Universitat Autònoma de Barcelona

Bachelor Thesis

Bachelor's Degree in Mathematics

# Variations of Principal Component Analysis for Extremes and Application to Climate Data

**Miguel Donderis de Vicente**

Supervisors

**Niclas Rieger, Álvaro Corral**

Year

**2021/2022**

Call

**June**

# Acknowledgements

Primero de todo, me gustaría darle las gracias a Álvaro Corral por la posibilidad que me ha dado de llevar a cabo mi Trabajo Final de Grado sobre un tema en el que siempre he tenido mucho interés, lo que me ha permitido entender una ligera parte de lo que supone el mundo de la investigación en sistemas complejos, así como por todos sus consejos a lo largo de este tiempo. También quiero darle las gracias especialmente a Niclas, que ha sido una ayuda esencial y constante a lo largo de todos estos meses, por su atención y por estar siempre dispuesto a ayudarme a aprender.

Finalmente, me gustaría dedicarle este trabajo a mi familia, en especial a mis hermanos Javier, Beatriz y Andrés, y sobre todo a mi madre, por todo el apoyo que me ha brindado a lo largo de estos años y porque gracias a todo su esfuerzo he podido estudiar lo que siempre he querido y soy quien soy a dia de hoy.

# Abstract

Throughout this thesis we will review one of the most well known multivariate analysis methods, known as Principal Component Analysis, or PCA, whose main objective is to reduce the dimensionality of a set of data vectors by finding a new set of variables maximizing the variance of the given data set. In addition, we will investigate a generalization of this method, called Maximum Covariance Analysis, or MCA, which maximizes the cross-covariance between two different sets of variables, and which reduces to PCA in case both data sets are equal.

The main goal of this work will be to compare the behaviour of the PCA standard decomposition to a recently developed extension, called Extreme Principal Component Analysis [6], allowing to work directly with extreme values. An additional contribution of this work will be to naturally extent the extreme framework in [6] to the analysis for co-varying extremes, giving rise to what we called Extreme Maximum Covariance Analysis, and therefore allowing us to compare also between the standard and extreme MCA framework.

With this purpose, we will implement these four methods in Python and apply them to different data. For comparing standard and extreme PCA, we will focus on precipitation over the Iberian Peninsula, and in the case of MCA we will introduce also sea surface salinity. Hence, our work will be focused on studying the spatial and temporal structure of the eigenvectors to which leads the eigendecomposition of the variance matrix in the standard framework and a matrix that summarizes tail pairwise dependence in the extreme framework.

# Contents

# CHAPTER 1

# Introduction

Extreme weather events such as heat waves, droughts or heavy rainfall have always posed risks to human society. This risk grows as global warming becomes more latent and with it extreme weather events increase. That is why it is more necessary than ever to make an extra effort to try to understand patterns that allow us to mitigate the disastrous effects of this extremes. Today, weather forecasting is pretty accurate when it comes to predicting the weather a couple days from now, but it's a completely different story when it comes to predicting the weather a couple weeks from now, or even months. In particular, trying to predict the amount of precipitation is a really difficult forecast because the amount of precipitation that falls depends, for example, on factors such as the sea surface temperature, which in turn fluctuates due to seasonal cycles, oscillations such as El Niño, or even climate change, and other more complex atmospheric variables.

For this, one of the most important tools is dimensionality reduction, since it allows us to focus on the most important spatial and temporal patterns, those that explain the most variability in the data, to carry out an exhaustive study of the distribution of the extreme phenomena we are interested in. The best well-known dimensionality reduction technique is Principal Component Analysis, which by solving an algebraic problem of calculating eigenvalues and eigenvectors for a positive and symmetric semi-definite matrix one can achieve the reduction of the dataset. On the other hand, Daniel Cooley and Yujing Jiang recently carried out a study [6] in which they proposed a method to analyze the extreme behavior of certain meteorological phenomena through a different base of vectors, which they stated was more efficient. The proposed method is analogous to Principal Component Analysis, but is based on extreme value analysis theory and instead of decomposing a covariance or correlation matrix, the mentioned basis of vectors is obtained by performing a proper decomposition of a matrix that

describes pairwise extremal dependence.

Hence, the first objective that we set ourselves is, since these two methods have been tested and work, to really understand what differences exist between them. To do this, we will use data from precipitation observations over the Iberian Peninsula and the east Mediterranean coast, and we will try to study the structure of the different eigenvectors and their time series that are obtained by decomposing the respective matrices. With this purpose, first we must introduce the most elementary mathematical concepts that underlie these two methods and we will review some of the properties that make of them so suitable for studying this type of tasks.

Another different dimensionality reduction method that allows to extract the dominant covariant patterns between two different data fields is called Maximum Covariance Analysis, which is a variant of Principal Component Analysis such that when the two input data fields are equal, Maximum Covariance Analysis reduces to Principal Component Analysis. The extension from one method to the other is very simple and is obtained from what is called Singular Value Decomposition. On the other hand, the extension of the method proposed by Daniel Cooley and Yujing Jiang is also quite simple, simply changing the input fields and similarly constructing a tail pairwise dependence matrix. Thus, the second part of this thesis will consist of implementing these two methods, for both the standard the extreme case, with the aim of detecting covariant patterns in two different data sets and compare their usefulness. Again, one of them will be precipitation over the Iberian Peninsula and the other will be surface salinity over the Atlantic Ocean and the Mediterranean Sea. To do this, we will again have to briefly introduce the extension of the mathematical theory that takes us from the previous methods to these.

Of course these are not the only useful methods among the climate community to reduce the complexity of studying weather forecasting, but they are the most widely used. That is why with this work we intend to contribute our grain of sand to improve or better understand the prediction of extreme weather events in the long term.

CHAPTER 2

# Methods

## 2.1 Principal Component Analysis

Principal Component Analysis is probably the oldest and best known of all the existing multivariate analysis techniques. It was first introduced in 1901 by Karl Pearson [10], as an analogous to the Principal Axis Theorem in mechanics, and developed by Harold Hotelling [4, 5] many decades later. Like almost all multivariate analysis methods, it was not widely used until the advent of the first electronic computers; but today it is widely available and is part of any statistical package.

The central idea of Principal Component Analysis, to which we will refer from now on as PCA to abreviate, is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while preserving as much variability as possible. This reduction is achieved by transforming this data into a new set of variables, called *Principal Components*, or PCs, which are uncorrelated with each other and which are ordered so that the first components retain most of the dispersion present in the total set of original data. The calculation of the PCs is reduced to solving an algebraic problem of calculating the eigenvalues and eigenvectors for a positive and symmetric semi-definite matrix, the variance matrix of the consider variable.

In this first section of the chapter we will give a formal definition of the Principal Components and give a brief overview of their algebraic derivation. Furthermore, we will review some of the most important algebraic and geometric properties underlying PCA, which make of it an optimal method for reducing the dimensionality of a dataset in order to examine its underlying structure. The results we present here are based and presented in more detail in Ian Joliffe's book [8].

### 2.1.1 Definition and derivation of the Principal Components

Let's suppose that $\mathbf{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ is a vector of $p$ random variables such that we want to study the structure of the covariances or correlations between them. Unless $p$ is very small number, or the structure of the variables is very simple, it will often not be very helpful to look directly at the $p$ variances and the $\frac{1}{2}p(p-1)$ covariances or simply consider the first few variables, as this will result on losing much information. Therefore, an alternative approach is necessary and the most optimal way is to look for a few variables that preserve most of the information given by these variances and covariances, which are called the Principal Components as we mentioned before. Now, we will review their definition and study their algebraic derivation.

The first step for finding the PCs is to look for a linear combination $\boldsymbol{\alpha}_1^\top \mathbf{x}$ of the elements of $\mathbf{x}$ having maximum variance, where $\boldsymbol{\alpha}_1 \in \mathbb{R}^p$ is a vector such that $\boldsymbol{\alpha}_1^\top \mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p = \sum_{j=1}^p \alpha_{1j}x_j$. Next, look for a second linear combination $\boldsymbol{\alpha}_2^\top \mathbf{x}$, uncorrelated with $\boldsymbol{\alpha}_1^\top \mathbf{x}$, that has maximum variance, and so on, so that at the $k$th step we are suposed to find a linear combination $\boldsymbol{\alpha}_k \mathbf{x}$ that has maximum variance subject to being uncorrelated with the $k-1$ previous linear combinations, $\boldsymbol{\alpha}_1^\top \mathbf{x}, \boldsymbol{\alpha}_2^\top \mathbf{x}, \ldots, \boldsymbol{\alpha}_{k-1}^\top \mathbf{x}$. The $k$th derived variable, $\boldsymbol{\alpha}_k^\top \mathbf{x}$, is called the $k$th *Principal Component*, as stated before. It is clear that up to $p$ PCs could be found, as much as the dimension of the original random vector, but one hopes that most of the variation of the random vector $\mathbf{x}$ will be accounted for by $m$ PCs, where $m \ll p$. Generally, if a set of $p$ variables has substantial correlations among them, then the first few PCs will account for most of the variation in the original variables. Conversely, the last few PCs identify directions of little variation, that is, near-constant linear relationships among the original variables.

Having defined the PCs, we now need to understand how to find them. Consider, for the moment, the case where the random vector $\mathbf{x}$ has known covariance matrix $\boldsymbol{\Sigma}$, the matrix whose $(i,j)th$ element is the covariance between the $i$th and the $j$th elements of the vector $\mathbf{x}$ when $i \neq j$, and the variance of the $i$th element when $i = j$. The more realistic case in which $\boldsymbol{\Sigma}$ is unknown follows by replacing it by a sample covariance matrix $\mathbf{S}$ and we will study this case later. Then, it turns out that the $k$th PC is given by $\mathbf{z}_k = \boldsymbol{\alpha}_k^\top \mathbf{x}$, $k = 1, \ldots, p$, where $\boldsymbol{\alpha}_k$ is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to the $k$th largest eigenvalue $\lambda_k$. Furthermore, if $\boldsymbol{\alpha}_k$ is chosen to be normalized, then we find that $\text{var}(\mathbf{z}_k) = \lambda_k$. The following algebraic derivation of the PCs is the standard one given in many multivariate analysis books, and it can be skipped by those familiarized with this method. Although is not unique, as the derivation of the PCs has also been solved as a geometric problem, it is the more straight and clear and so we decided to present this one.

Consider first the expression $\boldsymbol{\alpha}_1^\top \mathbf{x}$, where $\boldsymbol{\alpha}_1$ is chosen to maximize $\text{var}(\boldsymbol{\alpha}_1^\top \mathbf{x}) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1$. The first thing that we can note is that the maximum of this expression will not be achieved for a finite $\boldsymbol{\alpha}_1$, so we must impose a normalization constraint $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = \|\boldsymbol{\alpha}_1\| = 1$. The standard approach to maximize the given expression subject

to the normalization constraint is to use the Lagrange multipliers technique. Hence, we have that the expression to maximize is $\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 - 1)$, where $\lambda$ is the Lagrange multiplier. Now, differentiating with respect to $\boldsymbol{\alpha}_1$ gives

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1 = (\boldsymbol{\Sigma} - \lambda\mathbb{I}_p)\boldsymbol{\alpha}_1 = 0,$$

where $\mathbb{I}_p \in \mathbb{R}^{p\times p}$ is the identity matrix. The first thing that we note is that $\lambda$ is an eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ is its corresponding eigenvector. So as to decide which of the eigenvectors of $\boldsymbol{\Sigma}$ does this value correspond to, we recall that the quantinty to be maximized is the variance of the first PC, i.e.

$$\text{var}(\boldsymbol{\alpha}_1^\top \mathbf{x}) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^\top \lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = \lambda,$$

so $\lambda$ must be as large as possible. Therefore, we obtain that $\boldsymbol{\alpha}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$ and $\text{var}(\boldsymbol{\alpha}_1^\top \mathbf{x}) = \lambda_1$ is the largest eigenvalue of $\boldsymbol{\Sigma}$.

In general, the $k$th PC of the random vector $\mathbf{x}$ is given by the linear combination $\boldsymbol{\alpha}_k^\top \mathbf{x}$ such that $\text{var}(\boldsymbol{\alpha}_k^\top \mathbf{x}) = \lambda_k$, where $\lambda_k$ is the $k$th largest eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_k$ is its corresponding eigenvector. As we've already derived the expression for the case $k = 1$, we now will derive it for the case $k = 2$. The proof for the cases $k \geq 3$ is slightly more complicated but similar in the procedure, so we will not include it here.

The second PC is by definition given by the linear combination $\boldsymbol{\alpha}_2^\top \mathbf{x}$, which maximizes $\text{var}(\boldsymbol{\alpha}_2^\top \mathbf{x}) = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma}\boldsymbol{\alpha}_2$ subject to being uncorrelated with the first PC, $\boldsymbol{\alpha}_1^\top \mathbf{x}$, or equivalently subject to $\text{cov}(\boldsymbol{\alpha}_1^\top \mathbf{x}, \boldsymbol{\alpha}_2^\top \mathbf{x}) = 0$, where $\text{cov}(x, y)$ denotes the covariance between $x, y$. Now,

$$\text{cov}(\boldsymbol{\alpha}_1^\top \mathbf{x}, \boldsymbol{\alpha}_2^\top \mathbf{x}) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2^\top \lambda\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_2 = 0,$$

so any of these expression of the equality chain can be used to specify uncorrelation between the two first PCs. As again, we want to maximize $\text{var}(\boldsymbol{\alpha}_2^\top \mathbf{x})$ and we remember that we have to set a uncorrelation and a normalization constraint, so as to make sure that the variance is bounded. Then, the quantity to be maximized now is $\boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_1$, where $\lambda, \phi$ are the Lagrange multipliers. Differentiating with respect to $\boldsymbol{\alpha}_2$ and multiplying by $\boldsymbol{\alpha}_1^\top$ on the left we obtain

$$\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 0,$$

and since the first two terms are zero because of the uncorrelation restriction and $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$, we obtain that $\phi = 0$. Therefore, the equation is simplified to

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = (\boldsymbol{\Sigma} - \lambda\mathbb{I})\boldsymbol{\alpha}_2 = 0,$$

so $\lambda$ is once more an eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_2$ is its corresponding eigenvector. Again,

the quantity to maximize is $\text{var}(\boldsymbol{\alpha}_2^\top \mathbf{x}) = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 = \lambda$. Assuming that $\boldsymbol{\Sigma}$ does not have repeated eigenvalues that simplifies the derivation, we have that $\lambda$ is the second largest eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_2$ is its corresponding eigenvector. As stated before, it can be shown that for the third, fourth and so on PCs, the vectors $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, \ldots, \boldsymbol{\alpha}_p$ are the eigenvectors of $\boldsymbol{\Sigma}$ corresponding to $\lambda_3, \lambda_4, \ldots, \lambda_p$, the third, fourth until the smallest eigenvalue of $\boldsymbol{\Sigma}$, and furthermore $\text{var}(\boldsymbol{\alpha}_k^\top \mathbf{x}) = \lambda_k,\ k = 1, \ldots, p$, the $k$th largest eigenvalue.

It should be noted that sometimes the eigenvectors $\boldsymbol{\alpha}_k$ are referred to as the Principal Components. This usage, though sometimes defended, is confusing and it is preferable to reserve the term *Principal Components* for the derived variables $\boldsymbol{\alpha}_k^\top \mathbf{x}$, the projection of the data vector $\mathbf{x}$ onto the $k$th eigenvector $\boldsymbol{\alpha}_k$, which are also called Empirical Orthogonal Variables, and denote by *Empirical Orthogonal Functions*, or EOFs, the eigenvectors $\boldsymbol{\alpha}_k$, which are also called vector of coefficients or loadings. Hence, though we will make more emphasis on this once we start with the results analysis, to set notation we will refer to the eigenvectors as spatial patterns or EOFs, to the projection of the original data onto the eigenvectors, which is equivalent to the temporal evolution associated to spatial patterns, as PCs or PC time series and a mode corresponds to a pair of a EOF and its associated PC time series.

### 2.1.2 Optimal properties of Population Principal Components

Consider the derivation of the Principal Components we have introduced in the previous section, and denote $\mathbf{z}$ the vector whose $k$th element is $z_k$, the $k$th Principal Component, $k = 1, \ldots, p$. For ease of reading, henceforth we will denote always a Principal Component by PC. Unless stated otherwise, the $k$th PC will be taken to mean the component with the $k$th largest variance, with the corresponding interpretations for the $k$th eigenvalue and the $k$th eigenvector we introduced in the previous section. Then, we can write

$$\mathbf{z} = \mathbf{A}^\top \mathbf{x}, \tag{2.1}$$

where $\mathbf{A}^\top$ is the orthogonal matrix whose $k$th column, $\alpha_k$, the $kth$ eigenvector of $\boldsymbol{\Sigma}$. Hence, we can see that the PCs are defined by an orthonormal linear transformation of $\mathbf{x}$ given by $\mathbf{A}$, the matrix of eigenvectors of $\boldsymbol{\Sigma}$. If we remember the algebraic derivation presented in the previous section, we have seen that the equality

$$\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda} \tag{2.2}$$

is satisfied, where $\boldsymbol{\Lambda}$ is the diagonal matrix whose $k$th diagonal element is $\lambda_k$, the $k$th largest eigenvalue of $\boldsymbol{\Sigma}$, such that $\lambda_k = \text{var}(\boldsymbol{\alpha}_k^\top \mathbf{x}) = \text{var}(z_k)$. We observe two different ways of reexpressing this equality, which follow because $\mathbf{A}$ is a orthogonal matrix and that will be useful later, namely $\mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A} = \boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^\top$.

