat{m}(t)e^{-tx} \) twice with respect to \( t \), we have

\[
e^{-tx} \sum_k \left[x_k^2 - 2x_k x + x^2\right] e^{tx_k} = e^{-tx} \sum_k (x_k - x)^2 e^{tx_k} \geq 0,
\]

so as expected this functional is globally convex, and the extremum point is the (unique) minimum.

An important property of the survival function (and of any of its proper estimators) is the monotonicity which follows from the monotonicity of the c.d.f.:

\[
F(x_1) \leq F(x_2), \quad (\forall) \ x_1 \leq x_2 \Rightarrow S(x_1) \geq S(x_2), \quad (\forall) \ x_1 \geq x_2.
\]
We must verify that the large-deviations estimator obtained by the procedure outlined above preserves the monotonicity property. To that end, we differentiate the estimator with respect to $x$, using chain rule and the fact that the optimal value $t_*$ is a function of $x$: 

$$
\frac{d}{dx} \left[ \hat{m}(t_*) e^{-t_* x} \right] = \frac{dt_*}{dx} \cdot \frac{\partial}{\partial t} \left[ \hat{m}(t_*) e^{-t_* x} \right]_{t=t_*(x)} + \frac{\partial}{\partial x} \left[ \hat{m}(t_*) e^{-t_* x} \right] \tag{3.5.2}
$$

Since the first term in Equation (3.5.2) vanishes as $\hat{m}(t) e^{-tx}$ reaches its minimum at $t = t_*$, we obtain

$$
\frac{d}{dx} \left[ \hat{m}(t_*) e^{-t_* x} \right] = -t_* \left[ \hat{m}(t_*) e^{-t_* x} \right] \leq 0, \quad (\forall) t_* > 0.
$$

Thus, the estimator obtained has the required monotonicity property as a monotonically decreasing function of $x$.

These observations can now be summarized into the first algorithm of approximation for the empirical rate function in the case of samples of i.i.d. r.v.’s, given below in pseudocode:

**Algorithm I (case of i.i.d. r.v.)**

1) **INPUT** $x$

2) **INPUT** sample $S = \{x_1, x_2, \ldots, x_n\}$

   (i) **IF** $x < \text{MIN}(S)$, **THEN** RETURN 1;

   (ii) **ELSE IF** $x > \text{MAX}(S)$, **THEN** RETURN 0;

   (iii) **ELSE** $t_* \leftarrow \text{CONVEX-MINIMIZE} \text{ function } \hat{m}(t) e^{-tx}$

3) **RETURN** $\hat{m}(t_*) e^{-t_* x}$

**Remark 9** In general, step 2)-(iii) of the algorithm would require a separate discussion regarding the efficiency of numerical solvers used. This is due to the fact that the equation to solve in order to find the (global) minimum is transcendental, i.e. we are seeking to perform nonlinear convex minimization, also known as nonlinear programming (NLP). In several dimensions, such problems are known to have high computational complexity (for instance, even quadratic programming problems can be NP-hard [40]). Fortunately, in our case, the function to minimize depends only on one real variable ($t$), so the problem can be efficiently solved numerically by employing predefined numerical solvers such as *Mathematica’s NMinimize* procedure. Because of this fact and because the
main input is a list (the basic data structure for the software), implementing the algorithm above in Mathematica 9 is particularly convenient (e.g. there are also predefined functions for finding the minimum and the maximum in an unsorted list).

**Example 6** To illustrate the algorithm and compare its performance against that of the Kaplan-Meier estimator (the most popular choice in non-parametric survival analysis with samples of i.i.d. r.v.), we generate \( n = 50 \) observations drawn from the exponential distribution with parameter 1, \( X_k \sim \text{Exp}(1), k = 1, 2, \ldots, n \). Based on this sample, we then compute the Kaplan-Meier estimator for \( \hat{S}(x) \) as outlined in Chapter 2, and the large-deviations estimator \( \hat{m}(t_x)e^{-t_x x} \), and plot both of them along with the exact formula giving \( S(x) \) (computed in Example 1, Chapter 2), in Figure 6. It can be easily observed that, besides being much smoother (and thus, more realistic) than the Kaplan-Meier estimator, the large-deviations one is also a much better approximation for the exact survival function \( S(x) \).

![Figure 6](image)

**Figure 6.** Comparing the Kaplan-Meier and large-deviations estimator to the exact formula.

### 3.6 Approximating generalized rate functions (case of non-i.i.d. samples)

In this section, we discuss the general case when the sample of observations is not i.i.d., and we do not know the distributions they are drawn from. The conditions that we impose, however, allow to
make use of the generalized LDP, as covered by the Gärtner-Ellis theorem mentioned in the previous chapter, and in particular for the case of sample averages for observations originating from a finite-state Markov chain. The main idea behind these results is that, whenever the sample of non-i.i.d. r.v.
is obtained by computing sub-sample averages of i.i.d.’s (the case of the G-E theorem), or summing consecutive observations from a Markov chain, there is a way to re-structure the sample in such a way as to retrieve the original sequence of i.i.d it comes from, and then apply Cramér’s result. The theoretical foundation of this work is still under development [41], yet following an empirical point of view, we propose here a numerical algorithm derived from these ideas, and apply it in order to quantify its efficiency and overall performance.

In order to explain the main idea and rationale for this generalization of the previous algorithm, we must mention some recent results obtained in the field of convex optimization and surrogate functions for classifying-type problems [42]-[44]. In turn, this requires a collection of basic facts concerning stochastic processes, which we mention in the following.

### 3.6.1 Stochastic processes on \( \mathbb{R} \): a classification

We call a a stochastic process (SP) *stationary* if it has the property:

\[
\mathbb{E}(X_{s+r}X_{t+r}) = \mathbb{E}(X_sX_t) \Rightarrow K(s, t) = \tilde{K}(s - t),
\]  

(3.6.1)

where \( K \) is the *auto-correlation function* \( K = \text{Cov}(X_s, X_t) \). Then (by Bochner-Lesbegue theorem) there is a measure on \( \mathbb{R}, \mu \), such that:

\[
\tilde{K}(t) = \int e^{it\nu} \mu(\nu) d\nu.
\]  

(3.6.2)

In other words, the components of the autocorrelation function are Fourier transforms of well-behaved densities. This allows to use standard tools of *filtering, prediction and estimation*.

However, in most situations, the processes observed do not seem to satisfy this property, so it becomes useful to consider various degrees of generalizations for stationarity.

**Strongly harmonizable processes**

A non-stationary process \( X_t \) is strongly harmonizable if there exists a measure \( \mu \) on \( \mathbb{R}^2 \), such that

\[
\tilde{K}(s, t) = \int_{\mathbb{R}^2} e^{i(s\nu - t\eta)} \mu(\nu, \eta) d\nu d\eta,
\]  

(3.6.3)
Then we have the characterization:

**Theorem 3.11** $X_t$ is strongly harmonizable if there is a positive constant $C$ such that:

$$\left| \int_{\mathbb{R}^2} f(s, t)K(s, t)dsdt \right| \leq C||f||_{\infty},$$

(3.6.4)

for any finite-norm function $f$. Alternatively, $X_t$ is strongly harmonizable if it is the Fourier transform of another stochastic process on the reciprocal space:

$$X_t = \int_{\mathbb{R}} e^{it\nu} Z(\nu) d\nu,$$

(3.6.5)

where $Z$ satisfies $\mathbb{E}(Z(A), Z(B)) = \mu(A \times B)$, for any $A, B \subset \mathbb{R}_+$. If $Z$ satisfies $Z(A) \perp Z(B)$ if $A \cap B = \emptyset$, then $X$ is actually stationary.

This characterization is important because of the following:

**Theorem 3.12** Any strongly harmonizable process is asymptotically stationary (A.S.), i.e. there is a smooth limit

$$\tilde{K}(t) = \lim_{T \to \infty} \left[ \frac{1}{2T} \int_{-T}^{T} K(s + t, s)ds \right], \quad (\forall) t \in \mathbb{R}_+.$$

(3.6.6)

Also, with minimal supplementary conditions, estimation, prediction and filtering of asymptotically stationary processes is the same as for stationary ones.

**Weakly harmonizable processes**

**Definition 3.6.1** We call a function $u$ a bi-measure if it is the Fourier transform of any bounded function on $\mathbb{R}^2$.

**Theorem 3.13** $X$ is weakly harmonizable if it is a bi-measure. The following are equivalent characterizations:

i) $X$ is the Fourier transform of an arbitrary stochastic process on the reciprocal space;

ii) $X$ is the projection of a stationary process $Y$ from a higher-dimensional space.

**Conclusion:** Generically, a projection of stationary processes to lower dimensions leads to a weakly harmonizable process. Therefore, we lose all the important properties specific to stationary and asymptotically stationary processes, just by projecting.
Applications: estimation, prediction and filtering

We can now list the applications to estimation, prediction and filtering for asymptotically stationary processes:

• Estimation: assume $X$ is A.S. and that

$$\lim_{T \to \infty} \sup_{[-T, T]} \left| \frac{1}{2T} \int_{-T}^{T} K(s + t, s) ds \right| = 0,$$

(3.6.7)

$$\frac{1}{2T} \int_{-T}^{T} \left[ \mathbb{E}(||X_t||^4) \right]^{1/2} dt \quad \text{bounded (}\forall\text{ } T,$$

(3.6.8)

then

$$\mu_{n,N}(\nu) := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{4n^2} \int_{-n}^{n} \int_{-n}^{n} e^{-it\nu} \mathbb{E}[X_i(s + t)X_i(s)] ds dt$$

(3.6.9)

is a consistent estimator of $\mu$.

• Prediction: a least-squares predictor (LSP) of any A.S. process with limit auto-correlation $K$ is equally good as an LSP of a stationary process with auto-correlation $K$.

• Filtering: adding any stationary noise term to an A.S., with vanishing auto-correlation in the infinite-time limit, and with bounded fourth moment does not affect any estimator or predictor based on the A.S.

If the sample of observations (non-i.i.d. r.v.) corresponds to a stochastic process satisfying the conditions of Theorem 3.11, then we can efficiently construct an algorithm for approximating the generalized rate function [41]. Moreover, there is sufficient heuristic evidence [42]-[44] to justify applying the algorithm in the case where the sample satisfies the relaxed conditions of Theorem 3.13.

Algorithm II

1) INPUT $x$

2) INPUT sample $S = \{x_1, x_2, \ldots, x_n\}$

3) find the choice of sample (block) partitioning minimizing the total relative entropy for Gibbs distributions $P_t^{(j)}$ of subsamples
4) \( t_s \leftarrow \text{MAXIMIZE the total entropy } W(t, x) \)

5) \( \text{RETURN } I(x) = t_s x - W(t_s, x). \)

**Remark 10** It is useful to note that the procedure proposed here preserves the global convexity property, since block-updating and sample partitioning are performed using convex combinations of the elements of the original sample.

1) For a sample of i.i.d. r.v. (equivalently, a stationary process of null auto-correlation), the algorithm reduces to step 4. In turn, this is equivalent to Algorithm I applied to the entire sample.