One of the main goal of this thesis is to study is to understand the importance and utility of using PCA as a dimensionality reduction method for simplifying climate

variables and helping detect different spatial patterns. So far we've defined and studied the algebraic derivation of the PCs. Now, it turns out the orthonormal linear transformation of $\mathbf{x}$ given by the expression (2.1), which defines the vector of PCs that we had denoted by $\mathbf{z}$, which make of it a really powerful method, so now we will focus on briefly reviewing and discussing some of them.

First, let $\mathbf{x}$ be a random vector such that $\mathbf{x}$ has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{y} = \mathbf{B}^\top \mathbf{x} \in \mathbb{R}^p$ be the orthonormal linear transformation given by the Principal Components (2.1). Then, we know that $\mathbb{E}(y_j) = \mu_j$ and that $\text{var}(y_j) = \lambda_j$, $j = 1, \dots, p$. On the other hand, $\text{cov}(y_i, y_j) = 0$, $i \neq j$ and $\text{var}(y_1) \geq \text{var}(y_2) \geq \cdots \geq \text{var}(y_p) \geq 0$. Finally, the most important properties are that $\sum_{j=1}^p \text{var}(y_j) = \text{tr}(\boldsymbol{\Sigma})$, $\prod_{j=1}^p \text{var}(y_j) = |\boldsymbol{\Sigma}|$, $j = 1, \dots, p$. We will not prove this results as they are straightforward, although one can follow [3], but we will have to remember this properties, as they are closely related with the optimal propositions we will study. Let's begin, then, with the study of these properties.

**Proposition 2.1.1.** *For any integer $q$, $1 \leq q \leq p$, consider the orthonormal linear transformation*

$$\mathbf{y} = \mathbf{B}^\top \mathbf{x},$$

*where $\mathbf{y} \in \mathbb{R}^q, \mathbf{B}^\top \in \mathbb{R}^{q \times p}$, and let $\boldsymbol{\Sigma}_y = \mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B}$ be the variance-covariance matrix for $\mathbf{y}$. Then the trace of $\boldsymbol{\Sigma}_y$, denoted $\text{tr}(\boldsymbol{\Sigma}_y)$, is maximized by taking $\mathbf{B} = \mathbf{A}_q$, where $\mathbf{A}_q$ consists of the first $q$ columns of $\mathbf{A}$, the matrix of eigenvectors of $\boldsymbol{\Sigma}$.*

*Proof.* Let $\boldsymbol{\beta}_k$ be the $k$th column vector of the matrix $\mathbf{B}$. As we have seen from the derivation of the PCs, the columns of $\mathbf{A}$ form a basis for a $p-$dimensional space, so we can write any vector as $\boldsymbol{\beta}_k = \sum_{j=1}^p c_{jk} \boldsymbol{\alpha}_j$, $k = 1, \dots, q$, where $c_{jk}, j = 1, \dots, p, k = 1, \dots, q$ are appropiately defined constants. Hence, matricially we can express this as $\mathbf{B} = \mathbf{AC}$, where $\mathbf{C} \in \mathbb{R}^{p \times q}$ with $(j,k)$th element $c_{jk}$. We also observe that

$$\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B} = \mathbf{C}^\top \mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A} \mathbf{C} = \mathbf{C}^\top \boldsymbol{\Sigma} \mathbf{C} = \sum_{j=1}^p \lambda_j \mathbf{c}_j \mathbf{c}_j^\top,$$

where we denote $\mathbf{c}_j^\top$ the $j$th row of $\mathbf{C}$. For obtaining this last equality we have used equation (2.2) and the fact that $\boldsymbol{\Lambda}$ is a diagonal matrix. Therefore,

$$\text{tr}(\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B}) = \sum_{j=1}^p \lambda_j \text{tr}(\mathbf{c}_j \mathbf{c}_j^\top) = \sum_{j=1}^p \lambda_j \text{tr}(\mathbf{c}_j^\top \mathbf{c}_j) = \sum_{j=1}^p \lambda_j \mathbf{c}_j^\top \mathbf{c}_j = \sum_{j=1}^p \sum_{k=1}^q \lambda_j c_{jk}^2.$$

Now, $\mathbf{C} = \mathbf{A}^\top \mathbf{B}$, so $\mathbf{C}^\top \mathbf{C} = \mathbf{B}^\top \mathbf{A} \mathbf{A}^\top \mathbf{B} = \mathbf{B}^\top \mathbf{B} = \mathbb{I}_q$, because $\mathbf{A}$ is orthogonal and the columns of $\mathbf{B}$ are orthonormal. Hence, $\sum_{j=1}^p \sum_{k=1}^q c_{jk}^2 = q$,. The columns of the matrix $\mathbf{C}$ are also orthonormal and they can be thought as the first $q$ columns of a orthogonal matrix $\mathbf{D} \in \mathbb{R}^{p \times p}$. But, as the rows of $\mathbf{D}$ are orthonormal, i.e. $\mathbf{d}_j^\top \mathbf{d}_j = 1$, $j = 1, \dots, p$ and the rows of $\mathbf{C}$ consist of the first $q$ elements of the rows of $\mathbf{D}$, we can see that

$\mathbf{c}_j'\mathbf{c}_j \leq 1,\ j = 1, \ldots, p$, and therefore $\sum_{k=1}^{q} c_{jk}^2 \leq 1$. Now, $\sum_{k=1}^{q} c_{jk}^2 \leq 1$ is the coefficient of $\lambda_j$, and the sum of these coefficients is $q$ and none of the coefficients can exceed 1. By construction of the PCs, $\lambda_1 > \lambda_2 > \cdots \lambda_p$, and we note that $\sum_{j=1}^{p} \left( \sum_{k=1}^{q} c_{jk}^2 \right) \lambda_j$ will be maximized if we can find a set of coefficients $c_{jk}$ such that

$$\sum_{k=1}^{q} c_{jk}^2 = \begin{cases} 1, & j = 1, \ldots, q, \\ 0, & j = q+1, \ldots, p. \end{cases}$$

But if $\mathbf{B}^\top = \mathbf{A}_q^\top$, then

$$c_{jk} = \begin{cases} 1, & 1 \leq j = k \leq q, \\ 0, & \text{elsewhere}, \end{cases}$$

what satisfies the previous condition. Therefore, $\operatorname{tr}(\mathbf{\Sigma}_y)$ is maximized by taking $\mathbf{B}^\top = \mathbf{A}_q^\top$, as required. □

Now, in a similar way we have the following property that once understood the previous result, this one seems natural.

**Proposition 2.1.2.** *For any integer $q$, $1 \leq q \leq p$, consider the orthonormal linear transformation*

$$\mathbf{y} = \mathbf{B}^\top \mathbf{x},$$

*where $\mathbf{y} \in \mathbb{R}^q, \mathbf{B}^\top \in \mathbb{R}^{q \times p}$, and let $\mathbf{\Sigma}_y = \mathbf{B}^\top \mathbf{\Sigma} \mathbf{B}$ be the variance-covariance matrix for $\mathbf{y}$. Then $\operatorname{tr}(\mathbf{\Sigma}_y)$ is minimized by taking $\mathbf{B} = \mathbf{A}_q^*$, where $\mathbf{A}_q^*$ consists of the last $q$ columns of $\mathbf{A}$.*

*Proof.* The derivation of the PCs given in the section 2.1.1 can be easily turned around so as to succesively find linear functions of $\mathbf{x}$ whose variance is small as possible, instead of maximizing the variance as we have done, subject to the constraint of being uncorrelated with the previous linear functions. One may note that the process is similar and that the only change is that, instead of maximizing the variance, it must be minimized. Therefore, the solutions are again the eigenvectors of $\mathbf{\Sigma}$, but this time in reverse order, starting with the one corresponding to the smallest variance. Thus, the proof for the Proposition 2.1.1 can be easily adapted to prove this case. □

The importance of Proposition 2.1.1 is very remarkable, since it states that the trace of the covariance matrix of the orthonormal transformation $\mathbf{y} = \mathbf{B}^\top \mathbf{x}$, which is the sum of the variances of the different elements of $\mathbf{y}$, is maximized when the column vectors of the matrix $\mathbf{B}$ are the $q$ first eigenvectors with greatest eigenvalue of $\mathbf{\Sigma}$. Therefore, what in fact we obtain with this result is that the $q$ first PCs maximize the variance of the projections of the original data $\mathbf{x}$ onto an arbitrary $q$-dimensional space, which was the main goal of this technique. We could think that Proposition 2.1.2 doesn't give extra information, as if we know that the $q$ first PCs maximize the variance, it is clear that

the last $q$ PCs will minimize it. However, the statistical implication of this property is that, as these last PCs have variances as small as possible, they are useful in their own right, for example in detecting unsuspected near-constant linear relationships between the elements of $\mathbf{x}$, in selecting a subset of $\mathbf{x}$ or in the detection of outliers. However, we will not focus on these applications, but in the study of the components which explain the greatest variance.

The following Proposition is not an optimal result in the sense of the two previous we've seen, but it is a useful result for dealing with the covariance matrix.

**Proposition 2.1.3** (Spectral Decomposition of $\mathbf{\Sigma}$)**.** *Let* $\mathbf{x}$ *be a vector with known covariance matrix* $\mathbf{\Sigma}$*. Then*

$$\mathbf{\Sigma} = \lambda_1 \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^\top + \lambda_2 \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_2^\top + \cdots + \lambda_p \boldsymbol{\alpha}_p \boldsymbol{\alpha}_p^\top.$$

*Proof.* From equation (2.2) we know that we can write $\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top$, and therefore expanding the right-hand side of the matrix product, as $\mathbf{\Lambda}$ is a diagonal matrix and $\mathbf{A}$ is orthonormal, we get that $\mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top = \sum_{k=1} \lambda_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^\top$, what shows that $\mathbf{\Sigma} = \sum_{k=1} \lambda_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k'$, as required. □

Looking at the diagonal elements we see that $\text{var}(x_j) = \sum_{k=1}^{p} \lambda_k \alpha_{kj}^2$. The implication of the present result is that not only we can decompose the combined variances of all the elements of $\mathbf{x}$ into decreasing contributions due to each PC, but we can also decompose the whole covariance matrix into contributions of the form $\lambda_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^\top$ from each PC.

**Proposition 2.1.4.** *For any integer* $q$*,* $1 \leq q \leq p$*, consider the orthonormal linear transformation*

$$\mathbf{y} = \mathbf{B}^\top \mathbf{x},$$

*where* $\mathbf{y} \in \mathbb{R}^q, \mathbf{B}^\top \in \mathbb{R}^{q \times p}$*, and let* $\mathbf{\Sigma}_y = \mathbf{B}^\top \mathbf{\Sigma} \mathbf{B}$ *be the variance-covariance matrix for* $\mathbf{y}$*. If* $\det(\mathbf{\Sigma}_y)$ *denotes the determinant of the covariance matrix of* $\mathbf{y}$*, then* $\det(\mathbf{\Sigma}_y)$ *is maximized when* $\mathbf{B} = \mathbf{A}_q$*.*

*Proof.* Consider an arbitrary integer $k$ such that $1 \leq k \leq q$, and let $S_k$ be the subspace of $p$-dimensional vectors orthogonal to $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{k-1}$. Then, the dimension of the vector space $S_k$ is $\dim(S_k) = p - k + 1$. Furthermore, we know that the $k$th eigenvalue of $\mathbf{\Sigma}$, $\lambda_k$, satisfies

$$\lambda_k = \sup_{\boldsymbol{\alpha} \in S_k, \boldsymbol{\alpha} \neq \mathbf{0}} \left\{ \frac{\boldsymbol{\alpha}^\top \mathbf{\Sigma} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \boldsymbol{\alpha}} \right\}.$$

Suppose that $\mu_1 > \mu_2 > \cdots \geq \mu_q$ are the eigenvalues of $\mathbf{B}^\top \mathbf{\Sigma} \mathbf{B}$ and that $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_q$ are the corresponding eigenvectors. Let's denote by $T_k$ the subspace of $q$-dimensional vectors orthogonal to $\boldsymbol{\gamma}_{k+1}, \ldots, \boldsymbol{\gamma}_q$, so we observe that $\dim(T_k) = k$. Then, for any non-zero vector $\boldsymbol{\gamma}$ in $T_k$ it holds that $\boldsymbol{\gamma}^\top \mathbf{B}^\top \mathbf{\Sigma} \mathbf{B} \boldsymbol{\gamma} \geq \mu_k \|\boldsymbol{\gamma}\|$. Consider now the subspace $\tilde{S}_k$ of $p$-dimensional vectors of the form $\mathbf{B}\boldsymbol{\gamma}$, for any vector $\boldsymbol{\gamma} \in T_k$. Then we have that

$\dim(\tilde{S}_k) = \dim(T_k) = k$, because $\mathbf{B}$ preserves lengths of vectors. From the Grassmann formula for vector spaces, we have that

$$\dim(S_k \cap \tilde{S}_k) + \dim(S_k + \tilde{S}_k) = \dim(S_k) + \dim(\tilde{S}_k).$$

But as $\dim(S_k + \tilde{S}_k) \leq p$, $\dim(S_k) = p - k + 1$ and $\dim(\tilde{S}_k) = k$, necessarily we find that $\dim(S_k \cap \tilde{S}_k) \geq 1$. Therefore, what this tells us is that there exists at least a non-zero vector $\boldsymbol{\alpha} \in S_k$ of the form $\boldsymbol{\alpha} = B\boldsymbol{\gamma}$ for a given $\boldsymbol{\gamma} \in T_k$, so it follows that

$$\mu_k \leq \frac{\boldsymbol{\gamma}^\top \mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B} \boldsymbol{\gamma}}{\boldsymbol{\gamma}^\top \boldsymbol{\gamma}} = \frac{\boldsymbol{\gamma}^\top \mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B} \boldsymbol{\gamma}}{\boldsymbol{\gamma}^\top \mathbf{B}^\top \mathbf{B} \boldsymbol{\gamma}} = \frac{\boldsymbol{\alpha}^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \boldsymbol{\alpha}} \leq \lambda_k.$$

Hence, the $k$th eigenvalue of $\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B}$ is smaller than the $k$th eigenvalue of $\boldsymbol{\Sigma}$, $k = 1, \ldots, q$. This means that $\det(\boldsymbol{\Sigma}_y) = \prod_{k=1}^q \mu_k \leq \prod_{k=1}^q \lambda_k$, but if $\mathbf{B} = \mathbf{A}_q$ then $\det(\boldsymbol{\Sigma}_y) = \prod_{k=1}^q \lambda_k$ in this case, and therefore $\det(\boldsymbol{\Sigma}_y)$ is maximized when $\mathbf{B} = \mathbf{A}_q$. □

Similarly to Proposition 2.1.1, we note that the importance of Proposition 2.1.4 lies on the fact that the determinant of the covariance matrix of the orthonormal transformation $\mathbf{y} = \mathbf{B}^\top \mathbf{x}$, which is the productory of the variances of the different elements of $\mathbf{y}$, is maximized when the column vectors of the matrix $\mathbf{B}$ are the $q$ first eigenvectors with greatest eigenvalue of $\boldsymbol{\Sigma}$. Therefore, we obtain that the PCs maximize the variance of the projections in the different components, but in this case instead of obtaining that they maximize the trace, we obtain that it maximizes the determinat. All in all, we obtain another property that tells us that, if we choose to transform the data using the PCs, then we maximize the variance of the transformed data, which we remember was the goal of this technique.

The determinant of the covariance matrix, which is also called *generalized variance* can be used as a single measure of spread for a multivariate random vector, so from Proposition 2.1.4 we now know that this spread or variation is maximized if we consider the transformation given by the first $q$ PCs. Analogously to Proposition 2.1.2 we have the following equivalent result to Proposition 2.1.4.

**Proposition 2.1.5.** *For any integer $q$, $1 \leq q \leq p$, consider the orthonormal linear transformation*

$$\mathbf{y} = \mathbf{B}^\top \mathbf{x},$$

*where $\mathbf{y} \in \mathbb{R}^q, \mathbf{B}^\top \in \mathbb{R}^{q \times p}$, and let $\boldsymbol{\Sigma}_y = \mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B}$ be the variance-covariance matrix for $\mathbf{y}$. If $\det(\boldsymbol{\Sigma}_y)$ denotes the determinant of the covariance matrix of $\mathbf{y}$, then $\det(\boldsymbol{\Sigma}_y)$ is minimized when $\mathbf{B} = \mathbf{A}_q^*$, the last $q$ columns of the matrix of eigenvectors of $\boldsymbol{\Sigma}$.*

**Proposition 2.1.6.** *Suppose that we wish to predict each random variable $x_j$ in $\mathbf{x}$ by a linear function of $\mathbf{y}$, where $\mathbf{y} = \mathbf{B}^\top \mathbf{x}$. If $\sigma_j^2$ is the residual variance in predicting $x_j$ from $\mathbf{y}$, then $\sum_{j=1}^p \sigma_j^2$ is minimized if $\mathbf{B} = \mathbf{A}_q$.*

This result may not seem very interesting at first sight, but the implication of Proposition 2.1.6 is that, if we wished to get the best linear predictor of $\mathbf{x}$ in a $q$-dimensional subspace, in the sense of minimizing the sum over elements of $\mathbf{x}$ of the residual variances, then this optimal $q$-dimensional subspace is defined by the first $q$ PCs. In fact, we will see that Proposition 2.1.6 is the population equivalent of Proposition 2.1.8, which is a really powerful result and which gives another definition of the PCs. Hence, here we will not give a proof for this result but we will give it for Proposition 2.1.8.

It is interesting to note that the PCs are the only set of $p$ linear functions of $\mathbf{x}$ that are uncorrelated and have orthogonal vector of coefficients. A special case of Proposition 2.1.6 was pointed out by Hotelling in [4], where he noted that the first PCs derived from a correlation matrix are the linear functions of $\mathbf{x}$ that have greater mean square correlation with the elements of $\mathbf{x}$ than any other function.

**Proposition 2.1.7.** *The PCs define the principal axes of the family of p-dimensional ellipsoids* $\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} = \lambda$, *where* $\lambda \in \mathbb{N}$ *is a constant.*

*Proof.* As we have studied, the PCs are defined by the orthonormal transformation $\mathbf{z} = \mathbf{A}^\top \mathbf{x}$ and since $\mathbf{A}$ is a orthogonal matrix we can find the inverse transformation, given by $\mathbf{x} = \mathbf{A}\mathbf{z}$. Then, substituing into the equation that describes a family of $p$-dimensional ellipsoids we find that $(\mathbf{A}\mathbf{z})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{A}\mathbf{z}) = \mathbf{z}^\top \mathbf{A}^\top \boldsymbol{\Sigma}^{-1}\mathbf{A}\mathbf{z}$. It is known that the eigenvectors of $\boldsymbol{\Sigma}^{-}1$ are the same as those of $\boldsymbol{\Sigma}$, and that the eigenvalues are the reciprocal, if we assume they are all strictly positive. From equation (2.2) that $\mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A} = \boldsymbol{\Lambda}^{-1}$, and therefore $\mathbf{z}^\top \boldsymbol{\Lambda}^{-1}\mathbf{z} = \sum_{k=1}^{p} \frac{z_k^2}{\lambda_k} = \lambda$, which is the equation of an ellipsoid referred to its principal axes. □

Proposition 2.1.7 is not a optimal geometric property. In fact, it is mainly only useful to interpret the PCs if the random vector $\mathbf{x}$ has a multivariate normal distribution, as in this case the ellipsoids define contours of constant probability. Furthermore, the interpretation of PCs as defining the principal axes of ellipsoids of constant density was also mentioned by Hotelling in the paper [4], one of the earliest papers on this technique.

### 2.1.3 Optimal properties of Sample Principal Components

In this section we will study some interesting and important algebraic and geometric properties of PCs obtained from a sample covariance matrix, instead of from a population covariance matrix. This case is the more realistic, because, as we mentioned before, the covariance matrix $\boldsymbol{\Sigma}$ is unknown and we have to estimate it through the sample covariance matrix $\mathbf{S}$. Most of the properties are just the sample analogous of the population case, so they will be mentioned briefly, and we will finally introduce one important geometric result.

Before deepening into the properties by themselves, we need to set some notation. Suppose that we have $n$ different observations on the $p$-dimensional random vector $\mathbf{x}$,

which we will denote by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Let $z_{i1} = \mathbf{a}_1^\top \mathbf{x}_i$, $i = 1, \ldots, n$ and choose the vector of coefficients $\mathbf{a}_1^\top$ to maximize the sample variance $\frac{1}{n-1}\sum_{i=1}^n (z_{i1} - \bar{z}_1)^2$ subject to the normalization constraint $\|\mathbf{a}_1\| = 1$, where $\bar{z}_1$ refers to the mean. Next, let $z_{i2} = \mathbf{a}_2^\top \mathbf{x}_i$, $i = 1, \ldots, n$ and choose $\mathbf{a}_2^\top$ to maximize the sample variance of $z_{i2}$ subject to the normalization constraint $\|\mathbf{a}_2\| = 1$ and to being uncorrelated with $z_{i1}$. One may note that this process is similar to the one described in the Section 2.1.1, so continuing in an obvious manner we obtain a sample version of the definition of PCs. Therefore, $\mathbf{z}_k = \mathbf{a}_k^\top \mathbf{x} = (z_{1k}, \ldots, z_{nk})$ is defined as the $k$th sample PC, $k = 1, \ldots, p$, and $z_{ik}$ is the score of the $i$th observation on the $k$th PC. Moreover, the sample variance of the PC scores for the $k$th sample is $l_k$, the $k$th largest eigenvalue of the sample covariance $\mathbf{S}$ of $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, and $\mathbf{a}_k$,$k = 1, \ldots, p$ is the corresponding eigenvector.

Let's now consider the matrices $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{n \times p}$ to have $(i, k)$th elements equal to the value of the $k$th element $x_{ik}$ of $\mathbf{x}_i$ and $z_{ik}$ of $\mathbf{z}_i$. Then, $\mathbf{X}, \mathbf{Z}$ are related by $\mathbf{Z} = \mathbf{XA}$, where $\mathbf{A} \in \mathbb{R}^{p \times p}$ is the orthogonal matrix whose $k$th column is $\mathbf{a}_k$, where we will consider $\mathbf{X}$ is a matrix of centered values and $\mathbf{Z}$ is the matrix of PC scores. In fact, critical to PCA is the fact of considering standardized data, so as to avoid reescaling problems as we want to maximize variance.

Turning to the optimal algebraic properties listed in the Section 2.1.2, if we consider the orthonormal linear transformation $\mathbf{y}_i = \mathbf{B}^\top \mathbf{x}_i$, $i = 1, \ldots, n$, where $\mathbf{B} \in \mathbb{R}^{p \times q}$ is a orthonormal matrix, then Proposition 2.1.1, 2.1.2, 2.1.4 and 2.1.6 still hold, but replacing $\mathbf{\Sigma}_y$ with the sample covariance matrix of the observations $\mathbf{y}_i$, $i = 1, \ldots, n$, and with the matrix $\mathbf{A}$ defined as having as $k$th column the vector $\mathbf{a}_k$ and $\mathbf{A}_q, \mathbf{A}_q^*$ representing the first and last $q$ columns of $\mathbf{A}$, respectively. All proofs are similar to the ones given in section 2.1.2 by replacing the population quantities with the corresponding sample quantities. Regarding Proposition 2.1.6, to which we didn't give a proof, it will reappear later as an important result an will be proved. Regarding the spectral decomposition, Proposition 2.1.3 also holds for the sample case in the form

$$\mathbf{S} = l_1 \mathbf{a}_1 \mathbf{a}_1^\top + l_2 \mathbf{a}_2 \mathbf{a}_2^\top + \cdots l_p \mathbf{a}_p \mathbf{a}_p^\top.$$

The statistical implications of these properties are esentially the same as the ones stated, but they must now be viewed in a sample context.

However, most of the optimal properties of PCs specific to sample situation are geometric, or at least have an important geometric point of view. As with the algebraic properties, the two geometric properties that we introduced in the population case are still relevant for the sample PCs, although with slight modifications to the statistical implications. Proposition 2.1.7 is still valid for samples but by replacing $\mathbf{\Sigma}$ by $\mathbf{S}$. The ellipsoids $\mathbf{x}^\top \mathbf{S}^{-1} \mathbf{x} = \lambda$, where $\lambda$ is a given constant, do not longer have the interpretation of countours of fixed probability. In this case, these ellipsoids give contours of equal Mahalanobis distance from the sample mean $\bar{\mathbf{x}}$. That is why some authors have interpreted PCA as successively finding orthogonal directions for which the Mahalanobis distance from the set to a hypersphere enclosing all the data is minimized.

The next property is the sample equivalent to Proposition 2.1.6, and both are concerned with least squares linear regression of each variable $x_j$ on the $q$ variables contained in $\mathbf{y}$. After proving it, we will understand the great importance of this result, which is that PCA minimizes the squared distance between the original data and its projection.

**Proposition 2.1.8.** *Suppose that the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are transformed by the orthonormal linear transformation $\mathbf{y}_i = \mathbf{B}^\top \mathbf{x}_i$, $i = 1, \ldots, n$, where $\mathbf{B} \in \mathbb{R}^{p \times q}$ is a orthonormal matrix, so that $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are projections of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ onto a $q$-dimensional subspace. A measure of "goodness of fit" of this $q$-dimensional subspace to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ can be defined as the sum of squared perpendicular distances of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from the subspace. This measure is minimized when $\mathbf{B} = \mathbf{A}_q$.*

*Proof.* The vector $\mathbf{y}_i$ is an orthonormal projection of $\mathbf{x}_i$ onto a $q$-dimensional subspace defined by the matrix $\mathbf{B}$. Denote by $\mathbf{m}_i$ the position of $\mathbf{y}_i$ in terms of the original coordinates, and $\mathbf{r}_i = \mathbf{x}_i - \mathbf{m}_i$. We note that $\mathbf{m}_i$ is an orthonormal projection of $\mathbf{x}_i$ onto a $q$-dimensional subspace and that $\mathbf{r}_i$ is orthogonal to that subspace, so we have that $\mathbf{r}_i^\top \mathbf{m}_i = 0$. Furthermore, $\|\mathbf{r}_i\|$ is the squared perpendicular distance of $\mathbf{x}_i$ from the subspace so that the sum of squared perpendicular distances of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from the subspace is $\sum_{i=1}^n \|\mathbf{r}_i\|$. Now, we note that

$$\|\mathbf{x}_i\| = \|(\mathbf{m}_i + \mathbf{r}_i)\| = \|\mathbf{m}_i\| + \|\mathbf{r}_i\| + 2\mathbf{r}_i^\top \mathbf{m}_i = \|\mathbf{m}_i\| + \|\mathbf{r}_i\|.$$

Hence, it is clear that

$$\sum_{i=1}^n \|\mathbf{r}_i\| = \sum_{i=1}^n \|\mathbf{x}_i\| - \sum_{i=1}^n \|\mathbf{m}_i\|,$$

so we can see that for a set of observations, minimizing the sum of squared perpendicular distances is equivalent to maximization of $\sum_{i=1}^n \|\mathbf{r}_i\|$. It is known that distances are preserved under orthogonal transformations, so the squared distance $\|\mathbf{m}_i\|$ of $\mathbf{y}_i$ from the origin is the same in $y$ or $x$ coordinates, so the quantitity to be maximized is $\sum_{i=1}^n \|\mathbf{y}_i\|$. But

$$\begin{aligned}\sum_{i=1}^n \|\mathbf{y}_i\| &= \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{B}\mathbf{B}^\top \mathbf{x}_i = \operatorname{tr} \sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{B}\mathbf{B}^\top \mathbf{x}_i) = \sum_{i=1}^n \operatorname{tr}(\mathbf{x}_i^\top \mathbf{B}\mathbf{B}^\top \mathbf{x}_i) = \\ &= \sum_{i=1}^n \operatorname{tr}(\mathbf{B}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{B}) = \operatorname{tr}\left(\mathbf{B}^\top \sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{x}_i)\mathbf{B}\right) = \operatorname{tr}(\mathbf{B}^\top \mathbf{X}^\top \mathbf{X}\mathbf{B}) = \\ &= (n-1)\operatorname{tr}(\mathbf{B}^\top \mathbf{S}\mathbf{B}).\end{aligned}$$

Finally, from Proposition 2.1.1 we know that $\operatorname{tr}(\mathbf{B}^\top \mathbf{S}\mathbf{B})$ is maximized when $\mathbf{B} = \mathbf{A}_q$, obtaining the desired result. □

Instead of treating Proposition 2.1.8 as just another property of sample PCs, it can

also be viewed as an alternative geometric derivation of the PCs. Thus, the PCs are defined as the linear functions, or projections, of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ that successively define subspaces for which the sum of squared perpendicular distances of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from the subspace is minimized. This definition provides another way in which PCs can be interpreted as accounting for as much as possible of the total variation in the data, within a lower-dimensional space. In fact, this is essentially the approach adopted by Pearson in [10].

## 2.2 Extreme Principal Component Analysis

In this section we will study the method Daniel Cooley and Yujing Jiang described in [6] for analyzing extremal behaviour through the lens of a different basis of vectors than the one obtained through PCA. Nonetheless, the method is analogous to PCA but it is based on theory from extreme value analysis. Specifically, rather than decomposing a covariance or correlation matrix, the authors obtained the basis of vectors by performing an eigendecomposition of a matrix that describes pairwise extremal dependence, what they called the Tail Pairwise Dependence Matrix.

### 2.2.1 Regular variation. Definition and first properties

The theory of regularly varying functions is an essential analytical tool for dealing with heavy tails or long-range dependence. Roughly speaking, *regularly varying functions* are those functions which behave asymptotically like power functions, i.e. functions that are heavy tailed. In this section we will consider only real functions of a real single variable. Consideration of multivariate cases and probability concepts suggests introducing definitions in terms of vague convergence of measures, which is out of the scope of this thesis.

Let's begin with some basic definitions that will lead us towards the understanding of the regularly varying framework.

**Definition 2.2.1.** *A function $f : X \to \mathbb{R}$ is measurable if, for every real number $a$, the set*

$$\{x \in X : f(x) > a\}$$

*is measurable.*

When $X = \mathbb{R}$ with Lebesgue measure, or more generally any Borel measure, then all continuous functions are measurable.

**Definition 2.2.2.** *A measurable function $U : [0, \infty) \to [0, \infty)$ is called slowly varying at $\infty$ if for all $a > 0$,*

$$\lim_{t \to \infty} \frac{U(ta)}{U(t)} = 1.$$

**Definition 2.2.3.** *A measurable function $U : \mathbb{R}^+ \to \mathbb{R}^+$ is regularly varying at $\infty$ with index $\rho \in \mathbb{R}$, denoted by $U \in RV_\rho$, if for $x > 0$,*

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\rho.$$

*The exponent $\rho$ is called the exponent of variation.*

If $\rho = 0$, then $U$ is in fact a slowly varying function. Slowly varying functions are generically denoted by $L(x)$. If $U \in \mathrm{RV}_\rho$, then $\frac{U(x)}{x^\rho} \in \mathrm{RV}_0$, and if we set $L(x) = \frac{U(x)}{x^\rho}$, we note that it is always possible to represent a regularly varying function with index $\rho$ as $x^\rho L(x)$, where $L(x)$ is some slowly varying function. In fact, this is an important result called *Karamata's characterization theorem.*

**Theorem 2.2.1** (Karamata's characterization theorem)**.** *Every regularly varying function $U : \mathbb{R}^+ \to \mathbb{R}^+$ is of the form $U(x) = x^\beta L(x)$, where $\beta \in \mathbb{R}$ and $L$ is a slowly varying function.*

In probabiliy applicactions, sometimes people are concerned with distributions whose tails are regularly varying. Some examples of such functions are $1 - F(x) = x^{-\alpha}$, $x \geq 1, \alpha > 0$, or the extreme-value distribution $\Phi_\alpha(x) = e^{-x^{-\alpha}}$. In fact, if a random variable follows a Fréchet distribution, is a special case of the generalized extreme value distribution, it has cumulative distribution function given by $\mathbb{P}(X \leq x) = e^{-x^{-\alpha}}$, where $\alpha$ is a shape parameter. We observe that $\Phi_\alpha$ has the property $1 - \Phi_\alpha(x) \sim x^{-\alpha}$ as $x \to \infty$, in other words, the upper tail is a regularly varying function with index $\rho = -\alpha$.

### 2.2.2 A framework for multivariate extremes

The foundations of the Extreme Principal Component Analysis, to which we will also refer as xPCA, described in [6] is the framework of multivariate regular variation. Basically, a random vector that is multivariate regularly varying is one that is heavy-tailed in all its dimensions, namely, each random variable of the vector follows a distribution function whose upper tail is regularly varying, with not necessarily the same tail index. Formally, the definition of a regularly varying is the following.