2) The efficiency of step 3) is insured by performing the block-partitioning of the sample using the surrogate functions proposed in [44], rather than the actual functional to minimize (relative entropy). This is because the surrogate functions (in this case, subsample correlations) are much more efficient at solving the classifier-type problem posed by the block partitioning.

In the remainder of the thesis, we exemplify this algorithm (which contains the simpler case of sample of i.i.d.) on real data obtained from studies of climate change (modeling and global warming), and use these examples in order to comment on the usefulness of the proposed procedure.
Chapter 4

Predicting large-scale climate perturbations in the Southern Pacific

This chapter is rooted in a study based on monthly measurements of air pressure, taken at sea level, between 1882 and 1998, at the Darwin research station, in Northern Territory, Australia. These measurements are an important part of a global monitoring system, which includes measuring pressure and air temperature across the Pacific, from Darwin in the West to Galapagos Islands in the East. The normal weather patterns in this area were studied and described systematically for the first time by Sir Gilbert Thomas Walker [51, 52]. The air and water patterns he discovered are called the \textit{Walker circulation}. Anomalies (deviations from this pattern) include the phenomena known as El Niño and La Niña, which will be described briefly below. They are large-scale fluctuations, with global consequences, reaching all the way to continental USA.

Figure 7 is a schematic description of the normal weather patterns in Equatorial Pacific. Air is warmer, more moist, and with lower pressure at Darwin than in the East. The warm air evaporates, goes up in the atmosphere, then travels East, where it gets cooled off and drops back to sea level. The circulation is then completed by the colder air traveling from East to West, giving rise to the well-known \textit{trade winds}. They are so consistent that the water level is higher at Darwin than in Galapagos by about 60 centimeters. On the other hand, cold water goes up to the surface in the East, gets carried West by the winds, where it becomes warmer and drops to the bottom. From there, it circulates back East.

This is the Walker circulation. There are two ways in which it may be perturbed:

1) either the air/water become warmer in the East, so the whole Pacific is warmer than average. Likewise, the air pressure becomes lower everywhere – this is El Niño;

2) or the air/water become colder in the West, so the whole system is colder than average, with higher air pressure – this is La Niña.
Regardless of which mechanism happens, the fact that differences in temperature/pressure become smaller between the two end of the Pacific leads to decreased circulation, which further amplifies the anomaly, etc. Figure 8 shows a detailed temperature map for a cold and warm episode, respectively.

4.1 The Dynamical Linear Model

4.1.1 Theoretical considerations

We begin by reviewing the basic assumptions and results of Dynamical Linear Model (DLM) theory [48], relevant for the problem studied. The following summary of the DLM approach follows closely the [48] and references cited there.

Let us discretize the time step and transform all linear differential equations (in time) into differ-
ence equations, with unit time step. Then a general dynamical linear model is given by

\[ Y_t = F_t X_t + \nu_t, \quad \nu_t \sim N(0, V_t), \quad (4.1.1) \]
\[ X_t = G_t X_{t-1} + \omega_t, \quad \omega_t \sim N(0, W_t), \quad (4.1.2) \]

where again \( Y_t \) is the vector of observables (in our case it is a scalar, pressure), \( X_t \) is the (hidden) state vector, and the matrices \( G_t, F_t \) represent the linearizations of dynamical equations and projection, respectively. The noise terms are taken to be uncorrelated, unbiased, time-dependent Gaussians.

Figure 8: Temperature maps for El Niño and La Niña episodes.
Initial conditions are taken to be
\[ X_0 \sim N(\mu_0, \Sigma_0). \] (4.1.3)

Under these conditions, the time-dependent solution of the system (4.1.1) is given via the updating equations (here we define the collection of data at time \( t \) to consist of the observed variables \( D_t := \{Y_0, \ldots, Y_t\} \)):

\[
(X_t|D_{t-1}) \sim N(a_t, R_t), \quad a_t = G_t \mu_{t-1}, \quad R_t = G_t \Sigma_{t-1} G'_t + W_t, \quad (4.1.4)
\]

\[
(Y_t|D_{t-1}) \sim N(\phi_t, Q_t), \quad \phi_t = F_t a_t, \quad Q_t = F_t R_t F'_t + V_t, \quad (4.1.5)
\]

\[
(X_t|D_t) \sim N(\mu_t, \Sigma_t), \quad \mu_t = a_t + A_t e_t, \quad \Sigma_t = R_t - A_t Q_t A'_t, \quad (4.1.6)
\]

\[
A_t = R_t F'_t Q_t^{-1}, \quad e_t = Y_t - \phi_t. \quad (4.1.7)
\]

These updating equations assume complete knowledge of the parameters of the noise (here, only covariance matrices). However, we never really have this knowledge in reality. Hence, we must consider the updating equations unconditional on \( V_t \):

\[
(X_{t-1}|D_{t-1}) \sim T_{n(t-1)}(m_{t-1}, \Sigma_{t-1}) \quad (4.1.8)
\]

\[
(X_t|D_{t-1}) \sim T_{n(t-1)}(a_t, R_t) \quad (4.1.9)
\]

\[
(Y_t|D_{t-1}) \sim T_{n(t-1)}(\phi_t, Q_t) \quad (4.1.10)
\]

\[
(X_t|D_t) \sim T_{n(t)}(\mu_t, \Sigma_t), \quad (4.1.11)
\]

where \( n(t) = t \) (to avoid confusion). (In these equations, \( T_n \) stands for the \( T \) distribution with \( n \) degrees of freedom.)

**Forecasting (prediction)**

Let us introduce the forecast function

\[
f_t(k) = \mathbb{E}[Y_{t+k}|D_t], \quad k \geq 1. \quad (4.1.12)
\]

Then the following hold:

\[
(X_{t+k}|D_t) \sim T_{n(t)}(a_t(k), R_t(k)) \quad (4.1.13)
\]

\[
(Y_{t+k}|D_t) \sim T_{n(t)}(f_t(k) t, Q_t(k)), \quad (4.1.14)
\]
where
\[ f_t(k) = F_{t+k}a_t(k), \quad Q_t(k) = F_{t+k}R_t(k)F'_{t+k} + V_{t+k}, \quad (4.1.15) \]
and
\[ a_t(k) = G_{t+k}a_t(k-1), \quad R_t(k) = G_{t+k}R_t(k-1)G'_{t+k} + W_{t+k}, \quad (4.1.16) \]
Covariance matrices may also be forecasted in a similar manner.

### 4.1.2 Data analysis

Figure 9 is a plot of the 1400 data points, monthly measurements from 1882 to 1998. The horizontal axis shows the month in the series, and the vertical axis shows the air pressure, in units of 10 KiloPascal. The standard air pressure at sea level is considered to be 101 KiloPascal, or 10.1 on this scale. Over the duration of the experiment, measured values ranged from 3 to 15. It is of course most interesting to determine the seasonality of these observations.

To study the periodicity, a periodogram of the data was performed, shown in Figure 10. The horizontal axis shows the frequency, in inverse months, and the vertical axis is the spectral power, on linear scale. Clearly, there is only one harmonic to this data, a little below 0.1, more precisely at \( \frac{1}{12} \), so the only periodicity of the data is annual, every 12 months.

### 4.1.3 DLM formulation

Using this information, it is now possible to construct a minimal DLM model which would capture the seasonality in the data [48]. It is given by Eq. (7.1.2), where the model parameters are shown in Eq. (4.1.18), with the seasonality implemented through the block-diagonal component, with frequency 1/12.

\[ Y_t = F'_t \cdot \theta_t + \nu_t, \quad \theta_t = G_t \cdot \theta_{t-1} + \omega_t \quad (4.1.17) \]

\[ F'_t = (1 \ 1 \ 0), \quad G_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(2\pi f) & \sin(2\pi f) \\ 0 & -\sin(2\pi f) & \cos(2\pi f) \end{bmatrix}, \quad f = 1/12 \quad (4.1.18) \]

\[ \nu_t \sim N(0, V_t), \quad \omega_t \sim N(0, W_t), \quad \text{discount factor } \delta \quad (4.1.19) \]

We also specify the noise to be Gaussian, with zero mean. The variances \( V_t \) and \( W_t \) are all that is left to determine the model. We simplify the model by determining \( W_t \) through a discount factor.
Figure 9.: The 1400 sea level pressure data points.

\[ W_t = \frac{1 - \delta}{\delta} G_t \cdot C_{t-1} \cdot G_t', \quad \text{Cov}(\theta_t|D_t) = C_t. \]  

(4.1.20)

\( \delta \). It will be determined entirely by the variance matrix of the distribution of \( \theta_t \), given the data up to \( Y_t \), Eq. (4.1.20). We will then determine the optimal value of the discount factor \( \delta \) by computing the measures MSE (mean square error), MAD (mean absolute difference), and LIK (likelihood of data).
4.1.4 The MCMC algorithm

What remains to be determined now is the variance of the observational noise, $V_t$. This is given by nature, so it is unknown, and not clear how to choose it. On the other hand, it is very important, since it enters in all the updating equations for the DLM model.

We choose to estimate the unknown variance $V_t$ through the Monte Carlo Markov Chain (MCMC) algorithm, with the Metropolis-Hastings criterion [54]. In order to do that, we assume that $V_t$ has the form $V_t = \lambda_t V$, where $V^{-1}$ has a Gamma distribution [48], and $\lambda_t$ may take values 1 or $k^2$, for some constant $k$ (in this case, $k = 3$). The global algorithm of the program can be summarized as follows:
Take (see [48])

\[ V_t = \lambda_t V, \quad V_t^{-1} \sim \Gamma \left( \frac{n_t}{2}, \frac{n_t S_t}{2} \right) \]

Choose a value of \( \lambda_t, V \) as initial estimate

Update DLM for unknown, constant \( V_t \) with given initial estimate:

\[ f_t = F'_t \cdot G_t \cdot m_t, \quad Q_t = F'_t \cdot R_t \cdot F_t + S_{t-1}, \quad R_t = \text{Cov}(Y_t|D_t) \]

\[ e = Y_t - f_t, \quad S_t = S_{t-1} \cdot \left[ 1 + \frac{1}{n_t} \left( \frac{e^2}{Q_t} - 1 \right) \right], \quad n_t = n_{t-1} + 1 \]

Apply the Metropolis-Hastings criterion (a randomized Neyman-Pearson, UMP test) and decide whether to change \( \lambda_t \) or not:

- for given data \( \{Y_i\} \) and estimates \( \{f_i\} \), compute the likelihood functions

\[
L_{\lambda_t=1}(\{Y_i\}) = \exp \left[ -\frac{(Y_i - f_i)^2}{2V} \right], \quad L_{\lambda_t=k^2}(\{Y_i\}) = k^{-1} \exp \left[ -\frac{(Y_i - f_i)^2}{2Vk^2} \right]
\]

and form the variable \( r \in (0, 1), r = 0.95 \cdot L_{\lambda_t=1}/(0.95 \cdot L_{\lambda_t=1} + 0.05 \cdot L_{\lambda_t=k^2}) \);

- compare \( r \) with a random variable drawn from the uniform distribution on \([0, 1]\):

\[
\begin{cases}
\text{if } r < \mu \in \text{U}[0, 1], & \text{then } \lambda_t = 1 \\
\text{if } r \geq \mu \in \text{U}[0, 1], & \text{then } \lambda_t = k^2
\end{cases}
\]

- for a whole set of data \( \{Y_i\} \), estimates \( \{f_i\} \) and \( \{\lambda_t\} \), draw \( V \) from the appropriate gamma distribution again.