**Definition 2.2.4.** *A random vector $\mathbf{X}$ that takes values in $\mathbb{R}^p_+ = [0, \infty)^p$ is said to be regularly varying if there exists a sequence $b_n \to \infty$ and a limit measure $\nu_{\mathbf{X}}$ for sets in $[0, \infty)^p \backslash \{0\}$ such that $n\mathbb{P}(b_n^{-1}\mathbf{X} \in \cdot) \xrightarrow{v} \nu_{\mathbf{X}}(\cdot)$ as $n \to \infty$, where $\xrightarrow{v}$ denotes vague convergence in $M_+([0, \infty))$, the space of nonnegative Radon measures on $[0, \infty)^p \backslash \{0\}$.*

The formal definition of a regularly varying random vector requires from concepts from measure theory and vague convergence, which is out of the scope of this work as mentioned previously. Therefore, we will only introduce the most necessary tools for understanding the framework of xPCA. It can be shown that $b_n = L(n)n^{\frac{1}{\alpha}}$, where $L(\cdot)$

is some slowly varying function and $\alpha > 0$ is called the *tail index* of the random vector $\mathbf{X}$. We then write $\mathbf{X} \in \mathrm{RV}_+^p(\alpha)$ so as to denote a regularly varying random vector $\mathbf{X}$ with tail index $\alpha$.

The measure $\nu_{\mathbf{X}}$ has the *scaling property*, namely, $\nu_{\mathbf{X}}(aC) = a^{-\alpha}\nu_{\mathbf{X}}(C)$, for any set $C \subset [0,\infty)^p \backslash \{0\}$ and any $a > 0$. The scaling property implies that $\nu_{\mathbf{X}}$ can be more easely understood for sets defined by polar, rather than with cartesian coodinates. Hence, given any norm $\|\cdot\|$, the unit sphere is defined by $\mathbb{S}_{p-1}^+ = \{x \in \mathbb{R}_+^p : \|x\| = 1\}$. Now, we define the set $C(r, B) = \{x \in \mathbb{R}_+^p : \|x\| > r, \frac{x}{\|x\|} \in B\}$, for $r > 0$ and $B \subset \mathbb{S}_{p-1}^+$ a Borel set. Then, $\nu_{\mathbf{X}}\{C(r, B)\} = r^{-\alpha} H_{\mathbf{X}}(B)$, where $H_{\mathbf{X}}$ is termed the *angular measure* on $\mathbf{S}_{p-1}^+$. Consequently,

$$\nu_{\mathbf{X}}(dr \times d\omega) = \alpha r^{-\alpha-1} dr dH_{\mathbf{X}}(\omega). \tag{2.3}$$

Therefore, we can see that the probabilistic behaviour of a multivariate regularly varying random vector $\mathbf{X}$ is most easily described after a polar transformation, as the magnitude and direction of the vector are approximately independent for large observations.

Let $\mathbf{X} \in \mathbb{R}^p$ be a regularly varying random vector taking values on $[0,\infty)^p$. We work on the $p$-dimensional positive orthant, as this allows us to focus on the large values and ignore the small ones. A formal definition of regular variation for random vectors requires ideas of vague convergence as we have explained, so more details can be found in [2, 12], where the authors give a comprehensive treatment of regular variation. For the purposes of this thesis, it suffices to assume that if $A$ is a set consisting of large values, sufficiently far away from the origin, then

$$P(\mathbf{X} \in A) \propto \int_{(r,\omega)\in A} \alpha r^{-(\alpha+1)} dr dH_{\mathbf{X}}(\omega). \tag{2.4}$$

Here the symbol $\propto$ denotes "proportional to", $\alpha > 0$ and $r$ refers to the radial component of the location, $\boldsymbol{\omega}$ is a location on the unit sphere $\mathbb{S} = \{\boldsymbol{\omega} \in \mathbb{R}_+^p : \|\boldsymbol{\omega}\| = 1\}$, and $H$ is a measure on the unit sphere $\mathbb{S}$. The heavy-tailed nature of the distribution is shown in that $r$ in the integrand has power-law behaviour given by the parameter $\alpha$. As $\alpha$ decreases, the tail becomes heavier, and we can see that $\alpha$ is the reciprocal of $\xi$, the shape parameter of the Generalized Extreme Value distribution. If we assume (2.4), the probabilistic behaviour of extreme events is characterized by the tail index $\alpha$ and the angular measure $H$, which describes the tail dependence. We note that this integral also follows from equation (2.3) by assuming $\mathbf{X}$ is a regularly varying random vector with tail index $\alpha$, and that $\mathbf{X}$ follows a Fréchet distribution, but we will come to this fact later.

Asymptotic independence is a fundamental notion of tail dependence. Let $x_1(p)$ and $x_2(p)$ denote respectively the $p$th quantile of random variables $X_1$ and $X_2$. Then, if $\lim_{p\to 1} \mathbb{P}(X_2 > x_2(p) | X_1 > x_1(p)) = 0$ we affirm that $X_1$ and $X_2$ are asymptotically

independent, and that $X_1$ and $X_2$ are asymptotically dependent if the limit is greater than zero. If $X_1$ and $X_2$ are jointly regularly varying random variables and also asymptotically independent, then the mass of $H$ exists only on the axes. On the other hand, formally we say that a random variable $X_i$ has scale $b$ if $\lim_{x\to\infty} \mathbb{P}(X_i > x)/x^{-\alpha} = b$. If $X_i$ is a regularly varying random variable with unit scale, then it's clear that $bX_i$ will have scale $b$. In standard PCA, scale is described by the variance, but we have to be careful because variance speaks about the scale of the random variable from its center mean, whereas scale in the framework we are introducing it describes the behaviour in the random variable's tail.

In small dimensions, the angular measure $H$ can be chosen to be modeled parametrically or nonparametrically. However, in high dimensions rather than completely model the high-dimensional angular measure, we will introduce a matrix of bivariate tail dependencies, called the Tail Pairwise Dependence Matrix, that summarizes the tail dependence contained in the angular measure $H$. Finally, we must say that the regular variation framework described above requires that each of the variables is heavy-tailed with a common tail index $\alpha$. Often in extremes studies, the data does not exhibit this property, so transforming the marginal distributions is common and necessary to extremes studies and can be defended by theoretical results, see for instance [11]. We will come back to this point in Chapter 3.

### 2.2.3 The Tail Pairwise Dependence Matrix

Assume $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ is a $p$-dimensional regularly varying random vector with tail index $\alpha = 2$ and angular measure $H$. We note that, as the values of $\mathbf{X}$ become larger, after the polar transformation the magnitude given by the radial component $r$ in the integral (2.4) becomes independent of the angular component $\boldsymbol{\omega} \in \mathbb{S}$ of the random vector, so the pairwise dependence information is contained essentially in the measure $H$ that lives on $\mathbb{S}$. Let's consider the angular components of the random vector $\mathbf{X}$, which we denote by $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)$ that live on the unit sphere $\mathbb{S}_{p-1}^{+} = \{\omega \in \bar{\mathbb{X}}_{+}^{p} : \|\omega\|_2 = 1\}$, where $\|\cdot\|_2$ is the Euclidean norm, and define $\boldsymbol{\Sigma}_{\mathbf{X}} \in \mathbb{R}^{p\times p}$ a matrix of summary pairwise dependencies by letting

$$\sigma_{\mathbf{X},ij} = \int_{\mathbb{S}_{p-1}^{+}} \omega_i \omega_j dH(\omega), \quad \boldsymbol{\Sigma}_{\mathbf{X}} = (\sigma_{\mathbf{X},ij}), i, j = 1, \ldots, p. \tag{2.5}$$

We refer to $\boldsymbol{\Sigma}_{\mathbf{X}}$ as the Tail Pairwise Dependence Matrix of $\mathbf{X}$, or TPDM to abreviate, and $\sigma_{\mathbf{X},ik}$ corresponds to the extremal dependence measure, defined in the bivariate case in [9]. Furthermore, Cooley and Thibaud prove in [2] that in fact this is a positive semidefinite matrix.

Although this matrix focuses on extremal dependence due to its dependence on the angular measure $H$, the construction of the TPDM is similar to that of the standard covariance matrix, and therefore it seems natural to have similar properties. Hence, if

$X_i$ has scale $b$, then the $i$th diagonal element $\sigma_{ii}$ is $b^2$, and $\sigma_{ij} = 0$ if and only if $X_i$ and $X_j$ are asymptotically independent random variables.

A fact which is important for PCA is the fact that the covariance matrix is symmetric and positive definite, and therefore its eigenvectors are real and the eigenvalues are positive. In the case of $\mathbf{\Sigma_X}$, if the marginal distributions are transformed to have a common scale of one, then the TPDM behaves like a correlation matrix with diagonal entries of one. Another important property is that the TPDM is a completely positive matrix, namely, there exists a nonnegative matrix $B \in \mathbb{R}^{p+q}$, with $q \geq p$ such that $\mathbf{\Sigma_X} = \mathbf{B}\mathbf{B}^T$.

As with the covariance matrix, in the practice one has no way to calculate (2.5) and the matrix $\mathbf{\Sigma_X}$. Therefore, so as to estimate the TPDM let $\mathbf{x}_t, t = 1, \ldots, n$ be the transformed observations, for being regularly varying with tail index $\alpha = 2$, for all stations on day $t$. Then, the elements of the TPDM are estimated using pairs of $\mathbf{x}_t$ elements. Let's first define the radial and angular components

$$r_{t,ij} = \sqrt{x_{t,i}^2 + x_{t,j}^2}, \quad (\omega_{t,i}, \omega_{t,j}) = \frac{(x_{t,i}, x_{t,j})}{r_{t,ij}}.$$

We therefore estimate $\sigma_{\mathbf{X},ij}$ as

$$\hat{\sigma}_{\mathbf{X},ij} = \frac{2}{n_{ij,\text{exc}}} \sum_{t=1}^{n} \omega_{t,i}\omega_{t,j}\mathbb{I}(r_{t,ij} > r_{0,ij}), \quad \hat{\mathbf{\Sigma}}_{\mathbf{X}} = (\hat{\sigma}_{\mathbf{X},ij}), i, j = 1, \ldots, p. \tag{2.6}$$

where $r_{0,ij}$ is some high threshold for the radial components and $n_{ij,\text{exc}}$ is the number of observations whose $r_{ij}$ is greater than the corresponding high threshold. As done in [6], we choose $r_{0,ij}$ to correspond to the 0.98 quantile of the data. We note that the indicator function $\mathbb{I}(r_{t,ij} > r_{0,ij})$ forces the estimation to be based on the pairs with the largest radial component, which must be greater than $r_{0,ij}$. However, choosing this value involves the usual difficulties often found in choosing a threshold in an extreme value analysis.

One issue with this pairwise estimate is that $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$ is not guaranteed to be positive definite, so in this case we will have to find the nearest positive definite matrix to the estimated $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$ and then perform all the corresponding analysis with it.

### 2.2.4 Inner product space via transformation

In the following section we describe the framework for defining an inner product space on a given open set. We then use this framework to define a particular inner product space on its positive orthant, whose operations preserve regular variation. Daniel Cooley and Émeric Thibaud work [2] can be followed for more details on the construction of this vector space in the positive orthant.

Let $t$ be a bijection from $\mathbb{R}$ onto some open set $\mathbb{X} \subset \mathbb{R}$, and let $t^{-1}$ be its inverse. We will refer to $t$ as the *transform*. Let $\mathbb{X}^p$ be the set of $p-$dimensional vectors whose

elements lie on $\mathbb{X}$, namely, if $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{X}^p$, then $x_i \in \mathbb{X}$, $i = 1, \ldots, p$. We denote by $t(\mathbf{y})$ the elementwise application of the transform $t$ to the elements of a given vector $\mathbf{y} \in \mathbb{R}^p$, in such a way that other functions operating on vectors will be applied elementwise similarly. We define the vector addition of two elements $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}^p$ by

$$\mathbf{x}_1 \oplus \mathbf{x}_2 = t\left(t^{-1}(\mathbf{x}_1) + t^{-1}(\mathbf{x}_2)\right).$$

Given a real number $a \in \mathbb{R}$, we define the scalar multiplication by a vector $\mathbf{x} \in \mathbb{X}^p$ by

$$a \circ \mathbf{x} = t\left(at^{-1}(\mathbf{x})\right).$$

We finally define the additive identity in $\mathbb{X}^p$ as $0_{\mathbb{X}^p} = t(0)$ and the additive inverse of any $\mathbf{x} \in \mathbb{X}^p$ by $-x = t\left(-t^{-1}(x)\right)$. In the Appendix A we show that $\mathbb{X}^p$ is in fact a vector space.

Following the construction of the vector space $\mathbb{X}^p$, let $\mathbf{x}_j \in \mathbb{X}^p$ and $a_j \in \mathbb{R}$, $j = 1, \ldots, q$. Then, we define a linear combination in $\mathbb{X}^p$ as

$$a_1 \circ \mathbf{x}_1 \oplus a_2 \circ \mathbf{x}_2 \oplus \cdots \oplus a_q \circ \mathbf{x}_q = t\left(\sum_{j=1}^{q} a_j t^{-1}(\mathbf{x}_j)\right).$$

We note that, as $\mathbb{X}^p$ is a $p-$dimensional vector space, any set of $p$ vectors that are linearly independent in $\mathbb{X}^p$, i.e. vectors such that if $a_1 \circ \mathbf{x}_1 \oplus \cdots \oplus a_q \mathbf{x}_q = 0$, necessarily $a_j = 0$, $\forall j = 1, \ldots, q$, will form a basis for $\mathbb{X}^p$.

Let now be $A = (a_1, \ldots, a_q) \in \mathbb{R}^{p \times q}$ a matrix of real numbers. Then, given a vector $\mathbf{x} \in \mathbb{X}^q$ we define the matrix multiplication by $A$ as

$$A \circ \mathbf{x} = a_1 \circ \mathbf{x}_1 \oplus \cdots \oplus a_q \circ \mathbf{x}_q = t\left(At^{-1}(\mathbf{x})\right) \in \mathbb{X}^p.$$

If we denote $\mathbb{I} \in \mathbb{R}^p$ the identity matrix, then $\mathbb{I} \circ \mathbf{x} = t\left(It^{-1}(\mathbf{x})\right) = x$, and we note that if $B \in \mathbb{R}^{p' \times p}$, then $B \circ A \circ \mathbf{x} = B \circ t\left(At^{-1}(\mathbf{x})\right) = t\left(BAt^{-1}(\mathbf{x})\right) = BA \circ \mathbf{x} \in \mathbb{X}^{p'}$. Now that we've defined matrix multiplication in our particular vector space, we recall that as in $\mathbb{R}^p$, linear combinations can be written as matrix operations. However, because of the fact that constants $a_j$ lie in $\mathbb{R}$ and vectors $\mathbf{x}_j$ lie in $\mathbb{X}^p$ $\forall j = 1, \ldots, n$, the linear combination in this case becomes

$$a_1 \circ \mathbf{x}_1 \oplus \cdots \oplus a_q \circ \mathbf{x}_q = Y \circ t(a), \tag{2.7}$$

where $Y \in \mathbb{R}^{p \times q}$ is the matrix whose columns are $\mathbf{y}_j = t^{-1}(\mathbf{x}_i)$, $j = 1, \ldots, q$.

Given two vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}^p$, we define the scalar product in $\mathbb{X}^p$ by

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = \sum_{i=1}^{p} t^{-1}(\mathbf{x}_{1i}) t^{-1}(\mathbf{x}_{2i}).$$

In the Appendix A we prove that the conditions for being a well defined scalar product are met. We define the norm of a vector $\mathbf{x} \in \mathbb{X}^p$ as $\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}}$, and say that two vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}^p$ are orthogonal if $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = 0$, denoted by $\mathbf{x}_1 \perp \mathbf{x}_2$. Now, vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}^p$ and their preimages, $\mathbf{y}_1 = t^{-1}(\mathbf{x}_1), \mathbf{y}_2 = t^{-1}(\mathbf{x}_2) \in \mathbb{R}^p$ share the same inner product value. Consequently, $\|\mathbf{x}\| = \|\mathbf{y}\|_2$, with $\|\cdot\|_2$ the Euclidean norm, and $\mathbf{x}_1 \perp \mathbf{x}_2$ if and only if $\mathbf{x}_1 \perp \mathbf{x}_2 \in \mathbb{R}^p$.

Let's consider now a nonsigular matrix $S \in \mathbb{R}^{p \times p}$ and think of the linear transformation associated with it, $S : \mathbb{R}^p \to \mathbb{R}^p$, defined by $\mathbf{x} \mapsto S \circ \mathbf{x}$. We therefore define the inverse operator, and $S^{-1}$ its associated matrix, to be an application such that $S^{-1} \circ (S \circ \mathbf{x}) = S \circ (S^{-1} \circ \mathbf{x}) = \mathbf{x}$. We note that the inverse operator coincides with the usual multiplication by the inverse matrix.

If $S$ is a diagonalizable matrix, we can define the eigenvalue/eigenvector pair of $S$ to be $\lambda \in \mathbb{R}$ and $\mathbf{e} \in \mathbb{X}^p$ such that $S \circ \mathbf{e} = \lambda \circ \mathbf{e}$, where we assume that $\|\mathbf{e}\| = 1$. Now, if $\lambda$, $\mathbf{u}$ are its corresponding eigenvalue/eigenvector pair in $\mathbb{R}^p$, then it holds that $\lambda$ and $\mathbf{e} = t(\mathbf{u})$ form an eigenvalue/eigenvector pair in $\mathbb{X}^p$, i.e. $S \circ \mathbf{x} = t(S\mathbf{u}) = t(\lambda \mathbf{u}) = \lambda \circ \mathbf{e}$. Further, we assume that $S$ is a symmetric positive definite matrix, namely $\mathbf{y}^T S \mathbf{y} > 0$ for any $\mathbf{y} \in \mathbb{R}^p \backslash \{0\}$. We therefore define a positive quadratic form im $\mathbb{X}^p$, whose associated matrix is $S$, by $Q(S, \mathbf{x}) = \langle \mathbf{x}, S \circ \mathbf{x} \rangle$. In fact, we observe that

$$Q(S, \mathbf{x}) = \langle \mathbf{x}, t\left(St^{-1}(\mathbf{x})\right) \rangle = \sum_{i=1}^{p} \sum_{j=1}^{p} t^{-1}(\mathbf{x}_i) s_{ij} t^{-1}(\mathbf{x}_j) = \mathbf{y}^T S \mathbf{y},$$

where $\mathbf{y} = t^{-1}(\mathbf{x})$. Hence, we see that $\mathbf{x}$ and its inverse image share the same quadratic form with respect to a symmetric positive-definite matrix $S$. Consequently, relationships between the eigenvectors and eigenvalues of $S$ and bounds on the quadratic forms in $\mathbb{R}^p$ carry over to $\mathbb{X}^p$, yielding the following proposition whose proof follows from linear algebra results in $\mathbb{R}^p$ in [7].

In the following section we will apply the ideas of this section to the particular case of the transform $\tau : \mathbb{R} \to [0, \infty)$ defined by $\tau(y) = \log(1 + e^y)$. This bijection, known as the *softplus function* is mainly used in neural networks as it is continuous and infinitely differentiable. The functions maps the elements of $\mathbb{R}$ to elements of its positive orthant $\mathbb{R}_+$ and its inverse is $\tau^{-1}(x) = \log(e^x - 1)$. Importantly for our purposes, it holds that $\lim_{y \to \infty} \tau(y)/y = \lim_{x \to \infty} \tau^{-1}(x)/x = 1$, that is, the transform and its inverse have a negligible effect on large values. For our purposes, we will extend $\tau$ such that $\tau(-\infty) = 0$, $\tau^{-1}(0) = -\infty$ and $\tau(\infty) = \tau^{-1}(\infty) = \infty$. Then, we have that $\tau : \overline{\mathbb{R}}^p \to \overline{\mathbb{X}}^p$, where $\overline{\mathbb{R}}^p = [-\infty, \infty]^p$ and $\overline{\mathbb{X}}^p = [0, \infty]^p$.

### 2.2.5 Principal Component Analysis decomposition for extremes

As we have explained in the previous section, essential for the standard PCA method is the fact that the eigenvectors of the covariance matrix form an orthonormal basis for the $p-$dimensional reals, which is ordered in importance by the eigenvalues, which

yield the amount of variance explained by each eigenvector, of the covariance matrix. Critical to this method will be therefore obtaining an ordered orthonormal basis for the $p-$dimensional positive orthant.

So as to find this basis, we must first have defined a vector space for the $p$-dimensional positive orthant, which as explained before is obtained by applying the *softplus function* transformation $\mathbf{x} = \tau(\mathbf{y}) = \log(1 + e^{\mathbf{y}})$ componentwise to the vector $\mathbf{y} \in \mathbb{R}^p$. One important characteristic of this transformation, as we previously mentioned, is that $\tau(\mathbf{y}) \approx \mathbf{y}$ when $\mathbf{y} \to \infty$, and therefore the transformation is negligible for large values. Furthermore, vector addition and scalar multiplication of a vector are defined via this transformation, and regular variation is preserved by this particular transformation as well. This last property is studied more detail in [2].

Additionally, Cooley and Thibaud show in [2] that applying this transformation to the eigenvectors of the TPDM yields an orthonormal basis for the positive orthant and that this basis is ordered by eigenvalues that yield the scale explained by each eigenvector. We remember that in the TPDM framework, the scales refers to the behaviour in the random variable's tail. Now, let $\mathbf{\Sigma}_X = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, where $\mathbf{D}$ is the diagonal matrix whose elements are the eigenvalues of the TPDM, with $\lambda_1 \geq \lambda_p \geq 0$, and $\mathbf{U}$ is a matrix with columns $\mathbf{u}_i, i = 1, \ldots, p$ the corresponding eigenvectors. Then, the eigenvectors for the positive orthant are given after the transformation $\mathbf{e}_i = \tau(\mathbf{u}_i)$.

Finally, let $\mathbf{x}_t$ be the realization of the regularly varying random vector $\mathbf{X}$ with associated TPDM $\mathbf{\Sigma}_\mathbf{X}$ at time $t$. Let $\mathbf{v}_t = \mathbf{U}^\top \tau^{-1}(\mathbf{x}_t)$. Then $\mathbf{v}_t \in \mathbb{R}^p$ is the vectors of PCs for $\mathbf{x}_t$, that is, it is the vector of coefficients of the eigenbasis

$$\mathbf{x}_t = \mathbf{v}_{t,1} \circ \mathbf{e}_1 \oplus \cdots \oplus \mathbf{v}_{t,p} \circ \mathbf{e}_p,$$

with the transformed multiplication and addition.

The PCA decomposition becomes useful from the knowledge that most of the information in $\mathbf{x}_t$ is contained in the leading terms of the linear combination. In a standard PCA study, the leading eigenvectors are often visualized and interpreted, as the orthogonality constraint implies that the eigenvectors contain no redundant information, so interpretation is done sequentially. Here, each eigenvector is the direction of greatest variance remaining after the variance accounted for by the previous eigenvectors is removed. Furthermore, the time series of the leading PCs $\mathbf{v}_{t,i}$ can be investigated to find behavior in the often large-scale effects described by the corresponding eigenvectors. This last fact will be critical to our comparison between standard and extreme PCA methods.

## 2.3 Maximum Covariance Analysis

Maximum Covariance Analysis, to which we will refer from now on as MCA to abreviate, is a technique, similar to others such as Canonical Correlation Analysis, that tries to

find pairs of linear combinations of two sets of data vectors such that their covariances, instead of their correlations as in the case of the Canonical Correlation Analysis, are maximized. In other words, this method looks for patterns in two space-time datasets which explain a maximum fraction of the cross-covariance between them. We will now introduce the main theoretical background behind MCA. With this purpose we will explain the Singular Value Decomposition and study the optimality of the first mode of this decomposition. The results we present are esentially based on the book wrote by Wilks [15] and other lecture material written by Bretherton [1].

### 2.3.1 Singular Value Decomposition

The singular value decomposition, to which we will also refer as SVD, is a factorization of any real or complex matrix that generalizes the eigendecomposition of a square normal matrix, i.e. a matrix that commutes with its conjugate transpose, with an orthonormal eigenbasis to any $m \times n$ matrix.

Specifically, the singular value decomposition of an complex matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ is a factorization of the form $\mathbf{A} = \mathbf{U\Lambda V}^*$, where $\mathbf{U} \in \mathbb{C}^{m\times m}$ is an complex unitary matrix, $\mathbf{\Lambda} \in \mathbb{C}^{m\times n}$ is a rectangular diagonal matrix with non-negative real numbers on the diagonal, and $\mathbf{V} \in \mathbb{C}^{n\times n}$ is a complex unitary matrix. If $\mathbf{A}$ is real, then $\mathbf{U}$ and $\mathbf{V}$ can also be guaranteed to be real orthogonal matrices. In such contexts, the SVD is often denoted $\mathbf{U\Lambda V^T}$.

For the following, we will assume $\mathbf{A}$ is a real matrix. The diagonal entries of $\mathbf{\Lambda}$ are uniquely determined by the original matrix $\mathbf{A}$ and are known as the singular values of $\mathbf{A}$. The number of non-zero singular values is equal to the rank of $\mathbf{A}$. The columns of $\mathbf{U}$ and the columns of $\mathbf{V}$ are respectively called left-singular vectors and right-singular vectors of $\mathbf{A}$ and form two sets of orthonormal bases $u_1, \ldots, u_m$ and $v_1, \ldots, v_n$, and the SVD can be written as $\mathbf{A} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^\top$, where $r \leq \min(m, n)$ is the rank of M.

The importance of SVD for PCA is twofold. First, it provides a computationally efficient method for actually finding the PCs. It is clear that if we can find $\mathbf{X}, \mathbf{\Lambda}$ such that $\mathbf{X\Lambda X^T}$, then $\mathbf{\Lambda}$ and $\mathbf{X}$ will give us the eigenvectors and the square roots of the eigenvalues of $\mathbf{XX}^\top$, and therefore the coefficients and standard deviations of the principal components for the sample covariance matrix $\mathbf{S}$.

### 2.3.2 Optimality of the first Singular Value Decomposition mode

Let's consider $\mathbf{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ and $\mathbf{y} = (y_1, \ldots, y_q)^\top \in \mathbb{R}^q$ two vectors of $p$ and $q$ random variables. Consider now two data matrices $\mathbf{X} \in \mathbb{R}^{p\times n}$ and $\mathbf{Y} \in \mathbb{R}^{q\times n}$, the matrices that contain $n$ number of observations of each random vector, where $m$ and $q$ are respectively the number of $\mathbf{x}$ and $\mathbf{y}$ measurements at each time. Let $\mathbf{u}$ be an arbitrary unit column $p$-vector representing a pattern in the $\mathbf{x}$ field and $\mathbf{v}$ be an arbitrary unit column $q$-vector representing a pattern in the $\mathbf{y}$ field. Let the time series of their projection on the data be the row vectors $\mathbf{a} = \mathbf{u}^\top\mathbf{X}, \mathbf{b} = \mathbf{v}^\top\mathbf{Y} \in \mathbb{R}^{1\times n}$. So far,

this is the typical scheme we've seen in the PCA analysis. Then, MCA looks for optimal patters in the structure of the data $\mathbf{x}, \mathbf{y}$ such that the covariances of the time series $\mathbf{a}, \mathbf{b}$ is maximized, subject to the constraint that the vectors $\mathbf{u}$ and $\mathbf{v}$ are orthonormal. Now, we note that

$$\text{cov}(\mathbf{a}, \mathbf{b}) = \text{cov}(\mathbf{u}^\top \mathbf{X}, \mathbf{v}^\top \mathbf{Y}) = \frac{1}{n-1} \left( \mathbf{u}^\top \mathbf{X} (\mathbf{v}^\top \mathbf{Y})^\top \right) = \mathbf{u}^\top \mathbf{S_{xy}} \mathbf{v},$$

where $\mathbf{S_{xy}} = \frac{1}{n-1} \mathbf{X}\mathbf{Y}^\top$ is the sample covariance matrix between $\mathbf{x}$ and $\mathbf{y}$, whose $(i,j)$th element is the cross-covariance between $x_i \in \mathbf{x}$ and $y_j \in \mathbf{y}$ for a given observation. Computationally, the vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ are found through a Singular Value Decomposition of the matrix

$$\mathbf{S_{xy}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{u}_k$ is the $k$th column vector of $\mathbf{U}$ and is called left-singular vector, while $\mathbf{v}_k$ is the $k$th column vector of $\mathbf{V}$ and is called right-singular vector, and the elements of the diagonal matrix $\mathbf{\Sigma}$ are the covariances between the pairs of the previous linear combinations $\mathbf{a}, \mathbf{b}$, and are called singular values. Hence, the maximum value of $\text{cov}(\mathbf{u}^\top \mathbf{X}, \mathbf{v}^\top \mathbf{Y})$ is obtained from the leading mode of the SVD of $\mathbf{S_{xy}}$, with $\mathbf{u}_1$ and $\mathbf{v}_1$ being respectively the first left and right singular vector, and therefore $\text{cov}(\mathbf{u}_1^\top \mathbf{X}, \mathbf{v}_1^\top \mathbf{Y}) = \sigma_1$. The following modes of this decomposition maximize the mentioned covariance subject to the constraint that the vectors $\mathbf{u}_k, \mathbf{v}_k$ are orthogonal to the previous $k-1$ modes.

Each SVD mode explains an amount of $\sigma_k^2$, $k = 1, \ldots, r = \min(m, q)$ of the overall squared covariance of $\mathbf{S_{xy}}$, where $\sigma_k^2$ is the $k$th largest eigenvalue of the cross-covariance matrix. As for PCA, it can also be important to think of the importance of each SVD mode in terms of the explained squared covariance fraction from the total amount, which is $\frac{\sigma_k^2}{\sum_{k=1}^r \sigma_k^2}$.

Consider the SVD decomposition of the covariance matrix $\mathbf{S_{xy}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Now, we can express $\mathbf{u}$ in the basis of the left singular vectors, $\hat{\mathbf{u}} = \mathbf{U}^\top \mathbf{u}$.; and analogously we can express $\mathbf{v}$ in the basis of the right singular vectors $\hat{\mathbf{v}} = \mathbf{V}^\top \mathbf{v}$. We note that $\hat{\mathbf{u}}$ and $\hat{\mathbf{u}}$ are two unit vectors. Remember that the magnitude we wanted to maximize was $\text{cov}(\mathbf{u}^\top \mathbf{X}, \mathbf{v}^\top \mathbf{Y}) = \mathbf{u}^\top \mathbf{S_{xy}} \mathbf{v}$. Therefore, combining all three previous expressions we have that

$$\text{cov}(\mathbf{u}^\top \mathbf{X}, \mathbf{v}^\top \mathbf{Y}) = \mathbf{u}^\top \mathbf{S_{xy}} \mathbf{v} = \mathbf{u}^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{v} = \hat{\mathbf{u}}\mathbf{\Sigma}\hat{\mathbf{v}} = \sum_{k=1}^{r} \hat{u}_k \sigma_k \hat{v}_k,$$

where $r = \min(m, q)$ is the rank of the covariance matrix. Now, applying the triangular inequality and the fact that $\hat{\mathbf{v}}$ is a unitary vector, we have that

$$\sum_{k=1}^{r} \hat{u}_k \sigma_k \hat{v}_k \leq \left( \sum_{k=1}^{r} \hat{u}_k^2 \sigma_k^2 \right)^{\frac{1}{2}} \left( \sum_{k=1}^{r} \hat{v}_k \right)^{\frac{1}{2}} \leq \left( \sum_{k=1}^{r} \hat{u}_k^2 \sigma_k^2 \right)^{\frac{1}{2}} \leq \sigma_1 \left( \sum_{k=1}^{r} \hat{u}_k^2 \right)^{\frac{1}{2}} \leq \sigma_1.$$

Therefore, we obtain that $\sigma_1$, the greatest value of $\text{cov}(\mathbf{u}^\top \mathbf{X}, \mathbf{v}^\top \mathbf{Y}) = \sum_{k=1}^{r} \hat{u}_k \sigma_k \hat{v}_k$, is achieved when $\hat{u}_1 = 1$ and $\hat{u}_k = 0$, $k \geq 1$. Similarly, we can see that we must take $\hat{\mathbf{v}}_1 = 1$ and $\hat{\mathbf{v}} = 0$, $k \geq 1$, i.e. $\mathbf{u} = \mathbf{u}_1$ and $\mathbf{v} = \mathbf{v}_1$. Similar to the PCA derivation, one can prove that mode $n$ explains a maximum fraction of the spatial covariance between $\mathbf{u}$ and $\mathbf{v}$ subject to the rescriction of being uncorrelated to the first $n-1$ SVD modes. In this case, we have that the eigenvalue of this mode is $\sigma_k$, the $k$th largest eigenvalue of $\mathbf{\Sigma_{xy}}$.

We can finally define the time series of the original data associated with the $k$th SVD mode, which are vectors with covariance $\sigma_k$, the $k$th largest cross-covariance. These are $\mathbf{a}_k = \mathbf{u}_k^\top \mathbf{X}$ and $\mathbf{b}_k = \mathbf{v}_k^\top \mathbf{Y}$, which are not necessarily uncorrelated.

## 2.4 Extreme Maximum Covariance Analysis

In this last theory section we will briefly introduce the theoretical framework of Extreme Maximum Covariance Analysis, or xMCA. Here we will only deal with the derivation of the TPDM, which will be analogous to the one given in the case of xPCA. Remember that the extreme method consisted on, rather than decomposing a covariance or correlation matrix, obtaining the basis of vectors by performing an eigendecomposition of a matrix that describes pairwise extremal dependence. This decomposition will follow from the fact that the TPDM is a completely positive matrix, what allows us to perform a SVD decomposition.