Iterate until achieving convergence

Repeat everything for several values of \( \delta \) and choose optimal value

4.2 Filtering, smoothing and forecasting

4.2.1 Filtering recurrences and smoothing the data

Within the framework of the model presented, it is possible to perform retrospective analysis [48], by applying backwards recurrence relations, and computing the forecast functions with negative arguments:

From the backwards operator \( B_t \equiv C_t \cdot G'_{t+1} \cdot R_{t+1}^{-1} \),

\[ (4.2.1) \]
we derive the mean of filtered marginal distributions

\[ \mathbb{E}(\theta_{t-k} | D_t) = a_t(-k) = m_{t-k} + B_{t-k}[a_{t}(-k + 1) - a_{t-k}(1)]. \] (4.2.2)

The \( k \)-step backwards forecast function is then

\[ f_t(-k) = F'_{t-k}a_{t}(-k) \] (4.2.3)

Having these values, it is possible to detect and eliminate outliers, by doing testing based on the distributions of observations, state vectors and observational noise variance, in the presence of deleted observations [48]. In other words, observations which are not found inside a confidence interval centered on \( f_t(-k) \), are deleted, thus smoothing out the data series. This allows for better performance of our model when computing \( k \)-step forward forecast functions, indicated in the next section.

### 4.2.2 \( k \)-step ahead forecasting

For constant parameters \( F'_t, G_t \), the \( k \)-step forecast function

\[ f_t(k) = F'_t \cdot G^k_t \cdot m_t, \quad m_t = \mathbb{E}(\theta_t | D_t) \] (4.2.4)

can be expressed in scalar form as

\[ f_t(k) = m_{t,1} + m_{t,2} \cdot \cos \left( \frac{k\pi}{6} \right) + m_{t,3} \cdot \sin \left( \frac{k\pi}{6} \right). \] (4.2.5)

In our case, \( t = 1400 \) and \( k \) ranges from 121 to 132 (one year of forecast, ten years into the future). Using the model with the optimal value of \( \delta = 0.7 \) (although measures were rather insensitive to variations around this value), we computed the forecasts (4.2.5) for the year 2008 (ten years from the last year of observation, 1998).

### 4.3 Interpretation of the results and conclusions

The model investigated in this study has limitations of two different kinds. On one hand, its predictive power is limited by the fact that it takes into account only pressure data. The main issues related to this fact are summarized below:
Figure 11.: Predicted values for 12 consecutive months, ten years into the future.

- pressure anomalies alone do not indicate a major climate change, owing to the secondary nature of this variable. The El Niño/La Niña phenomena are determined by temperature variations, more precisely by consecutive deviations of 0.5°C or more, for five month or longer\(^1\) [55];

- among the factors ignored is the influence of salinity, which is very relevant because of the accelerating rate of melting of glaciers in Antarctica; other factors come from the fact that Equatorial Pacific is not a closed system, but coupled to other systems, like Northern Pacific.

Additionally, the model may be too incomplete from the point of view of its design. Instead of using the model matrix \(G_t = \text{BlockDiag}[1, J_2(1, \pi/6)]\), a second discount factor could have been introduced, giving the matrix \(G_t^{(\delta')} = \text{BlockDiag}[\delta', J_2(1, \pi/6)]\), which has the potential to

\(^1\)U.S. Department of Commerce National Oceanic and Atmospheric Administration.
capture more accurately the fluctuations in the data, and therefore might not predict a net trend like our current model does.

After completing this project in the form presented thus far, we attempted to modify the algorithm in order to implement such a generalized, discount factor-based model. Within the existing time constraints, the preliminary tests for the new model did not lead to significantly different predictions for 121 to 132 months ahead in time; it is possible that a more detailed analysis would show a qualitatively different conclusion.

### 4.4 Using the generalized algorithm for prediction of La Niña

In order to compare the efficiency of the numerical methods proposed in Chapter 3, and applicable for the seasonal data characteristic for the problem discussed here, we performed an independent analysis for the year 1998. Then, we compared the results of the analysis to the forecast provided by the DLM approach. Below are listed the procedural steps which constitute the large-deviations study:

- Run Algorithm II for data type detection: 12-month periodicity detected
- Apply algorithm I to evaluate rate function of large deviations for the 12 subsamples (data through 1997)
- Estimate probability of large deviations in pressure (at level defined by the La Niña/El Niño phenomenon), \((p - \bar{p})/\bar{p} > 2\%\)

When applied to the 12 months of 1998, the procedure yields the following estimates:

<table>
<thead>
<tr>
<th></th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>.86</td>
<td>.96</td>
<td>.39</td>
<td>.62</td>
<td>.89</td>
<td>.57</td>
<td>.61</td>
<td>.78</td>
<td>.92</td>
<td>.83</td>
<td>.68</td>
<td>.74</td>
</tr>
</tbody>
</table>

Since the DLM forecast for 1998 also predicts an El Niño episode, consistent with the meteorological record, we conclude that the algorithm performs as well as the DLM approach, while being considerably less demanding in computational terms.
Chapter 5

Estimating the atmospheric CO₂ levels - US data and model with interactions

In this chapter (and the following two), we embark on a manifold study of risk factors (attributable variables) relevant for the variation of atmospheric CO₂ levels for the period since 1960. In the present analysis, we used real yearly data that has been collected from 1959 to 2008 for the continental United States. The air samples were collected at Mauna Loa Observatory, Hawaii, and the CO₂ emission data was obtained from Carbon Dioxide Information Analysis Center (CDIAC) at Oak Ridge National Laboratory Division of U.S. Department of Energy.

The analysis presented here consists of two parts: we first partially replicate the comprehensive study performed in [45] in order to select the relevant variables and their interactions, and then we perform the surface-response analysis (nonlinear modeling) to the model obtained in [45].

5.1 Regression analysis and model building

5.1.1 Second-order model: parameter determination and validation

One of the underlying assumptions to construct the model is that the response variable should follow Gaussian distribution. It is known [45] that the CO₂ in the atmosphere does not follow the Gaussian distribution.

Therefore, the Box-Cox transformation is applied to the CO₂ atmosphere data to filter the data to be normally distributed. After the Box-Cox filter, we retest the data and it shows our data will follow normal distribution; thus, we proceed to estimate the coefficients of the contributable variables for the transformed CO₂ atmosphere data.

We can proceed to estimate the approximate coefficients of the contributable variables for transformed CO₂ in the atmosphere and obtain the coefficient of all possible interactions.

At the same time, we can determine the significant contributions of both attributable variables and interactions. We begin with seven attributable variables as previously defined as $X_2, \ldots, X_8$
in the dataset (since the values listed as $X_1$ in the dataset are not relevant, $X_1$ being just the sum of all variables $X_i, i \geq 2$), and arrive by applying the stepwise forward selection procedure at a model with only five relevant variables (subsequently renamed $x_1, x_2, x_3, x_4, x_5$, corresponding to the original variables $X_2, X_3, X_5, X_6, X_8$), and fifteen 2$^{nd}$ order interactions between each pair.

We find that only five interactions are statistically relevant at $\alpha = 0.01$ level.

Thus the result of estimation becomes the quadratic model with interactions (fully consistent with the results of [45]):

$$[\overline{CO}_2]^{-2.376} = \beta_0 + \sum_{i=1}^5 \beta_i x_i + \sum_{i \leq j = 1}^5 \beta_{ij} x_i x_j,$$

where the measure for goodness-of-fit ($R^2 = 0.9973$ and the $p$–value less than 0.0001), as well as parameters $\{\beta_k\}$, are found from the SAS output: $x_1 =$ Gas Fuels, $x_2 =$ Liquid Fuels, $x_3 =$ Gas Flares, $x_4 =$ Cement, $x_5 =$ Bunker. Their corresponding coefficients are given in Table 2.

The only non-zero interaction coefficients are obtained as follows (after rescaling by a global scale factor of $10^{-19}$):

$$\beta_{13} = -2.107, \beta_{23} = 5.593, \beta_{24} = -2.559, \beta_{35} = -58.22, \beta_{45} = 20.49$$
Table 2: Linear regression coefficients for attributable variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intercept</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{17} \times \beta$</td>
<td>3.196 $\cdot$ 10^8</td>
<td>-2.586</td>
<td>-129.6</td>
<td>-1939</td>
<td>6922</td>
<td>-896.1</td>
</tr>
</tbody>
</table>

Therefore, we can write our model in matrix notation (where prime denotes transposition) as

$$Y = \beta_0 + \beta' \cdot X + X' \cdot B \cdot X,$$

(5.1.1)

with the obvious identifications

$$X' = (x_1, \ldots, x_5), \quad \beta' = (\beta_1, \ldots, \beta_5),\quad B_{ij} = B_{ji} = \frac{1}{2} \beta_{ij} \ (i < j).$$

More precisely, the vector of coefficients $\beta$ (up to a scale factor of $10^{17}$) and the symmetric matrix $B$ (up to a scale factor of $10^{19}$) have the forms:

$$\beta = \begin{bmatrix} -2.586 \\ -129.6 \\ -1939 \\ 6922 \\ -896.1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & -1.0535 & 0 & 0 \\ 0 & 0 & 2.7965 & -1.2795 & 0 \\ -1.0535 & 2.7965 & 0 & 0 & -29.11 \\ 0 & -1.2795 & 0 & 0 & 10.245 \\ 0 & 0 & -29.11 & 10.245 & 0 \end{bmatrix}.$$  

In order to perform the surface response analysis for this model, we must bring it to the simplest expression, by finding first its normal form and then its canonical decomposition. Since these operations require inverting the matrix of second-order interactions, we need a preliminary calculation in order to determine its eigenvalues and corresponding orthonormal eigenvectors.

5.2 Eigenvalue analysis of the second-order interactions matrix

We recall that $\lambda_k, V_k \ (k = 1, \ldots, 5)$ are the eigenvalues and normalized eigenvectors of the matrix $B$ if they solve the systems of linear equations:

$$B \cdot V_k = \lambda_k V_k, \quad V_k' \cdot V_p = \delta_{kp},$$
with \( \delta_{ij} \) the Kronecker symbol, defined by \( \delta_{ij} = 1 \) if \( i = j \) and \( \delta_{ij} = 0 \) otherwise. Then the matrix \( B \) has the \textit{principal-value decomposition} (c.f. [48, Appendix §C])

\[
B = \sum_{k=1}^{5} \lambda_k V_k V_k'.
\]  

(5.2.1)

Since the matrix \( B \) has the form

\[
B = \begin{bmatrix}
0 & 0 & a & 0 & 0 \\
0 & b & c & 0 & 0 \\
a & b & 0 & 0 & d \\
0 & c & 0 & 0 & e \\
0 & d & e & 0 & 0
\end{bmatrix}, \quad a, b, c, d, e \in \mathbb{R},
\]

it follows from a general calculation that its eigenvalues are symmetric with respect to the origin: \( \lambda_1, 2 > 0, \lambda_3 = 0, \lambda_4 = -\lambda_2, \lambda_5 = -\lambda_1 \), so

\[
\lambda_1 > \lambda_2 > 0 > \lambda_4 > \lambda_5.
\]

More precisely, the eigenvalues of a matrix of this form are given by:

\[
\lambda_{1,5} = \pm \sqrt{\frac{s^2 + \sqrt{s^4 - 4p^2}}{2}}, \quad \lambda_{2,4} = \pm \sqrt{\frac{s^2 - \sqrt{s^4 - 4p^2}}{2}}, \quad \lambda_3 = 0,
\]

where \( s^2 = a^2 + b^2 + c^2 + d^2 + e^2 \) and \( p^2 = a^2(e^2 + c^2) + (be - cd)^2 \).