### 2.4.1 The Tail Pairwise Dependence Matrix

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_q)^T$ be a $p$-dimensional and $q$-dimensional regularly varying random vector with tail index $\alpha = 2$ and angular measure $H$. As previously, the pairwise dependence information is contained in the measure $H$ that lives on $\mathbb{S}$. Let's consider the angular components of the random vector $\mathbf{X}, \mathbf{Y}$, which we denote by $\boldsymbol{\omega}_\mathbf{X} = (\omega_1^\mathbf{X}, \ldots, \omega_p^\mathbf{X})$, $\boldsymbol{\omega}_\mathbf{Y} = (\omega_1^\mathbf{Y}, \ldots, \omega_q^\mathbf{Y})$ that live on the unit sphere $\mathbb{S}_{p-1}^+ = \{\omega \in \bar{\mathbb{X}}_+^p : \|\omega\|_2 = 1\}$, where $\|\cdot\|_2$ is the Euclidean norm, and define $\mathbf{\Sigma_{XY}} \in \mathbb{R}^{p \times q}$ a matrix of summary pairwise dependencies by letting

$$\sigma_{\mathbf{XY},ij} = \int_{\mathbb{S}_{p-1}^+} \omega_i^\mathbf{X} \omega_j^\mathbf{Y} dH(\omega_\mathbf{X}, \omega_\mathbf{Y}), \quad \mathbf{\Sigma_{XY}} = (\sigma_{\mathbf{XY},ij}), i = 1, \ldots, p, j = 1, \ldots, q. \tag{2.8}$$

We refer to $\mathbf{\Sigma_{XY}}$ as the TPDM of $\mathbf{X}$ and $\mathbf{Y}$, which measures the tail dependence between these two random vectors, which we recall is in fact a positive semidefinite matrix.

So as to estimate this bivariate TPDM, let $\mathbf{x}_t, t = 1, \ldots, n_x$ and $\mathbf{y}_l, l = 1, \ldots, n_y$ be the transformed observations, for being regularly varying with tail index $\alpha = 2$ for all stations on day $t$. Then, the elements of the TPDM are estimated using pairs of $\mathbf{x}_t, \mathbf{y}_l$

elements. Let's now define the radial and angular components

$$r_{tl,ij} = \sqrt{x_{t,i}^2 + y_{l,j}^2}, \quad (\omega_{t,i}^{\mathbf{X}}, \omega_{l,j}^{\mathbf{Y}}) = \frac{(x_{t,i}, y_{l,j})}{r_{tl,ij}}.$$

We therefore estimate $\sigma_{\mathbf{XY},ij}$ as

$$\hat{\sigma}_{\mathbf{XY},ij} = \frac{2}{n_{ij,\text{exc}}} \sum_{t=1}^{n_x} \sum_{l=1}^{n_y} \omega_{t,i}^{\mathbf{X}} \omega_{l,j}^{\mathbf{Y}} \mathbb{I}(r_{tl,ij} > r_{00,ij}), \tag{2.9}$$

so that $\hat{\mathbf{\Sigma}}_{\mathbf{XY}} = (\hat{\sigma}_{\mathbf{XY},ij}), i = 1, \ldots, p, j = 1, \ldots, q$, where $r_{00,ij}$ is some high threshold for the radial components and $n_{ij,\text{exc}}$ is the number of observations whose $r_{ij}$ is greater than the corresponding high threshold. Again, as done in [6], we choose $r_{00,ij}$ to correspond to the 0.98 quantile of the data. We note that the indicator function $\mathbb{I}(r_{tl,ij} > r_{00,ij})$ forces the estimation to be based on the pairs with the largest radial component, which must be greater than $r_{0,ij}$. However, choosing this value involves the usual difficulties often found in choosing a threshold in an extreme value analysis.

Again, one issue with this pairwise estimate is that the estimate $\hat{\mathbf{\Sigma}}_{\mathbf{XY}}$ is not guaranteed to be positive definite, so in that case we will have to find the nearest positive definite matrix to the estimated $\hat{\mathbf{\Sigma}}_{\mathbf{XY}}$ and then perform all the corresponding analysis with it.

### 2.4.2 Maximum Covariance Analysis decomposition for extremes

As we saw, critical to the xPCA method is the fact of obtaining an ordered orthonormal basis for the $p$-dimensional positive orthant, analogously to PCA. This basis is obtained by applying the transformation $\mathbf{x} = \tau(\mathbf{y}) = \log(1 + e^{\mathbf{y}})$ componentwise to the vector $\mathbf{y} \in \mathbb{R}^p$. Remember that one important characteristic of this transformation is that $\tau(y) \approx y$ when $y \to \infty$, and therefore the transformation is negligible for large values both for $\mathbf{X}$ and $\mathbf{Y}$. Moreover, vector addition and scalar multiplication of a vector are defined via this transformation, and regular variation is preserved by this particular transformation as well.

Cooley and Thibaud [2] show that applying this transformation to the eigenvectors of the TPDM yields an orthonormal basis for the positive orthant, which is ordered by eigenvalues that yield the tail dependence scale explained by each eigenvector. Let $\mathbf{\Sigma}_{\mathbf{XY}} = \mathbf{U}\mathbf{D}\tilde{\mathbf{U}}^\top$ be the SVD of the TPDM, which exists because it is a completely positive matrix, where $\mathbf{D}$ is a diagonal matrix of eigenvalues with $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$, with $r = \min(p, q)$, and $\mathbf{U}$ is a matrix with columns $\mathbf{u}_i, i = 1, \ldots, p$ and $\tilde{\mathbf{U}}$ is a matrix with columns $\tilde{\mathbf{u}}_j, j = 1, \ldots, q$ being respectively the left and right singular vectors. Then the eigenvectors for the positive orthant are $\mathbf{e}_i = \tau(\mathbf{u}_i)$ and $\tilde{\mathbf{e}}_j = \tau(\tilde{\mathbf{u}}_j)$, $i = 1, \ldots, p$, $j = 1, \ldots, q$.

Let $\mathbf{x}_t, \mathbf{y}_t$ be respectively the realization of the regularly varying random vectors

$\mathbf{X}, \mathbf{Y}$ with associated Tail Pairwise Dependence Matrix $\mathbf{\Sigma_{XY}}$ at time $t$. Now, let $\mathbf{v}_t = \mathbf{U}^\top \tau(\mathbf{x}_t)$ and $\tilde{\mathbf{v}}_t = \tilde{\mathbf{U}}^\top \tau(\mathbf{y}_t)$. Then $\mathbf{v}_t$ and $\tilde{\mathbf{v}}_t$ are the vector of PCs for $\mathbf{x}_t$ and $\mathbf{y}_t$, that is, they are the vector of coefficients of the orthonormal eigenbasis in the positive orthant, i.e.

$$\mathbf{x}_t = \mathbf{v}_{t,1} \circ \mathbf{e}_1 \oplus \cdots \oplus \mathbf{v}_{t,p} \circ \mathbf{e}_p,$$

$$\mathbf{y}_t = \tilde{\mathbf{v}}_{t,1} \circ \mathbf{e}_1 \oplus \cdots \oplus \tilde{\mathbf{v}}_{t,q} \circ \mathbf{e}_q,$$

with the transformed multiplication and addition.

# CHAPTER 3

# Applications to climate data

## 3.1 Data

Once we have introduced and explained the theoretical mathematical framework of this thesis, it's time to explain the applications we have studied. First, in this section we will review the different type of data we have used for our study and the different preprocessing mechanisms we have applied to it. As we will see, the preprocessing depends both on the method applied, whether it is standard or its extreme variation, and the temporal resolution of the dataset.

### 3.1.1 Precipitation

In the study carried out by Daniel Cooley and Yujing Jiang [6], the authors focused on extreme precipitation events, and with this purpose they used daily precipitation data over the United States between the years 1950 and 2016 from the Global Historical Climatology Network (GHCN) dataset. The authors limited their investigation to the months of August, September and October, as they correspond to the height of hurricane activity in the Atlantic Ocean, when most extreme events take place on the east coast of the United States. The dataset they used contains data from 1140 stations and over 6164 days in the analyzed dataset.

Our first step was to reproduce the results from [6], so as to verify if the implementation of the method was correct. Inspired by their results, we decided to extend their analysis to different regions, in particular the Iberian Peninsula and the east Mediterranean coast of the same peninsula. With this purpose we used the E-OBS daily gridded meteorological dataset, which as its name indicates is a daily gridded land-only observational dataset over Europe whose station data are sourced directly from the European

National Meteorological and Hydrological Services between the years 1950 and 2019. Again, we focus on precipitation data, which in this dataset consists of total daily amount of rain, snow and hail measured as the height of the equivalent liquid water. This dataset also contains data from wind speed, temperature or humidity, which could be interesting for other kind of studies. We limited our investigation to the months from July to November, as they correspond to the months of maximum precipitation in the east Mediterranean coast of the Iberian Peninsula and the height of Cold Drop phenomena, when some extreme events over the Iberian Peninsula take place. The dataset has a spatial resolution of $0.25^\circ \times 0.25^\circ$ and consists of a daily time series of 25567 timepoints that goes, as said, from the year 1950 to 2019, and a number of 464 and 201 longitude and latitude gridpoints, respectively. However, as we choose to focus on the Iberian Penisula, and later of the east Mediterranean coast, we restrict our attention to latitudes between $[34, 45]$ and longitudes $[-10, 5]$ for the first case and latitudes between $[36.5, 43]$ and longitudes $[-4, 4]$ in the second case.

### 3.1.2 Sea surface salinity

The second goal of this work is finding patterns in different time series which explain a maximum fraction of the cross-covariance between two variables. In our case, we want to investigate the patterns between extreme precipitation and sea surface salinity. Hence, one of the considered variables will be the same precipitation data over the Iberian Peninsula we've just explained, while the second one will be sea surface salinity. With this purpose, we use Multi Observation Global Ocean Sea Surface Salinity and Sea Surface Density data from the Copernicus Marine Service, which consists of global analyses of the Sea Surface Salinity (SSS) and Sea Surface Density (SSD) obtained through a multivariate optimal interpolation algorithm that combines Soil Moisture Ocean Salinity satellite images and in situ salinity measurements with satellite SST information. The dataset has a weekly temporal resolution that goes back from the year 1993 until 2019 and a spatial resolution of $0.25^\circ \times 0.25^\circ$. Therefore, the dataset consists of a time series of 1408 timepoints and a resolution of 600 and 280 longitude and latitude gridpoints, respectively, but we restrict out attention to the data contain in latitudes between $[25, 55]$ and longitudes between $[-25, 20]$.

## 3.2 Preprocessing

As mentioned before, the different kind of preprocessing depends on the method applied and the temporal resolution of the dataset. Therefore, we will differentiate between the preprocessing applied to the precipitation and to the sea surface salinity, as in the first case the data recorded has daily resolution while in the second case it is weekly.

Contrary to the standard PCA framework, the regular variation framework described for the xPCA requires that each of the random variables considered is heavy-tailed with a common tail index $\alpha$. However, it is unsual for climate data to exhibit

this property, so transforming the marginal distributions is common to extreme studies. In fact, transforming the data is common outside the extreme analysis, as for example data may be transformed to be approximately Gaussian in many modeling frameworks. Hence, we will need to apply a transformation to the empirical distribution function of the data in the case of the extreme framework, while this is not necessary in the standard framework. Therefore, in each of this cases, we will also make a difference between the method applied, whether it is standard or extreme PCA and standard or extreme MCA.

### 3.2.1 Precipitation extremes

As done in [6], for comparing PCA and xPCA we choose to analyze precipitation data that corresponds to a 3-day moving average of the daily precipitation amount. If we denote by $z_{t,i}$ the observed precipitation on day $t$ at station $i$, we then define

$$x_{t,i}^{\text{orig}} = z_{t,i} + z_{t+1,i} + z_{t+2,i}$$

to be the moving averaged data, where the superscript simply denotes that $x_{t,i}^{\text{orig}}$ is in the original scale before further transformations of the marginal distributions. Selecting a 3-day moving average avoids some problems of a single extreme precipitation event being partially recorded over two separate days or being recorded in different but close stations, but this induces dependence in the $x_{t,i}^{\text{orig}}$, an important fact for the further analysis. Furthermore, we deciced to take a 3-day moving average as it was the original length for extreme precipitation events Daniel Cooley and Yujing Jiang wished to explore in [6], but extreme PCA could be applied to a moving averaged data of any duration of interest.

As mentioned before, the regularly varying framework leading to the Tail Pairwise Dependence Matrix assumes that each univariate marginal distribution is regular varying with tail index $\alpha = 2$, and so far we have only applied a moving average to the data so as to avoid some problems related to detection. Therefore, it is still necessary to make a transformation of the distribution function of the data to satisfy the condition of being regularly varying. To ensure that this is true for all data, we perform the transformation

$$x_{t,i} = G^{-1}\left(\hat{F}_i(x_{t,i}^{\text{orig}})\right), \tag{3.1}$$

where $G(x) = e^{-x^{-2}}$ is the cumulative distribution function, or cdf, of a Frechét random variable with scale 1 and tail index $\alpha = 2$, and $\hat{F}_i$ is the estimated cumulative distribution function of the data from station $i$, in other words the marginal distribution for station $i$. Therefore, this transforms the moving averaged data value $x_{t,i}^{\text{orig}}$ to a variable with Frechét distribution $x_{t,i}$, which is known to be regularly varying with tail index $\alpha = 2$. Now, as said $\hat{F}_i$ is an estimate of the cdf of precipitation at station $i$. The simplest method for obtaining $\hat{F}_i$ is to estimate it as the empirical cdf, which

is an unbiased estimate of the marginal distribution, and is generally represented as a step function. However, due to the dependence induced by the 3-day moving average, this kind of transformation is not ideal since it would ignore this dependence. Then, as performed in [6], we decided to estimate $\hat{F}_i$ as a linearly interpolated cdf, what still does not alleviate the dependence induced by the moving average. To finally solve this problem, we decided to take the average of three linearly interpolated cdf estimates, obtained from three subsequence given by the first, second and third terms of the moving average $x_{t,i}^{\text{orig}}$, respectively. This way of estimating $\hat{F}_i$ has been shown in the supplementary material of [6] to better retain the clustering generated by the moving average, and therefore the mentioned dependence. Finally, it is important to mention that the transformed data is treated as independent and identically distributed.

In summary, the first step both for PCA and xPCA will be to perform a moving average on the precipitation daily data. In the case of PCA it will also be necessary to standardize the data, and for the extreme framework we will need to perform transformation (3.1) to the data so as be regularly varying and have same scale and tail index $\alpha = 2$.

### 3.2.2 Covarying extremes in precipitation and sea surface salinity

The big difference between precipitation and the sea surface salinity data is that the first one has daily recordings while in the second case these are weekly. In the case of precipitation data it made sense to take a moving average of the data so as to avoid problems in the recording of extreme events, but in a dataset with weekly resolution it does not make sense to perform this first step of the preprocessing, as extreme events nearly never exceed a duration of a week. Furthermore, we cannot perform an analysis over two datasets that have daily and weekly resolution, so we will have to resample the precipitation dataset, taking the weekly sum so that the temporal resolution of both datasets coincides to be weekly. Furthermore, as stated before, in the case of xMCA the regularly varying framework leading to the Tail Pairwise Dependence Matrix assumes that each univariate marginal distribution is regular varying with tail index $\alpha = 2$, which is not necessarily satisfied by the sea surface salinity data nor the precipitation. Therefore, we again need to apply transformation (3.1) to both data, so as to transform them to a variable with Frechét distribution, again considering the interpolation of three linearly cumulative distribution functions of the data as estimate of the cdf $\hat{F}_i$. Again, the transformed data is treated as independent and identically distributed.

In summary, for the case of the MCA and xMCA the first step of the preprocessing, performing a moving average, can be skipped. Instead, we will have to resample the precipitation dataset to have weekly temporal resolution, as in the case of the sea surface salinity. In the case of MCA it will also be necessary to standardize the data, and for xMCA we will need to perform transformation (3.1) to the data so as to be regularly varying with same scale and tail index $\alpha = 2$.

# CHAPTER 4

# Results

The main goal of this thesis is to perform a comparison between the standard and extreme PCA methods, and similarly with the standard and extreme MCA methods. With this purpose, in the following section we will present the first six eigenvectors, or EOFs, and their corresponding PC time series for each region and method, respectively, after performing the transformations required and decomposing the corresponding matrices. So as to set notation, we will call eigenvectors, spatial patterns or EOFs the vectors obtained by decomposing the covariance matrix or the TPDM. On the other hand, we will refer to the projection of the data onto these eigenvectors, which is equivalent to the temporal evolution associated to the spatial patterns, as the PCs or PC time series. Finally, we will call a mode a pair of an EOF and its corresponding PC time series. In the later sections we will do a deeper discussion comparing both methods. The Python script to perform the analysis can be found in the Appendix B.

## 4.1 PCA and xPCA for precipitation

As we explained in Chapter 3, we applied PCA and xPCA methods for precipitation over the Iberian Peninsula and a particular region, the east Mediterranean coast. We limited our investigation to the months from July to November, as they correspond to the Cold Drop season in the east Mediterranean coast of Spain, when most extreme precipitation events take place there.