Indeed, upon computing numerically the eigenvalues (using the SAS RSREG procedure or Mathematica’s Eigensystem procedure), we arrive at

\[
\lambda_1 = -\lambda_5 = 31.0277 \times 10^{-19}, \quad \lambda_2 = -\lambda_4 = 0.446626 \times 10^{-19}, \quad \lambda_3 = 0,
\]

(5.2.2)

up to the software numerical precision.

Another general result is that the eigenvector corresponding to the null eigenvalue \( \lambda_3 = 0 \) has the form

\[
V_3' = \left( \frac{be - cd}{ac} x_5, -\frac{e}{c} x_5, 0, 0, x_5 \right), \quad x_5 \in \mathbb{R},
\]

that is to say its third and fourth entries are identically zero. Specifically for our model, the normalized eigenvector \( V_3 \) becomes

\[
V_3' = (0.619629, -0.778849, 0, 0, -0.0972326).
\]
The other four orthogonal and normalized eigenvectors are found to be

\[ V_1 = \begin{bmatrix} -0.022643 \\ 0.0697857 \\ 0.666881 \\ -0.235096 \\ -0.70329 \end{bmatrix}, \quad V_2 = \begin{bmatrix} -0.554542 \\ -0.437981 \\ 0.235096 \\ 0.666881 \\ -0.0256057 \end{bmatrix}, \]

\[ V_4 = \begin{bmatrix} 0.554542 \\ 0.437981 \\ 0.235096 \\ 0.666881 \\ 0.0256057 \end{bmatrix}, \quad V_5 = \begin{bmatrix} 0.022643 \\ -0.0697857 \\ -0.0256057 \\ 0.70329 \end{bmatrix}. \]

Since \( B \cdot V_3 = 0 \), it is useful to decompose the vector \( X \) into the component parallel to \( V_3, X_\parallel \), and the component perpendicular to \( V_3, X_\perp \):

\[ X = X_\parallel + X_\perp, \quad X_\parallel = (V_3' \cdot X)V_3, \quad X_\perp' \cdot X_\parallel = 0. \quad (5.2.3) \]

Then, we also have

\[ B \cdot X = B \cdot X_\perp, \quad X_\parallel = (0.619629x_1 - 0.778849x_2 - 0.0972326x_5)V_3, \quad (5.2.4) \]

so we conclude that the “neutral” component of \( X, X_\parallel \) (associated with the zero eigenvalue \( \lambda_3 \)), does not depend at all on the attributable variables \( x_3 \) and \( x_4 \), but only on the linear combination

\[ z_3 := V_3' \cdot X = 0.619629x_1 - 0.778849x_2 - 0.0972326x_5. \quad (5.2.5) \]

We will return to this important fact when discussing applications in the last section.

### 5.2.1 Canonical analysis of the quadratic model

Let \( B^- \) represent the symmetric generalized inverse of the matrix \( B \) ([48, Appendix §C])

\[ B^- = \sum_k \lambda_k^{-1} V_k V_k', \]

where the “primed” sum is taken only over non-zero eigenvalues (excluding \( \lambda_3 \) in our case). Then clearly from (5.2.4),

\[ B^- \cdot V_3 = 0, \quad B^- \cdot X_\parallel = 0. \]

52
Using the decomposition (5.2.3), the model (7.1.2) becomes
\[ Y = \beta_0 + (\beta' \cdot V_3)z_3 + \beta' \cdot X \perp + X' \perp \cdot B \cdot X \perp \]

In order to bring this expression to its normal form, we begin by shifting the variable \(X\) by a constant term
\[ \hat{X} = X + \frac{1}{2}B^{-} \cdot \beta. \]

**Remark 11** This transformation does not change the “parallel” component since
\[ V_3' \cdot \hat{X} = V_3' \cdot X, \quad V_3' \cdot B^{-} = 0. \]

We obtain the model
\[ Y = \beta_0 + (\beta' \cdot V_3)z_3 + \beta' \cdot \hat{X} \perp - \frac{1}{4}B^{-} \cdot \beta + \hat{X}' \perp \cdot B \cdot \hat{X} \perp \]
where we have used the property \( B^{-} \cdot B \cdot B^{-} = B^{-}. \) Since \( B \cdot B^{-} \hat{X} \perp = \hat{X} \perp, \)
\[ Y = \beta_0 - \frac{1}{4}B^{-} \cdot \beta + (\beta' \cdot V_3)z_3 + \hat{X}' \cdot B \cdot \hat{X}, \]
so we are now working with the normal quadratic form \( \hat{X}' \cdot B \cdot \hat{X}. \) Using again (7.1.3), the quadratic form \( \hat{X}' \cdot B \cdot \hat{X} \) becomes
\[ \hat{X}' \left( \sum_{k=1}^{5} \lambda_k V_k V_k' \right) \hat{X} = \sum_{k=1}^{5} \lambda_k (\hat{X}' V_k)(V_k' \hat{X}) = \sum_{k=1}^{5} \lambda_k |V_k' \cdot \hat{X}|^2 = \sum_{k=1}^{5} \lambda_k z_k^2, \]
where we have introduced the **canonical coordinates**
\[ z_k := V_k' \cdot \hat{X}, \quad k = 1, 2, \ldots, 5. \] (5.2.6)

We note that this coordinate change is consistent with (5.2.5) and Remark 11.

To conclude, we have the canonical form of the model
\[ Y = Y_0 + (\beta' \cdot V_3)z_3 + \lambda_1(z_1^2 - z_5^2) + \lambda_2(z_2^2 - z_4^2), \] (5.2.7)
and specifically for our data:
\[ Y - Y_0 = 186.47 \times 10^{-17} z_3 + 31.03 \times 10^{-19}(z_1^2 - z_5^2) + 0.45 \times 10^{-19}(z_2^2 - z_4^2). \]
In the following section we will determine the various types of confidence regions for pairs of variables for this model. As a preliminary step in this procedure, we must first find the *stationary point* of the model, defined generically as the point in attributable variables-space, where all the partial derivatives of the response variable $Y$, with respect to each independent variable, are simultaneously equal to zero (also known as the critical point or the zero-gradient point).

For the quadratic model (7.1.2), this condition becomes simply

$$\frac{\partial Y}{\partial x_k} = 0 \Rightarrow \beta' + 2X' \cdot B = 0 \Rightarrow B \cdot X = -\frac{1}{2} \beta$$

Using (5.2.3) and (5.2.4), the equation becomes

$$B \cdot X_\perp = -\frac{1}{2} \beta \Rightarrow X_\perp = -\frac{1}{2} B^{-1} \cdot \beta \Rightarrow \hat{X}_\perp = 0.$$  

Together with (7.1.5), this gives for the stationary point $z_1 = z_2 = z_4 = z_5 = 0$.

### 5.2.2 Confidence region shapes and conic sections

In order to distinguish between various types of shapes the confidence regions may have, we now specialize to a pair of variables $(z_i, z_j)$ from the normal quadratic form written in canonical variables, and impose the inequality

$$|Y - Y_0| \leq M, \quad M > 0,$$
leading to
\[ |\lambda_i z_i^2 + \lambda_j z_j^2| \leq M, \]
which defines the confidence region centered at \((0, 0)\). We find the following cases, corresponding to classes of conic sections:

**Extremum point, elliptical region: all eigenvalues have the same sign**

If \(\lambda_i, j\) are either all positive or all negative, the point \((0, 0)\) is a point of minimum or of maximum, respectively. The inequality becomes
\[
|\lambda_i| z_i^2 + |\lambda_j| z_j^2 \leq M \Rightarrow \frac{z_i^2}{M/|\lambda_i|} + \frac{z_j^2}{M/|\lambda_j|} \leq 1, \tag{5.2.8}
\]
which defines the interior of an ellipse of semiaxes \(\sqrt{M/|\lambda_i|}, \sqrt{M/|\lambda_j|}\) (see Figure 13, right panel). The confidence region is given parametrically by:
\[
z_i = \sqrt{\frac{M}{|\lambda_i|}} r \cos(\theta), \quad z_j = \sqrt{\frac{M}{|\lambda_j|}} r \sin(\theta), \quad 0 \leq r \leq 1, \quad \theta \in [0, 2\pi]. \tag{5.2.9}
\]

**Figure 14.** Confidence regions for the hyperbolic case.
Saddle-point, hyperbolic region: non-zero eigenvalues of different signs

If, say, $\lambda_i > 0$ and $\lambda_j < 0$, then $(0, 0)$ is a saddle point, and the inequality becomes

$$-M \leq |\lambda_i| z_i^2 - |\lambda_j| z_j^2 \leq M,$$

which defines the set of orthogonal hyperbolas (see Figure 14)

$$\frac{z_i^2}{M/|\lambda_i|} - \frac{z_j^2}{M/|\lambda_j|} \leq 1, \quad \frac{z_j^2}{M/|\lambda_j|} - \frac{z_i^2}{M/|\lambda_i|} \leq 1. \quad (5.2.10)$$

The intersection of these conditions defines a region that looks like an elongated rectangle (elongated “corners”, the domain defined by the blue and green curves in Figure 14) and can be approximated with a rectangular shape. The confidence region is given parametrically by:

$$z_i = \sqrt{\frac{M}{|\lambda_i|}} r \cosh(t), \quad z_j = \sqrt{\frac{M}{|\lambda_j|}} r \sinh(t), \quad -1 \leq r \leq 1, \quad t \in \mathbb{R}. \quad (5.2.11)$$

Flatness point, conical region: some eigenvalues are zero (degenerate)

Let now $\lambda_j \to 0$ in the previous case, and we obtain a point of “flatness” or degenerate point, where the inequality becomes

$$\lambda_i z_i^2 \leq M, \quad z_j \in \mathbb{R},$$

which corresponds to the conical degeneration of a hyperbolic region (the domain defined by the blue and red curves in Figure 14), i.e. an infinite strip domain, Figure 15:

$$|z_i| \leq \sqrt{\frac{M}{|\lambda_i|}} \quad (5.2.12)$$

5.2.3 Specific analysis for the atmospheric CO$_2$ quadratic model

In the case under discussion, since one of the eigenvalues have all possible values ($\lambda_{1,2} > 0, \lambda_{4,5} < 0, \lambda_3 = 0$), we obtain all possible cases for the confidence regions:

- the confidence region in any of $z_{1,2,4,5}$ and $z_3$ is flat (degenerate) shape;
- the confidence region in ($z_{1,2}$ and $z_{4,5}$) is of hyperbolic (rectangular) shape;
- the confidence region in any pair ($z_i, z_j$), $i, j = 1, 2$ or $i, j = 4, 5$ is of elliptical (round) shape.
5.3 Applications: predictions based on nonlinear analysis

5.3.1 Order of magnitude analysis for the canonical variables

We begin addressing the list of applications described in the Introduction by first providing numerical estimates for each term in the specific model derived earlier:

\[ Y - Y_0 = 186.47 \times 10^{-17} z_3 + 31.03 \times 10^{-19} (z_1^2 - z_5^2) + 0.45 \times 10^{-19} (z_2^2 - z_4^2). \]

From the data, as well as from the estimate of the response variable at the stationary point, we obtain (at order of magnitude) the following estimate for \( Y_0 \):

\[ Y_0 \sim (\text{CO}_2)^{-2.376} \sim O(10^{-7}) \]

Using the IPCC recommendation for CO₂ emissions reduction [47], of 20% – 30% through 2020, we obtain an annual variation of the order of 3%, which means a variation of the order of 10% for the response variable \( Y \) (note that increasing CO₂ corresponds to decreasing \( Y \)). Therefore, it is reasonable and relevant to work with variations of the order \( M \sim 0.1 \times Y_0 = O(10^{-8}) \).

At this order of magnitude, a simple estimate for the variation of the canonical variable \( z_3 \) gives us the value of \( |z_3| \sim 10^{15} \times M = O(10^7) \). By contrast, applying the formulas (7.1.9), (7.1.20),
(5.2.12) and the numerical values for $\lambda_{1,2}$, we obtain the order of magnitude

$$|z_k| \sim O\left(\sqrt{\frac{M}{|\lambda_k|}}\right) \sim O(10^5), \quad k = 1, 2, 4, 5.$$ 

This indicates that, while the canonical variable $z_3$ (5.2.5) may be allowed to fluctuate up to order $10^7$ around the origin, the other canonical variables are much more restricted, by up to 2 orders of magnitude less. Since the variable $z_3$ does not contain any contribution from the attributable variables $x_3, x_4$, this analysis singles them out in a two-fold way:

(i) their variation (no matter how small) always contributes to the quadratic part of the response variable, and

(ii) the order of magnitude allowed for their variations, at given threshold $M \sim 0.1 \times Y_0$, is about 100 times smaller than what is allowed for the linear combination $z_3$ (5.2.5).

Therefore, we arrive at the following conclusion with direct practical applications:

**Example 7** For variations of the CO$_2$ levels at the order of magnitude stipulated by IPCC (around 2% per year), the linear combination of attributable variables $z_3 = 0.619629x_1 - 0.778849x_2 - 0.0972326x_5$ can be considered to be basically “free” compared to the other canonical variables, i.e. it may have fluctuations of up to order $O(10^6)$ without having a significant effect on the CO$_2$ levels.

Moreover, we can estimate the order of magnitude of $M$ at which the variable $z_3$ stops being “free” with respect to the other variables, from the simple comparison

$$10^{15} \times M \sim O\left(\sqrt{\frac{M}{|\lambda_k|}}\right) = 10^9 \times \sqrt{M} \Rightarrow M \sim 10^{-12} \Rightarrow \frac{M}{Y_0} \sim 10^{-5} = 0.001\%$$

In other words, unless we are concerned with yearly variations of the CO$_2$ levels not exceeding 0.001% of the current levels (an accuracy not realistic for our present measurement and prediction capabilities), Conclusion 1 holds.

### 5.3.2 Managing CO$_2$ emissions: accountability policies and metrics

Throughout this subsection, we let the values of the attributable variables $X' = (x_1, x_2, x_3, x_4, x_5)$ be measured from the stationary point $X_s = -\frac{1}{2} B^- \cdot \beta$. In other words, instead of $x_1$ we use the shifted value $x_1 + \left(\frac{1}{2} B^- \cdot \beta\right)_1$, instead of $x_2$ we use the shifted value $x_2 + \left(\frac{1}{2} B^- \cdot \beta\right)_2$, etc.
Starting from the model

\[ Y - Y_0 = 186.47 \times 10^{-17} z_3 + 31.03 \times 10^{-19} (z_1^2 - z_5^2) + 0.45 \times 10^{-19} (z_2^2 - z_4^2), \]

and the defining relations for the linear combinations

\[
\begin{align*}
    z_1 &= -0.022643 x_1 + 0.0697857 x_2 + 0.666881 x_3 - 0.235096 x_4 - 0.70329 x_5, \\
    z_2 &= -0.554542 x_1 - 0.437981 x_2 + 0.235096 x_3 + 0.666881 x_4 - 0.0256057 x_5, \\
    z_4 &= 0.554542 x_1 + 0.437981 x_2 + 0.235096 x_3 + 0.666881 x_4 + 0.0256057 x_5, \\
    z_5 &= 0.022643 x_1 - 0.0697857 x_2 + 0.666881 x_3 - 0.235096 x_4 + 0.70329 x_5,
\end{align*}
\]

and using Conclusion 1 (which allows to neglect the term proportional to \( z_3 \) from the model), we arrive at the following equation:

\[ M \simeq 124.12 \times 10^{-19} u_1 v_1 + 1.8 \times 10^{-19} u_2 v_2, \]

where

\[
\begin{align*}
    u_1 &= 0.666881 x_3 - 0.235096 x_4, \\
    u_2 &= 0.235096 x_3 + 0.666881 x_4, \\
    v_1 &= -0.022643 x_1 + 0.0697857 x_2 - 0.70329 x_5, \\
    v_2 &= -0.554542 x_1 - 0.437981 x_2 - 0.0256057 x_5,
\end{align*}
\]

which together with \( z_4 \) form a new set of orthogonal coordinates in \( \mathbb{R}^5 \) (just like \( \{ x_k \} \) and \( \{ z_k \} \)).

In order to implement a constraint at given value of \( M \), we may choose to set either the product \( u_1 v_1 = 0 \) or \( u_2 v_2 = 0 \), and solve for the remaining term. This choice will provide a direct procedure for comparing the relative weight of one attributable variable versus another.

**Example 8** We choose to set \( u_1 = 0 \), which leads to the conclusion that a variation of 1 unit in the attributable variable \( x_3 \) is offset by a variation of \( 0.666881/0.235096 \approx 2.837 \) units in the variable \( x_4 \). The new variable \( u_2 \) now becomes

\[ u_2 = (0.235096 + 0.666881 \cdot 2.837) x_3 \approx 2.127 x_3 \]

Choosing \( M \sim 10^{-8} \) again, we obtain the inequality

\[ |u_2 v_2| \leq 5.5 \times 10^{10} \Rightarrow |v_2| \leq \frac{2.586}{|x_3|} \times 10^{10}. \]
Specifically, consider the situation where we wish to increase the value of $x_3$ (Gas Flares) by $10^3$. From the previous analysis, in order for the total emissions not to exceed 2% of yearly average ($M \sim 10^{-8}$), an increase of 1000 units in Gas Flares may be accompanied by an increase in $x_4$ (Cement) of 2127 units, while the linear combination $v_2$ must satisfy

$$|0.554542x_1 + 0.437981x_2 + 0.0256057x_5| \leq 25.86 \times 10^6.$$ 

The values of the linear combinations $z_3, v_1$ remain arbitrary in this case:

$$v_1 = -0.022643x_1 + 0.0697857x_2 - 0.70329x_5 \in \mathbb{R},$$

$$z_3 = 0.619629x_1 - 0.778849x_2 - 0.0972326x_5 \in \mathbb{R}.$$ 

### 5.4 Predicting large deviations in the CO$_2$ levels using Algorithm I

In order to estimate the probability of large deviations of CO$_2$ levels, recall that the response variable, $Y$, and the CO$_2$ level (denoted $C$ for convenience) are related by the power-law transformation $Y = C^{-\gamma}, \gamma = 2.376$.

For a given threshold $M \geq C_0$, the large deviation $C > M$ means $(C - C_0)/C_0 > (M - C_0)/C_0 := r$. Assume that $r \leq 1$, then the inequality becomes after some manipulations

$$\frac{Y_0 - Y}{Y_0} \geq 1 - \left(\frac{1}{1 + r}\right)^{\frac{1}{\gamma}} \simeq \frac{r}{\gamma} \left[1 - \frac{r}{2\gamma}\right] \left(1 - r + \frac{r^2}{4}\right) + O((r/\gamma)^3) \quad (5.4.1)$$

Because the combination $r/\gamma$ is small, we can further approximate (at order $O((r/\gamma)^3)$) that the deviations in $C$ correspond to the deviations in $Y$ (relative to $Y_0$) of the order $r \frac{2}{\gamma} \left[1 - \frac{r}{2\gamma}\right]$. This allows to map between relative variations in $C$ and $Y$. However, because of the model symmetry, $(Y - Y_0)/Y \simeq \tilde{X}' \cdot H(\log Y) \cdot \tilde{X}$ ($H$ is the Hessian matrix, evaluated at the stationary point $Y_0$), calculated based on the empirical observations, can be considered as a sample of i.i.d. r.v. Therefore, applying the algorithm of Chapter 3 and mapping back using (5.4.1), we can compute Table 2.

<table>
<thead>
<tr>
<th>$(M - C)/C_0$</th>
<th>2%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(C &gt; M)$</td>
<td>0.73</td>
<td>0.48</td>
<td>0.02</td>
<td>0.007</td>
</tr>
</tbody>
</table>
Chapter 6

Estimating the atmospheric CO$_2$ levels - EU data and comparison to US model

Continuing the analysis performed in the previous chapter regarding the atmospheric CO$_2$ modeling for the US data, we now repeat the model building study for the EU case.

The analysis presented here consists of two steps: first, we partially replicate the comprehensive study performed in [45, 46] in order to select the relevant variables and their interactions, and then we validate and analyze the best second-order model. Based on this result, we will perform the surface-response analysis (curvature effects) of the model in the next section.

The following 23 EU countries$^1$ were used in this study$^2$: Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Spain, Sweden, United Kingdom.

The aim of this study is two-fold: first, we focus on the determination of the optimal second-order model (including interactions as well as quadratic terms) which provides the best fit for the data and features robust validation; second, we will use the model derived here in order to determine multi-variables confidence regions and obtain estimates for the allowed fluctuations in attributable variables, given a maximal range of change in the response variable. The second-order model we will obtain in the current study provides us with a list of relevant variables, ranked according to their statistical significance (quantified by the respective contribution to the total variability), which we employ in order to compare with the similar result obtained in the case of the U.S. data [45, 46].

---

$^1$As of June 2013, there are 27 member states of the EU. Slovenia and the Baltic states were excluded from the present study since there was no individual data available for the period during which they were part of former Yugoslavia, and former Soviet Union, respectively. However, their contribution to the CO$_2$ emissions is relatively small, as it can be seen from Figure 17.

$^2$Figure 17 was reproduced under the Creative Commons Attribution-Share Alike 3.0 Unported license.
Figure 16: Schematic representation of the carbon cycle (DOE report [47] and supporting documentation).