We remember that the eigenvectors, to which we will also refer as EOFs, $\mathbf{u}_i$, $i = 1, \ldots, p$ obtained through the standard decomposition of eigenvectors and eigenvalues of the estimate of the TPDM, $\hat{\mathbf{\Sigma}}_x$, are transformed to $\mathbf{e}_i = t(\mathbf{u}_i)$, $i = 1, \ldots, p$, which form and ordered orthonormal basis of the positive orthant $\mathbb{R}^p_+$, while no transforma-

tion is needed for the eigenvectors obtained using the standard PCA method after the decomposition of the covariance matrix. As in standard PCA, due to the orthogonality constraint of the EOFs, the interpretation is harder as we increase the number of EOFs considered in the interpretation, see for instance [13], so choosing a low number of eigenvectors which account for a great part of the variance will make our task easier.

In the presentation of the results we will look to each of the first six EOFs and their corresponding PCs, both for PCA and xPCA. Furthermore, we also include a red cross mark in the coefficient of the PC time series that corresponds to the day 12/09/2019, when historic torrential rainfall occurred in the southeast of the Iberian Peninsula. To facilitate latter comparison, we include the first PCA mode next to the corresponding first xPCA mode, and so on until the sixth mode.

### 4.1.1 Iberian Peninsula

The first region of interest is the Iberian Peninsula, which includes both countries of Portugal and Spain. As we have chosen to reduce the time series to the months from July to November, we will concentrate our attention on detecting extreme events in the east Mediterranean coast of Spain, that usually take place during these months. In Figure 4.1 we show the results obtained.

The first eigenvectors for both methods, Figures 4.1a, 4.1b, are completely negative, but as their corresponding PCs are negative they can actually be considered as completely positive. This is due to the fact that both the covariance matrix and the TPDM are completely positive matrices. Another noticeable feature is that there is little variation among the values for the different gridpoints. Taken by itself, this EOF tells us that precipitation could be quite homogeneous in all the Iberian Peninsula. However, we know this is not true, and in fact when combined with higher modes that give preference to extreme events on the east coast, this will allocate the precipitation to the Mediterranean coast and diminish the signal for the rest of the country.

Regarding the second and third EOFs for PCA, Figures 4.1c, 4.1e, we appreciate that now a dipole structure arises in both methods, which is natural in PCA decomposition as explained in [13]. Regarding PCA, the second EOF shows large negative values for the southeast coast of Spain and not so large for the rest of the Mediterranean coast, and moderate high levels in the northwest region of the Iberian Peninsula. Meanwhile, the third EOF shows the opposite trend, with high low negative values on the northeast coast of Spain and low positive values of the southwest of the Iberian Peninsula. This last mode could seem less interesting for our purposes, as it gives the same values to the Cantabrian and the Catalan coast, and usually extreme precipitation events during the considered dates mainly take place in the Mediterranean coast. Regarding the second and thir EOFs for xPCA, Figures 4.1d, 4.1f, the second EOF shows homogeneous large negative values in the whole east coast, and negative positive values for the western half of the Iberian Peninsula. As for the third EOF, we can see that it shows high positive values in the northern half of the peninsula and negative values in the southern half of

the peninsula. Therefore, if combined with negative coefficients, the second eigenvector of PCA and second and third eigenvectors of xPCA will allocate precipitation in the east Mediterranean coast, while the third eigenvector of PCA will do the same for the Cantabrian and northeastern coast.

As for the fourth and fifth and sixth EOFs, Figures 4.1g, 4.1h, 4.1i, 4.1j, 4.1k, 4.1l, we can see that they show a tripole structure, again what is expected in the PCA decomposition. Furthermore, we note that these three eigenvectors have a quite similar structure, except for a sign in the case of the fourth mode, contrary to what we have seen with the second and third eigenvectors. The fourth eigenvector of PCA, Figure 4.1g, shows positive values in the northwest and the southeast, while it shows negative values in nearly the whole east Mediterranean Coast and center of Spain. The same happens for the fourth mode of xPCA, Figure 4.1h, but with the values inverted. The fifth EOF both for PCA and xPCA, Figures 4.1i, 4.1j, shows the opposite behaviour, with negative values in the southwest coast and the northeast coast and positive values in the north of Spain and the east Mediterranean coast. Finally, the sixth EOF in both cases, Figures 4.1k, 4.1l, shows a gradient between the west coast, center of the peninsula and the east coast. This last mode could be the most interesting, as it shows a homogeneous behaviour along the whole Mediterranean coast.

### 4.1.2 East Mediterranean coast of the Iberian Peninsula

Now that we have investigated the structure of the different EOFs in the whole Iberian Peninsula, both for the standard and the extreme PCA methods, we will focus on the east Mediterranean coast of the Iberian Peninsula. We show, as in the previous case, the first six EOFs and their correspoding PC time series in Figure 4.2. Remember that we decided to investigate extreme precipitation events during the Cold Drop season over Spain, and those happen mainly in the east coast of the Iberian Peninsula, which is why we decided to focus on this region. The description of the EOFs will be very similar to the one made in the previous case and we will stick to interpreting the structure so as to compare the results later.

The first EOF for both cases, Figures 4.2a, 4.2b show again a completely negative homogeneous state along all the region for both cases, but so is their PCs. As a consequence, this mode will be useful when looking at positive patterns during the entire observation period. Again, if considered alone the result of these modes will be that precipitation should be quite homogeneous along all the region, but when combined with modes that give more weight to some regions then precipitation will be allocated there.

Regarding the second and third EOFs, we see that their spatial structure is very similar. The dipole structure that we mentioned in the analysis of the EOFs of the Iberian Peninsula again arises in these two EOFs. We see that, in both cases, the second eigenvector, see Figures 4.2c,4.2d, distinguishes between the north and south regions, while the third EOF, see Figures 4.2e, 4.2f, distinguishes between the interior

region and the coast. We also note that, although the structure is very similar, the weights given to specific small regions is slightly different.

So as to the fourth, fifth and sixth EOFs, they again show the tripole structure we mentioned before. In this case, the fourth eigenvectors, see Figures 4.2g, 4.2h, seem similar with small differences in the middle region of the tripole and also with the values inverted. Regarding the fifth eigenvector, Figures 4.2i, 4.2j, they are practically identical except for the fact that the xPCA mode gives slightly more negative weight to the southeast coast region, and therefore much homogeneous along the whole coast. Finally, the sixth eigenvectors, Figures 4.2k, 4.2l, are quite similar in structure again and we observe that the values are inverted.

### 4.1.3 Reconstruction

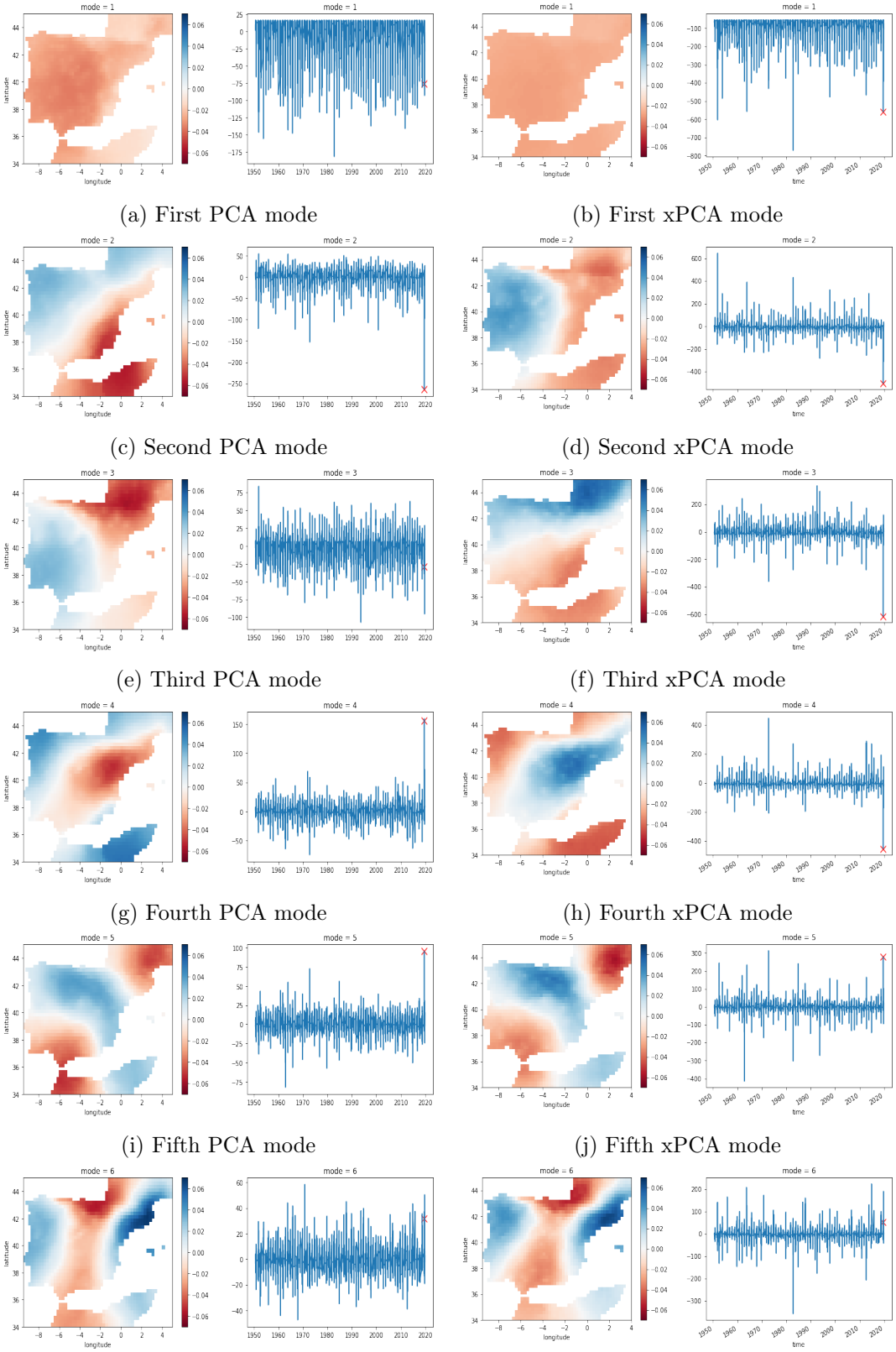
We finally include in Figure 4.3 a reconstruction of the precipitation over the whole Iberian Peninsula during the day 12/09/2019 using the first six modes of PCA and xPCA. Note that here we mention modes, as for doing the reconstruction we need both the EOFs and their corresponding PCs. In the Figure 4.3a we include the reconstruction using the PCA modes, while in the Figure 4.3b the reconstruction for the case of the xPCA modes. On the left of each figure we have the preprocessed data and on the right the reconstruction. As the preprocessed data is different, it's clear that the reconstruction will also differ. In Figure 4.4 we include a reconstruction of the precipitation during the same day but using all modes. As expected for both modes, both reconstructions for PCA, see Figure 4.4a and xPCA, see Figure 4.4b are nearly perfect. Finally, in the Figure 4.5a we have the variance and the cumulative variance explained by the first six modes for PCA, while in the Figure 4.5b we include the tail dependence scale explained by the first six modes. Something remarkable is that the cumulative tail dependence scale explained by the xPCA method is slightly higher than the variance explained by PCA.

Analogously, we include in Figures 4.6 a reconstruction of the precipitation over the east Mediterranean coast during the day 12/09/2019 using the first six modes of PCA and xPCA and in 4.7 a reconstruction using all modes. Finally, in Figure 4.8 we include the variance and the tail dependence scale explained by the first six modes of PCA and xPCA, respectively.

## 4.2 MCA and xMCA for precipitation and sea surface salinity

The first singular vectors obtained from xMCA of weekly precipitation and sea surface salinity data exhibit a rather homogeneous pattern for both cases, as we can see in Figure 4.9b. The values are slightly larger for Portugal and Galicia while the coefficients for sea surface salinity do not show any particular regional focus. Regarding the first

singular vectors of MCA, see Figure 4.9a, we appreciate a similar homogeneous structure for precipitation, but as for the sea surface salinity we see that the homogeneous pattern from the xMCA EOF in sea surface salinity is lost and its structure seems much more complicated, with a negative region in the Cantabrian sea and northeastern coast of Spain and positive in the southeast coast. Regarding the second pair of singular vectors, we see that in the case of MCA, see Figure 4.9c, now there is a strong negative region in the west of the Iberian Peninsula which corresponds to the region of highest precipitation, while the east and south of Spain, the driest zones, now are positive. If we now pay attention to the sea surface salinity mode, we can see that the Mediterranean sea has negative values, while Atlantic ocean has positive values. Therefore, we can appreciate an inverse relationship that maps regions of high precipitation to low values of sea surface salinity and dry regions to high values of sea surface salinity. If we now look at the xMCA EOFs, see Figure 4.9d, we can see that the Mediterranean coast has strong positive values while the center and west of the Peninsula have negative values that are stronger in the zones of highest precipitation. Furthermore, we can see again that the east coast has negative values, which are stronger now, while the Atlantic ocean has low intensity values. Therefore, we note that the inverse relationship that we had mentioned before still holds and that now that the precipitation values are higher, the sea surface salinity values are higher too.




(a) First PCA mode

(b) First xPCA mode

(c) Second PCA mode

(d) Second xPCA mode

(e) Third PCA mode

(f) Third xPCA mode

(g) Fourth PCA mode

(h) Fourth xPCA mode

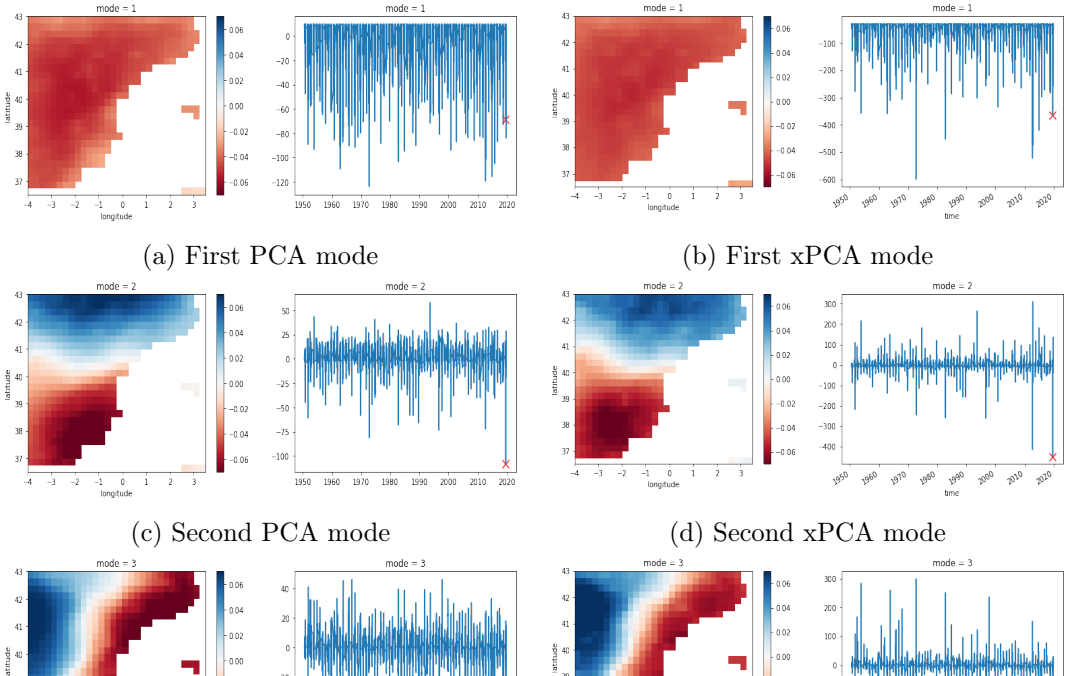
(i) Fifth PCA mode

(j) Fifth xPCA mode

(k) Sixth PCA mode

(l) Sixth xPCA mode

Figure 4.1: First six PCA and xPCA modes (EOFs and PCs) for precipitation over Spain during the months from July to November and during the years 1950-2019. The red cross corresponds to the day 12/09/2019.