Figure 17: EU CO₂ emissions, in thousands of metric tons [50].
6.1 Regression analysis and model building

As explained in the previous chapter, the dependent variable must be transformed in order to ensure normality of the data, by means of the Box-Cox transformation. After the Box-Cox filter, we retest the data and it shows our data will follow normal distribution; thus, we proceed to estimate the coefficients of the contributable variables for the transformed CO\textsubscript{2} atmosphere data. The parameter of the transformation is the same as the one used in the two previous studies [45, 46].

We can proceed to estimate the approximate coefficients of the contributable variables for transformed CO\textsubscript{2} in the atmosphere and obtain the coefficients of all possible interactions, using the multivariate regression procedure and corresponding goodness-of-fit measures.

We begin with seven attributable variables and apply the SAS stepwise forward selection procedure to a model with six relevant variables and 21 second-order interactions between each pair and self-interactions. Introducing the notation \(x_1 = \text{Liquid Fuels} (\text{Li}), x_2 = \text{Gas Fuels} (\text{Ga}), x_3 = \text{Gas Flares} (\text{Fl}), x_4 = \text{Bunker} (\text{Bu})\) for the relevant attributable variables, we find that only three of the variables (Liquid Fuels \(x_1\), Gas Fuels \(x_2\), and Gas Flares \(x_3\)) and only 3 interactions (Ga:Bu, Li:Fl, Li:Bu) and two quadratic terms (Li\(^2\), Bu\(^2\)) are statistically relevant at \(\alpha = 0.01\) level.

The result of estimation becomes the quadratic model with interactions:

\[
\hat{CO_2}^{-2.376} = 0.00000123 + (710.85Fl - 30.64Ga - 3.4501Li) \times 10^{-13} +
(37.34Ga \cdot Bu + 1.35Li \cdot Li - 65.12Bu \cdot Bu -
133.05Li \cdot Fl - 5.35Li \cdot Bu) \times 10^{-18}.
\]

The quality of the fit for this quadratic model is evidenced by high value of both \(R^2\) and \(R^2_{\text{adjusted}}\) which are the key criteria to evaluate the model fitting. In terms of the total (\(SS_t\)), regression (\(SS_r\), and error (\(SS_e\)) sums of squares, we have the standard formulas

\[
R^2 = \frac{SS_r}{SS_t}, \quad R^2_{\text{adjusted}} = 1 - \frac{SS_e/df_e}{SS_t/df_t},
\]

with \(df_{e,t}\) the degrees of freedom for the chi-squared distributions for error and total, respectively [48]. We also employ the prediction of residual error sum of squares (PRESS) statistics which will evaluate how good the estimation will be if each time we remove one data point. If the index \(i\)
covers all the observations and \( \hat{y}_{(i)} \) is the predicted value when the observation is omitted, then

\[
PRESS = \sum_{i} (y_i - \hat{y}_{(i)})^2.
\]

Table 4: Statistical evaluation criteria for model (6.1.1)

<table>
<thead>
<tr>
<th></th>
<th>( R^2 )</th>
<th>( R_{\text{adj}}^2 )</th>
<th>( PRESS )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9979</td>
<td>0.9975</td>
<td>1.636 \times 10^{-20}</td>
</tr>
</tbody>
</table>

According to these goodness-of-fit measures, the model we have obtained is high quality and reliable for predictions. The model becomes:

\[
\left[\hat{C}O_2\right]^{-2.376} = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i \leq j=1}^{4} \beta_{ij} x_i x_j,
\]

with the corresponding ranks determined by the stepwise SAS procedure are given in Table 5, along with the coefficients in the final regression model.

Therefore, we can write our model in matrix notation (where prime denotes transposition) as:

\[
Y = \beta_0 + \beta' \cdot X + X' \cdot B \cdot X,
\]

(6.1.2)

with the obvious identifications

\[
X' = (x_1, \ldots, x_4), \: \beta' = (\beta_1, \ldots, \beta_4), \: B_{ij} = B_{ji} = \frac{1}{2} \beta_{ij} \: (i < j).
\]

More precisely, the vector \( \beta \) (up to an overall scale factor of \( 10^{-13} \)), and the symmetric matrix \( B \) (up to an overall scale factor of \( 10^{-18} \)) have the forms:

\[
\beta = \begin{bmatrix}
-3.4501 \\
-30.635 \\
710.848 \\
0
\end{bmatrix}, \quad B = \begin{bmatrix}
2.7113 & 0 & -133.05 & 0 \\
0 & 0 & 0 & 37.3391 \\
-133.05 & 0 & 0 & 0 \\
0 & 37.3391 & 0 & -130.23
\end{bmatrix}
\]

These results will be used to perform the surface response analysis for the model, in the next chapter. The matrix formulation of the quadratic expression (6.1.1) will be the starting point for finding its canonical decomposition, based on the eigenvalue analysis.
Table 5: Ranking by statistical relevance for attributable variables and interactions.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variable</th>
<th>Name</th>
<th>$\beta \times 10^{-18}$</th>
<th>Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ga ($x_2$)</td>
<td>Gas Fuels</td>
<td>$-30.635 \times 10^5$</td>
<td>48.72%</td>
</tr>
<tr>
<td>2</td>
<td>Ga:Bu ($x_2 \times x_4$)</td>
<td>Gas $\times$ Bunker</td>
<td>37.3391</td>
<td>12.41%</td>
</tr>
<tr>
<td>3</td>
<td>Li$^2$ ($x_1^2$)</td>
<td>Liquid $\times$ Liquid</td>
<td>1.35565</td>
<td>11.79%</td>
</tr>
<tr>
<td>4</td>
<td>Bu$^2$ ($x_3^2$)</td>
<td>Bunker</td>
<td>$-65.115$</td>
<td>7.78%</td>
</tr>
<tr>
<td>5</td>
<td>Fl ($x_3$)</td>
<td>Gas Flares</td>
<td>$710.848 \times 10^5$</td>
<td>6.66%</td>
</tr>
<tr>
<td>6</td>
<td>Li:Fl ($x_1 \times x_3$)</td>
<td>Liquid $\times$ Flares</td>
<td>$-133.05$</td>
<td>5.06%</td>
</tr>
<tr>
<td>7</td>
<td>Li:Bu ($x_1 \times x_4$)</td>
<td>Liquid $\times$ Bunker</td>
<td>$-5.3501$</td>
<td>4.71%</td>
</tr>
<tr>
<td>8</td>
<td>Li ($x_1$)</td>
<td>Liquid Fuels</td>
<td>$-3.4501 \times 10^5$</td>
<td>2.86%</td>
</tr>
</tbody>
</table>

6.1.1 Validation of the fitted model

Based on the standard diagnostics provided by the SAS regression procedure, we can quantify the reliability and accuracy of the model (6.1.1). Specifically, the normalized predicted value for residuals, residual quantile plot, and the predicted value for the dependent variable, all indicate that the model is accurately describing the total variability in the data, and that the canonical normality assumptions are satisfied.

To further assess the robustness of (6.1.1), we employed multiple cross-validation by partitioning the full data set into smaller, equal-sized sets, and then fitting the model using all but one of the subsets. The predicted values for the missing observations (from the subset removed from the analysis) are then computed and quantified using their mean value and dispersion.

Specifically, we divided the data set into 49 data sets and use all 48 sets to construct the model and validate the model using the one left out. Then we repeat the procedure 48 times. The mean of the residuals is $5.427845e-22$ and the variance of the residuals is $2.806e-44$.

6.2 Comparing the US and EU models

Using the results obtained in [46], it is possible develop a comparison between the US and EU quadratic models for attributable variables and interactions; in particular, it is possible to compare the relative relevance of the main single-factor variables and of the main interactions (see Table 6
below and Table 3.4 from [45]).

### Table 6: Comparison of statistical relevance for attributable variables and interactions, US vs. EU.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variable in US</th>
<th>Variable in EU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Liquid</td>
<td>Gas</td>
</tr>
<tr>
<td>2</td>
<td>Liquid:Cement</td>
<td>Gas:Bunker</td>
</tr>
<tr>
<td>3</td>
<td>Cement:Bunker</td>
<td>Liquid:Liquid</td>
</tr>
<tr>
<td>4</td>
<td>Bunker</td>
<td>Bunker:Bunker</td>
</tr>
<tr>
<td>5</td>
<td>Cement</td>
<td>Gas Flares</td>
</tr>
<tr>
<td>6</td>
<td>Gas Flares</td>
<td>Liquid:Gas Flares</td>
</tr>
<tr>
<td>7</td>
<td>Gas</td>
<td>Liquid:Bunker</td>
</tr>
<tr>
<td>8</td>
<td>Gas:Gas Flares</td>
<td>Liquid</td>
</tr>
</tbody>
</table>

6.2.1 Specific features of CO$_2$ contributors for EU data

As Table 6 indicates, the most significant risk factor found when studying the US data (Liquid fuels) is the least relevant in the case of the EU data; likewise, the main attributable variable obtained for the EU data (Gas fuels) is the least relevant single-factor in the case of the US data (rank 7 out of 8). In view of the importance of the variable Gas fuel for the EU data, we indicate the individual contributions at country-level, for the year 2008, in Table 7.

This important difference between the rankings of risk factors for EU and US suggests two directions for further comparisons:

i) for the purpose of CO$_2$ emission reductions, a very different picture emerges from the EU study versus the US study. Consider the information shown on Tables 5, 6 and 7: almost 50% of the total variability in the atmospheric CO$_2$ is due to Gas Fuels alone, which ranks first among the risk factors; furthermore, more than 25% of the total emissions for this factor is due to a single country, Germany (for year 2008).

ii) when developing criteria for trade-off of single risk factors (as in “cap-and-trade” schemes and beyond), specific information from each continent must be employed, as discussed in the
next section. While this point further emphasizes the regional aspect of CO$_2$ analysis, it also illustrates the inherent limitations of applying carbon-accountability policies from one continent to another.

**Table 7:** Contributions by country to gas fuel emissions, for year 2008 (in thousand metric tons of carbon).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Total</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GERMANY</td>
<td>85457</td>
<td>27.19%</td>
</tr>
<tr>
<td>2</td>
<td>POLAND</td>
<td>58395</td>
<td>18.58%</td>
</tr>
<tr>
<td>3</td>
<td>UK</td>
<td>37306</td>
<td>11.87%</td>
</tr>
<tr>
<td>4</td>
<td>CZECH REP.</td>
<td>21021</td>
<td>6.69%</td>
</tr>
<tr>
<td>5</td>
<td>ITALY</td>
<td>16855</td>
<td>5.36%</td>
</tr>
<tr>
<td>6</td>
<td>SPAIN</td>
<td>14669</td>
<td>4.67%</td>
</tr>
<tr>
<td>7</td>
<td>FRANCE</td>
<td>13383</td>
<td>4.26%</td>
</tr>
<tr>
<td>8</td>
<td>ROMANIA</td>
<td>9805</td>
<td>3.12%</td>
</tr>
<tr>
<td>9</td>
<td>GREECE</td>
<td>8946</td>
<td>2.85%</td>
</tr>
<tr>
<td>10</td>
<td>BULGARIA</td>
<td>8058</td>
<td>2.56%</td>
</tr>
<tr>
<td>11</td>
<td>NETHERLANDS</td>
<td>7346</td>
<td>2.34%</td>
</tr>
<tr>
<td>12</td>
<td>FINLAND</td>
<td>5478</td>
<td>1.74%</td>
</tr>
<tr>
<td>13</td>
<td>BELGIUM</td>
<td>4361</td>
<td>1.39%</td>
</tr>
<tr>
<td>14</td>
<td>SLOVAKIA</td>
<td>4205</td>
<td>1.34%</td>
</tr>
<tr>
<td>15</td>
<td>DENMARK</td>
<td>4062</td>
<td>1.29%</td>
</tr>
<tr>
<td>16</td>
<td>AUSTRIA</td>
<td>3896</td>
<td>1.24%</td>
</tr>
<tr>
<td>17</td>
<td>HUNGARY</td>
<td>3275</td>
<td>1.04%</td>
</tr>
<tr>
<td>18</td>
<td>PORTUGAL</td>
<td>2644</td>
<td>0.84%</td>
</tr>
<tr>
<td>19</td>
<td>SWEDEN</td>
<td>2530</td>
<td>0.80%</td>
</tr>
<tr>
<td>20</td>
<td>IRELAND</td>
<td>2519</td>
<td>0.80%</td>
</tr>
<tr>
<td>21</td>
<td>LUXEMBOURG</td>
<td>86</td>
<td>0.03%</td>
</tr>
<tr>
<td>22</td>
<td>CYPRUS</td>
<td>29</td>
<td>0.01%</td>
</tr>
<tr>
<td>23</td>
<td>MALTA</td>
<td>0</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
6.2.2 Regionalization of atmospheric CO$_2$ analysis

One relevant variable for the US market (Cement, together with its interactions to Liquid and Bunker) is completely absent in the case of the EU data. Together with the fact that, at the same level of statistical significance, the US model requires five variables while the EU data leads to a model with four, this shows that the two models are fundamentally different, in the sense that production, dynamics, and global interactions of the man-made factors responsible for atmospheric CO$_2$, are essentially different in the case of EU and UE.

This indicates that, when developing specific guidelines for industry regulations and accountability criteria, translating regulations and policies between the EU and US markets must be done with considerable caution. The differences found here mandate, in fact, that a specific approach must be developed in the EU case, and that adopting US-based policy directly may be unwarranted and outright misguided. We will quantify this aspect of the comparison in a forthcoming publication, aimed at comparing the possible carbon-production management in the case of the US versus the EU markets.
In the current chapter, we start from the second-order model developed in the previous chapter and perform its surface reponse analysis, leading to canonical two-dimensional confidence regions, and to specific comparisons between canonical variables, much as it was done in [46], in the case of US data.

### 7.1 The model, parameters, and descriptive quantities

We recall the final second-order model found earlier to provide a good fit for the data and to have robust features for prediction and estimation:

\[
[C\bar{O}_2]^{-2.376} = 0.00000123 + (710.85Fl - 30.64Ga - 3.4501Li) \times 10^{-13} + (37.34Ga \cdot Bu + 1.35 Li \cdot Li - 65.12 Bu \cdot Bu - 133.05 Li \cdot Fl - 5.35 Li \cdot Bu) \times 10^{-18}.
\]

Using the notation \( x_1 = \) Liquid Fuels (Li), \( x_2 = \) Gas Fuels (Ga), \( x_3 = \) Gas Flares (Fl), \( x_4 = \) Bunker (Bu) for the relevant attributable variables, the model becomes

\[
[C\bar{O}_2]^{-2.376} = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i \leq j=1}^{4} \beta_{ij} x_i x_j,
\]

or in matrix notation (where prime denotes transposition), (7.1.1) becomes

\[
Y = \beta_0 + \beta' \cdot X + X' \cdot B \cdot X,
\]

with the obvious identifications
More precisely, the vector $\beta$ (up to an overall scale factor of $10^{-13}$), and the symmetric matrix $B$ (up to an overall scale factor of $10^{-18}$) have the forms:

$$\beta = \begin{bmatrix} -3.4501 \\ -30.635 \\ 710.848 \\ 0 \end{bmatrix}, \quad B = \begin{bmatrix} 2.7113 & 0 & -133.05 & 0 \\ 0 & 0 & 0 & 37.3391 \\ -133.05 & 0 & 0 & 0 \\ 0 & 37.3391 & 0 & -130.23 \end{bmatrix}$$

In order to perform the surface response analysis for this model, we must bring it to the simplest expression, by finding first its normal form and then its canonical decomposition. Since these operations require inverting the matrix of second-order interactions, we need to perform a preliminary calculation in order to determine its eigenvalues and corresponding orthonormal eigenvectors.

### 7.1.1 Eigenvalue analysis of the second-order interactions matrix

We recall that $\lambda_k, V_k$ ($k = 1, \ldots, 4$) are the eigenvalues and normalized eigenvectors of the matrix $B$ if they solve the systems of linear equations:

$$B \cdot V_k = \lambda_k V_k, \quad V_k' \cdot V_p = \delta_{kp},$$

with $\delta_{ij}$ the Kronecker symbol, defined by $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. Then the matrix $B$ has the principal-value decomposition (c.f. [48, Appendix §C])

$$B = \sum_{k=1}^{4} \lambda_k V_k V_k'.$$  \hfill (7.1.3)

For the matrix $B$ found above, upon computing numerically the eigenvalues (using the SAS RSREG procedure or Mathematica’s Eigensystem procedure), we arrive at

$$\lambda_1 = -140.176, \; \lambda_2 = 134.413, \; \lambda_3 = -131.701, \; \lambda_4 = 9.94612,$$  \hfill (7.1.4)

up to the software numerical precision and the overall scale factor $10^{-18}$. 
The four orthogonal and normalized eigenvectors are found to be

\[
V_1 = \begin{bmatrix}
0 \\
-0.257397 \\
0 \\
0.966306
\end{bmatrix}, \quad V_2 = \begin{bmatrix}
0.7107 \\
0 \\
-0.703495 \\
0
\end{bmatrix}, \\
V_3 = \begin{bmatrix}
-0.703495 \\
0 \\
-0.7107 \\
0
\end{bmatrix}, \quad V_4 = \begin{bmatrix}
0 \\
-0.966306 \\
0 \\
-0.257397
\end{bmatrix}.
\]

7.1.2 Canonical analysis of the quadratic model

Let \( B^{-1} \) represent the inverse of the matrix \( B \) ([48, Appendix §C])

\[
B^{-1} = \sum_{k=1}^{4} \lambda_k^{-1} V_k V_k',
\]

and start from the model (7.1.2)

\[
Y = \beta_0 + \beta' \cdot X + X' \cdot B \cdot X.
\]

In order to bring this expression to its normal form, we begin by shifting the variable \( X \) by a constant term

\[
\hat{X} = X + \frac{1}{2} B^{-1} \cdot \beta.
\]

Since \( B \) is a non-singular matrix, we obtain the model

\[
Y = \beta_0 + \beta' \cdot \hat{X} - \frac{1}{4} \beta' \cdot B^{-1} \cdot \beta + \hat{X}' \cdot B \cdot \hat{X} - \beta' \cdot B \cdot B^{-1} \cdot \hat{X},
\]

where we have used the property \( B^{-1} \cdot B = I \). Therefore,

\[
Y = \beta_0 - \frac{1}{4} \beta' \cdot B^{-1} \cdot \beta + \hat{X}' \cdot B \cdot \hat{X},
\]

so we are now working with the normal quadratic form \( \hat{X}' \cdot B \cdot \hat{X} \). Using again (7.1.3), the quadratic form \( \hat{X}' \cdot B \cdot \hat{X} \) becomes

\[
\hat{X}' \left( \sum_{k=1}^{4} \lambda_k V_k V_k' \right) \hat{X} = \sum_{k=1}^{4} \lambda_k (\hat{X}' V_k)(V_k' \hat{X}) = \sum_{k=1}^{4} \lambda_k |V_k' \cdot \hat{X}|^2 = \sum_{k=1}^{4} \lambda_k z_k^2,
\]
where we have introduced the canonical coordinates

$$z_k := V_k' \cdot \hat{X}, \quad k = 1, 2, 3, 4.$$  \hspace{1cm} (7.1.5)

To conclude, we have the canonical form of the model

$$Y - Y_0 = \left( -140.176z_1^2 + 134.413z_2^2 - 131.701z_3^2 + 9.94612z_4^2 \right) \times 10^{-18},$$  \hspace{1cm} (7.1.6)

with $z_k$ given in (7.1.5).

To find the stationary point of the model, defined generically as the zero-gradient point, we must solve simultaneously for all $k = 1, \ldots, 4$:

$$\frac{\partial Y}{\partial x_k} = 0 \Rightarrow \beta' + 2X' \cdot B = 0 \Rightarrow B \cdot X = -\frac{1}{2}\beta,$$

which is equivalent to

$$B \cdot \hat{X} = 0 \Rightarrow \hat{X} = 0,$$

because $B$ is non-degenerate. Together with (7.1.5), this gives the stationary point as the origin of the $z$ coordinates, $z_1 = z_2 = z_3 = z_4 = 0$. In the original variables, the stationary point is found to be:

$$X_s = \frac{1}{2}B^{-1} \cdot \beta = \begin{bmatrix} -534271 \\ -286155 \\ -8294.32 \\ -82045.4 \end{bmatrix},$$  \hspace{1cm} (7.1.7)

up to an overall scale factor of $10^5$.

### 7.1.3 Confidence region shapes and conic sections

We repeat here the discussion regarding confidence region types presented in [46]. In order to distinguish between various types of shapes the confidence regions may have, we now specialize to a pair of variables $(z_i, z_j)$ from the normal quadratic form written in canonical variables, and impose the inequality

$$|Y - Y_0| \leq M, \quad M > 0,$$

leading to

$$\left| \lambda_i z_i^2 + \lambda_j z_j^2 \right| \leq M,$$
which defines the confidence region centered at \((0, 0)\). We find the following cases, corresponding to classes of conic sections:

**Extremum point, elliptical region: all eigenvalues have the same sign**

If \(\lambda_{i,j}\) are either all positive or all negative, the point \((0, 0)\) is a point of minimum or of maximum, respectively. The inequality becomes

\[
|\lambda_i|z_i^2 + |\lambda_j|z_j^2 \leq M \Rightarrow \frac{z_i^2}{M/|\lambda_i|} + \frac{z_j^2}{M/|\lambda_j|} \leq 1, \tag{7.1.8}
\]

which defines the interior of an ellipse of semiaxes \(\sqrt{M/|\lambda_i|}, \sqrt{M/|\lambda_j|}\). The confidence region is given parametrically by:

\[
z_i = \sqrt{\frac{M}{|\lambda_i|}} r \cos(\theta), \quad z_j = \sqrt{\frac{M}{|\lambda_j|}} r \sin(\theta), \quad 0 \leq r \leq 1, \quad \theta \in [0, 2\pi]. \tag{7.1.9}
\]

This is applicable for any pair of eigenvalues from \(\{\lambda_1, \lambda_3\}\) or from \(\{\lambda_2, \lambda_4\}\).

**Example 9** Determining specific numerical regions for CO\(_2\) fluctuations at levels discussed by IPCC [47].

In order to maintain consistency in comparing the models obtained for US [46] versus EU (this work), we compute the parameters of elliptical confidence regions for variables \(z_1, z_3\) and \(z_2, z_4\), corresponding to yearly CO\(_2\) level fluctuations of 3\% (see the discussion in [46, §4.1] and supporting documentation in [47]). As shown in [46, §4.1], this range of values corresponds to the order of magnitude \(M \sim 10^{-8}\), so we arrive at the equations

\[
0.7107\hat{X}_1 - 0.703495\hat{X}_3 = 0, \quad -0.966306\hat{X}_2 - 0.257397\hat{X}_4 = 0, \tag{7.1.10}
\]

\[
-0.257397\hat{X}_2 + 0.966306\hat{X}_4 = 8446.24 \cdot r \cos(\theta), \tag{7.1.11}
\]

\[
-0.703495\hat{X}_1 - 0.7107\hat{X}_3 = 8713.76 \cdot r \sin(\theta), \tag{7.1.12}
\]

which give the solution

\[
x_1 = 534271 - 6130.09 \cdot r \sin(\theta), \tag{7.1.13}
\]

\[
x_2 = 286155 - 2174.03 \cdot r \cos(\theta), \tag{7.1.14}
\]

\[
x_3 = 8294.32 - 6192.87 \cdot r \sin(\theta), \tag{7.1.15}
\]
\[ x_4 = 82045.4 + 8161.64 \cdot r \cos(\theta), \] (7.1.16)

where \(0 \leq r \leq 1, \theta \in [0, 2\pi]\). It is important to note that this polar parametrization (in terms of the polar coordinates \(r, \theta\)) provides us with a confidence region more restrictive than just a product of maximal confidence intervals for individual variables \(x_1, x_2, x_3, x_4\). The maximal confidence intervals would simply be

\[
\begin{align*}
|x_1 - 534271| &\leq 6130.09, \ |x_2 - 2286155| \leq 2174.03, \\
|x_3 - 8294.32| &\leq 6192.87, \ |x_4 - 82045.4| \leq 8161.64,
\end{align*}
\] (7.1.17) (7.1.18)

but the actual elliptical region will not include the set of minimal values \(x_1 = 528141, x_2 = 283981, x_3 = 2101.45, x_4 = 73883.8\), for instance.

**Saddle-point, hyperbolic region: non-zero eigenvalues of different signs**

If, say, \(\lambda_i > 0\) and \(\lambda_j < 0\), then \((0, 0)\) is a saddle point, and the inequality becomes

\[-M \leq |\lambda_i|z_i^2 - |\lambda_j|z_j^2 \leq M,
\]

which defines the set of orthogonal hyperbolas (see Figure 14)

\[
\frac{z_i^2}{M/|\lambda_i|} - \frac{z_j^2}{M/|\lambda_j|} \leq 1, \quad \frac{z_j^2}{M/|\lambda_j|} - \frac{z_i^2}{M/|\lambda_i|} \leq 1.
\] (7.1.19)

The intersection of these conditions defines a region that looks like an elongated rectangle (elongated “corners”, the domain defined by the blue and green curves in Figure 14) and can be approximated with a rectangular shape. The confidence region is given parametrically by:

\[
z_i = \sqrt{\frac{M}{|\lambda_i|}} r \cosh(t), \quad z_j = \sqrt{\frac{M}{|\lambda_j|}} r \sinh(t), \quad -1 \leq r \leq 1, \ t \in \mathbb{R}.
\] (7.1.20)

This would give confidence regions for any choice \(\lambda_i \in \{\lambda_1, \lambda_3\}\) and \(\lambda_j \in \{\lambda_2, \lambda_4\}\).

**Example 10** As before, we compute specific confidence regions corresponding to the IPCC recommended values for yearly CO2 fluctuations.

Repeating the calculation performed in the previous example, for the case of hyperbolic confidence regions, we obtain (again, for \(M \sim 10^{-8}\)) the conditions

\[ x_1 = 534271 - 6130.08 \cdot r \sinh(t), \] (7.1.21)

74
\[ x_2 = 286155 - 2174.03 \cdot r \cosh(t), \quad (7.1.22) \]
\[ x_3 = 8294.32 - 6067.93 \cdot r \sinh(t), \quad (7.1.23) \]
\[ x_4 = 82045.4 + 8161.64 \cdot r \cosh(t), \quad (7.1.24) \]

with \(-1 \leq r \leq 1, t \in \mathbb{R}\).

Notice that this does not provide an actual confidence region (the domain defined is unbounded).

However, we can extract from the conditions above specific linear relationships between the variables that can be used for comparison purposes. Such linear relationships can be used to find equivalencies between variables \(x_1, x_3\) and \(x_2, x_4\). We perform this numerical analysis in Section 7.2.2, and indicate how to interpret the results.

7.2 Conclusions and predictions based on nonlinear analysis

Throughout this subsection, we let the values of the attributable variables \(X' = (x_1, x_2, x_3, x_4)\) be measured from the stationary point \(X_s = -\frac{1}{2}B^{-1} \cdot \beta \) (7.1.7).

7.2.1 Nonlinear analysis of contributing factors

Starting from (7.1.6)
\[ Y - Y_0 = (-140.176 z_1^2 + 134.413 z_2^2 - 131.701 z_3^2 + 9.94612 z_4^2) \times 10^{-18}, \]
and the power-law transformation
\[ Y = (\text{CO}_2)^{-2.376}, \]
we first make the important remark that increasing/decreasing CO\(_2\) is equivalent to decreasing/increasing \(Y\).

Next, using the defining relations for the linear combinations \(z_k = V_k' \cdot \hat{X}\), with \(V_k\) given in §7.1.1, we notice that the combinations \(z_1, z_3\) contribute to increase the CO\(_2\) emissions via interactions, while \(z_2, z_4\) actually decrease it. Given that (measured from the stationary point \(X_s\)),
\[ z_1 = -0.257397 Ga + 0.966306 Bu, \quad z_2 = 0.7107 Li - 0.703495 Fl, \]
\[ z_3 = -0.703495 Li - 0.7107 Fl, \quad z_4 = -0.966306 Ga - 0.257397 Bu, \]
we notice that $z_1$, which is mostly a combination of Gas Fuels and Bunker, has the most damaging effect. Along with the fact that $x_1$ (Gas Fuels) ranks first among significant attributables in the second-order model, we can conclude that Gas-related sources seem to be the most significant factors responsible to the atmospheric CO$_2$ for the European countries studied here.

7.2.2 Relative importance of attributable variables

Finally, we can estimate the correct combinations between attributable variables $x_1 - x_4$ which would keep the CO$_2$ level constant, based on our model. It is particularly useful to observe that the variables $z_1$, $z_4$ are linear combinations only of attributables Ga, Bu, while $z_2$, $z_3$ are derived from the attributables Li, Fl. Therefore, it is natural to impose the conditions

$$
\lambda_2 z_2^2 - |\lambda| z_3^2 = 0, \quad -|\lambda_1| z_1^2 + \lambda_4 z_4^2 = 0,
$$

from which we obtain the hyperplane equations

$$
z_1 = \pm \sqrt{\frac{\lambda_4}{\lambda_1}} z_4, \quad z_2 = \pm \sqrt{\frac{|\lambda_3|}{\lambda_2}} z_3.
$$

These equations (using §7.1.1) lead to the linear relationships between Ga-Bu and Li-Fl given below:

$$
-0.257397Ga + 0.966306Bu = \pm 0.266372913(-0.966306Ga - 0.257397Bu)
$$

$$
0.7107Li - 0.703495Fl = \pm 0.989860283(-0.703495Li - 0.7107Fl).
$$

From these equations it is possible to develop an equivalence between different attributables, and to use such identities in order to develop policy and accountability criteria. The only acceptable solutions (selected by positivity of proportionality coefficients) yield:

$$
Ga = 1.74388Bu, \quad Li = 98.1284Fl. \quad (7.2.1)
$$

In other words, under a “CO$_2$ trade” policy developed under these guidelines, one unit of Gas fuel is equivalent to 1.74388 unit of Bunker, while one unit of Liquid Fuel can be replaced by 98.1284 units of Gas Flares. It is important to note that this “conversion formula” corresponds to the condition $M = 0$, i.e. no variation in the CO$_2$ levels. For any other value of $M$, the formulas would provide different conversion values, as we show below.


7.2.3 Comparing the US and EU models

In order to complete the comparison between the US and EU models, initiated in [46] and in the previous chapter, we evaluate conversion rates between attributable variables, corresponding to the same range of CO₂ level fluctuations as mandated by the IPCC ($M \sim 10^{-8}$ as shown in [46, §4.1]). As mentioned above, for a given value of $M \neq 0$, the conversion rates found earlier (for $M = 0$) are not valid anymore. Instead, we start from the relations (7.1.21)-(7.1.24) (derived specifically for $M \sim 10^{-8}$), and arrive at the linear relations established from these models.

\[
\frac{x_1 - 534271}{6130.08} = \frac{x_3 - 8294.32}{6067.93} = r \sinh(t), \quad (7.2.2)
\]

\[
-\frac{x_2 - 286155}{2174.03} = \frac{x_4 - 82045.4}{8161.64} = r \cosh(t). \quad (7.2.3)
\]

Therefore, we conclude that under these conditions, one unit variation of $x_1$ (Liquid) corresponds to $6130.08/6067.93 \simeq 1.01$ units variations of $x_3$ (Gas Flares). Recall that in the US study [46] we concluded that 1000 units of Gas Flares can be equated to 2127 units of Cement; the current study shows that in the case of EU, one unit of Liquid is equivalent to approximately 1.01 units of Gas Flares. However, a direct comparison of the various trading values cannot be derived, which is yet another indication that such studies must be performed regionally, and that application of uniform policies is not supported by the data.
Chapter 8
Conclusions and perspectives

In the current work, we have presented a novel method for estimating (analytically and numerically) the probability of large deviations in the context of survival analysis and climate modeling. In particular, we have discussed the efficiency and robustness of the new methodology when applied to various types of data. The results are promising and point towards more general areas where this approach may be applicable. We give a short list of specific areas of interest and directions of research which may be pursued in this context:

1) Approximations using skewed-tail kernel functions;

2) Estimating probability of large deviations (rare events) in environmental problems (floods, rainfalls);

3) Applications to health data: estimating survival probabilities for various types of cancers;

4) Adapting the large-deviations methodology for reliability analysis;

5) Employing large-deviations methodology for estimation of 3 (or 4)-parameter families of distributions (Weibull, Johnson);

6) Deriving the optimal transport rate for stochastic transport models in molecular biology;

7) Designing a generalized algorithm for samples of non-i.i.d. from unknown distributions.

The problems listed above require an approach that should cover both the analytical aspect and the computational component, since it is especially important to provide efficient ways to estimate the quantities of interest. Some of these studies are ongoing, as indicated by the bibliography attached.
References


81