(a) First PCA mode

(b) First xPCA mode



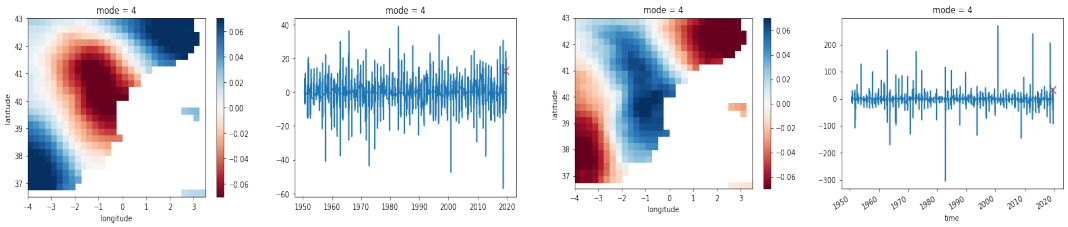

(c) Second PCA mode

(d) Second xPCA mode



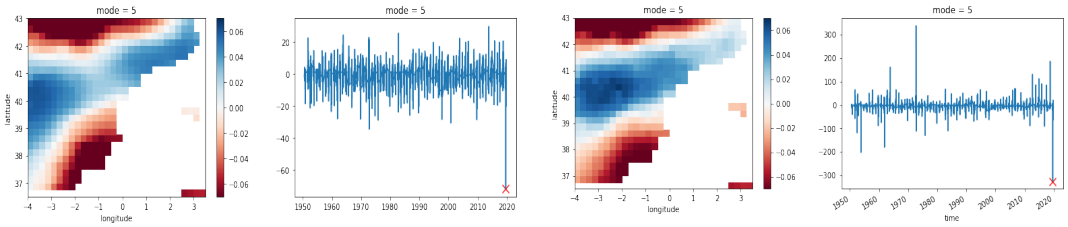

(e) Third PCA mode

(f) Third xPCA mode



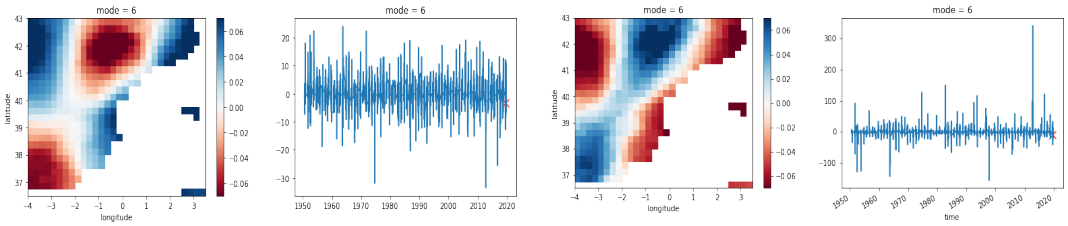

(g) Fourth PCA mode

(h) Fourth xPCA mode




(i) Fifth PCA mode

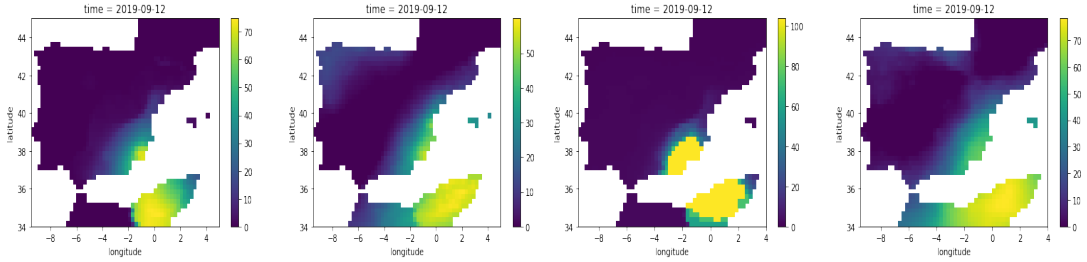
(j) Fifth xPCA mode




(k) Sixth PCA mode

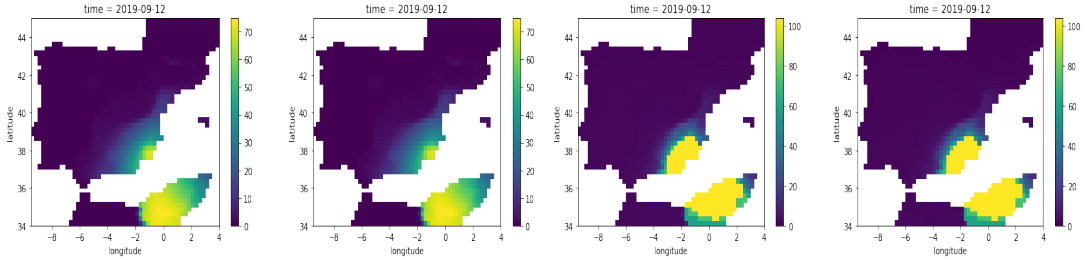
(l) Sixth xPCA mode

Figure 4.2: First six PCA and xPCA modes (EOFs and PCs) for precipitation over the Mediterranean east coast during the months from July to November and during the years 1950-2019. The red cross corresponds to the day 12/09/2019.



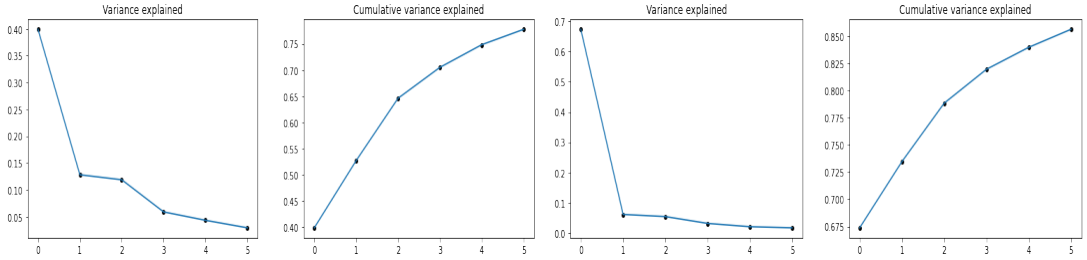

(a) Reconstruction using PCA (b) Reconstruction using xPCA

Figure 4.3: Reconstruction of precipitation during the day 12/09/2019 over the Iberian Peninsula using the six first modes.



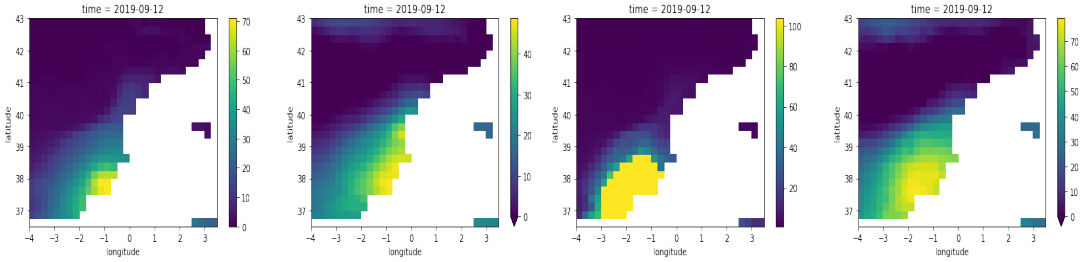

(a) Reconstruction using PCA (b) Reconstruction using xPCA

Figure 4.4: Reconstruction of precipitation during the day 12/09/2019 over the Iberian Peninsula using all modes.



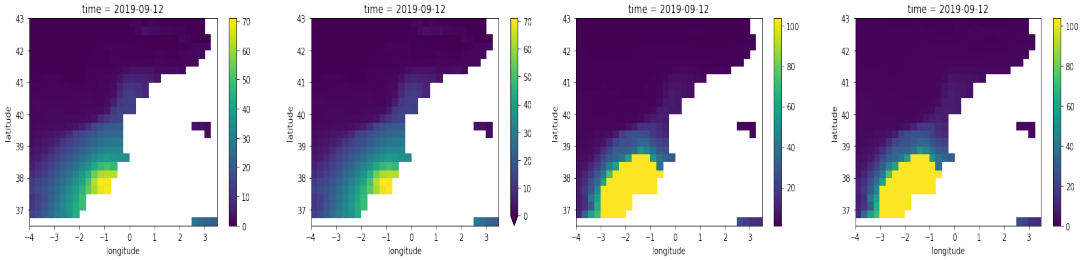

(a) Variance explained by PCA (b) Tail dependence scale explained by xPCA

Figure 4.5: Variance and tail dependence explained by the first six modes for precipitation during the day 12/09/2019 over the Iberian Peninsula.



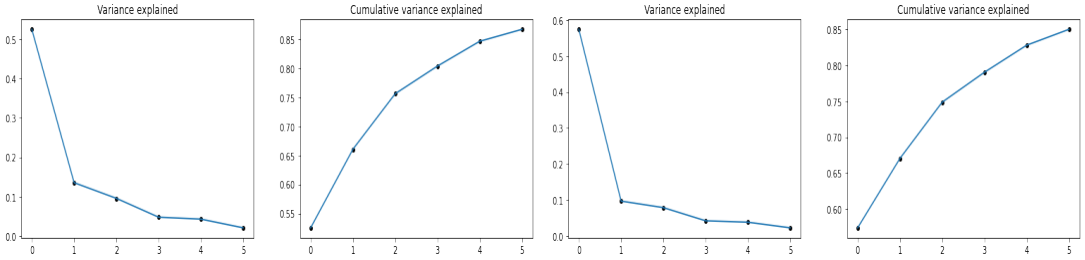

(a) Reconstruction using PCA (b) Reconstruction using xPCA

Figure 4.6: Reconstruction of precipitation during the day 12/09/2019 over the east Mediterranean coast using the six first modes.



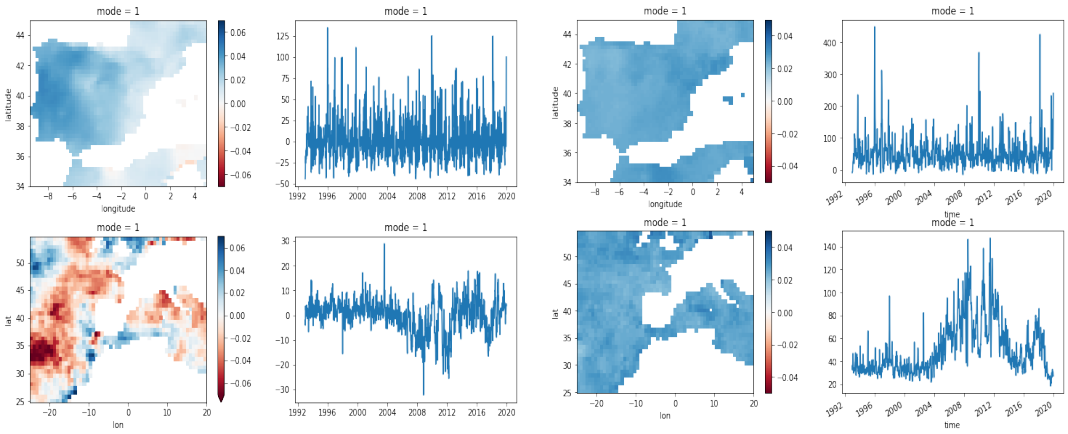

(a) Reconstruction using PCA (b) Reconstruction using xPCA

Figure 4.7: Reconstruction of precipitation during the day 12/09/2019 over over the east Mediterranean coast using all modes.




(a) Variance explained by PCA (b) Tail dependence scale explained by xPCA

Figure 4.8: Variance and tail dependence explained by the first six modes for precipitation during the day 12/09/2019 over the east Mediterranean coast.




(a) First MCA mode

(b) First xMCA mode

(c) Second MCA mode

(d) Second xMCA mode

Figure 4.9: Two first pairs of singular vectors for MCA and xMCA for precipitation over the Iberian Peninsula and sea surface salinity over the Atlantic ocean and Mediterranean sea for all months during the years 1993-2019.

# CHAPTER 5

# Discussion

Throughout this thesis we have reviewed one of the most studied multivariate analysis methods, known as PCA, whose main objective is to reduce the dimensionality of a set of variables by finding a new set of variables maximizing the variance of the given data set. In addition, we investigated a generalization of PCA, called MCA, which maximizes the cross-covariance between two different sets of variables and which reduces to PCA in case both data sets are equal. The main goal of this work was to compare these results to a recently developed extension, called xPCA [6], allowing to work directly with extreme values. An additional contribution with this work has been to naturally extent the framework described in [6] to the analysis for co-varying extremes, giving rise to what we called xMCA.

First, we applied the PCA and xPCA methods for precipitation data over two different regions, the Iberian Peninsula and the east Mediterranean coast of this Peninsula. The objective with this was to obtain the EOFs, or eigenvectors of the covariance standard matrix and the TPDM, so as to compare their structure and PCs, the time series of their coefficients of the projection of data onto the eigenvectors. We had as a reference the study carried out by Cooley and Jiang in [6], in which they studied the decomposition of the TPDM for the case of precipitation over the contiguous United States, placing special emphasis on extreme events during the hurricane season on the east coast.

In our case we decided to focus on precipitation over the Iberian Peninsula and reproduce the same analysis from scratch, but focusing on the Cold Drop season, when most extreme precipitation events take place on the east Mediterranean coast. In Figure 4.1 we presented the figures corresponding to the first six EOFs and their corresponding PC time series. As we have mentioned previously, as the number of modes increases their interpretation becomes more difficult due to the orthogonality constraint, so lim-

iting our study to a low number of EOFs and PCs is the best option. We previously observed that the first EOF for both PCA and xPCA are similar and homogeneous throughout the whole region, the second and third EOFs are different or at least seem to be inverted, and finally again the fourth, fifth and sixth EOFs are very similar. These results are quite surprising, because the spatial structure of the different EOFs is all in all quite similar, except possibly for the second and third. It is very difficult to say that we actually expected both eigendecomposition to yield EOFs with a similar spatial structure, what could be happening because the data has some special properties, notably that variance and tail pairwise dependencies have similar modes of variations. However, if we look at the coefficient of the PC times series for each EOF and for the day 12/09/2019 marked with a red cross, when a storm caused by a Cold Drop left historical torrential rains in the south of the Valencian Community and Murcia, we can see that the coefficients are far more large for the case of the extreme EOFs than for the standard EOFs. In particular, if we look at coeficcient of the xPCA PCs for this particular day, see Figures 4.1b, 4.1d, 4.1f, 4.1h, 4.1j, we note that in all of them this coefficient is a value that stands out with even very high values, while in the case of PCA only in the PCs from Figures 4.1d, 4.1h, 4.1j we can detect this extreme event. Therefore, we note that in the case of the PCA modes, these coefficients tend to be much more homogeneous throughout all the time series, and although one can note when heavy rainfall took place looking at some PCs its value does not differ much from the whole time series. Meanwhile, in the case of the xPCA PC time series the values for the extreme precipitation events are really high, allowing us to detect when historic precipitation events took place. This is something which was expected, as the EOFs from xPCA are found by diagonalizing a matrix which accounts for tail pairwise extreme dependence, instead of simply variance, so we note that these modes give more importance to extreme events, specially in the time series of their PCs coefficients.

As we decided to focus on studying extreme precipitation events during the months from July to November and these usually take place in the Mediterranean coast, we decided to make a zoom and apply PCA and xPCA over the east Mediterranean coast, what include the Valencia Community, Catalonia and Murcia. In Figure 4.2 we presented the first six EOFs for both PCA and xPCA methods and their corresponding PC time series, so now we can carry out a similar discussion as with the case of the whole Iberian Peninsula. Again, we include a cross mark in the PC time series for each EOF for the particular case of the day 12/09/2019. Before, we noticed that the first three EOFs, both for PCA and xPCA, surprisingly have a very similar spatial structure and that the fourth, fifth and sixth differ slightly, but keep again similar spatial patterns. Moreover, we can see that the typical structure of monopole for the first mode, a dipole for the second and third and a tripole for the last three is identical in all the EOFs. As for the case of the whole Iberian Peninsula, we note that the first six eigenvectors keep a very similar spatial structure, so by looking at the spatial patterns we are not able to detect special patterns for detecting extreme events. However, if we now look at the

PCs, and in particular to the values that are marked with the red cross as it is the day we're interesd in, we can see again that the values corresponding to the time series of the xPCA stand out much more than the ones from the PCA time series. In particular, by looking at the PCs in Figures 4.2b, 4.2d, 4.2f, 4.2j we are able to detect a special extreme event during the day with the crossed mark, while in the case of PCA only the PCs in Figures 4.2c, 4.2i gives as an intuition of the mentioned event. Again, we note that throughout the whole time series, the values of the PC time series of the PCA eigenvectors are quite homogeneous, but for the xPCA the values quickly shoot up to reach very high values for the corresponding day of historic precipitation. Therefore, as we stated before, although the spatial structure of the modes for both methods is very similar, by looking at the PC time series we note that the xPCA method can help us detect much easier and with much more convictino these extreme precipitation events.

Therefore, the PCA variant method for extremes studied here and presented in [6] offers similar spatial patterns as PCA, but when it comes to detecting extreme events it becomes a more reliable method.

As we mentioned in the results, we also decided to study one of the PCA variants known as Maximum Covariance Analysis, or MCA, which is related to PCA through the decomposition known as Singular Value Decomposition and that it's a very useful tool for detecting coupled patterns between two different geophysical fields. With this objective, we used two different types of data, on the one hand daily precipitation over the Iberian Peninsula and on the other sea surface salinity over the Atlantic Ocean and the Mediterranean Sea. Due to the complexity involved in analyzing different covariant trends, we decided to restrict our study to the first two modes for both data fields. In this case we will not focus on any day in particular, but we will limit ourselves to a brief analysis of the spatial structure and the temporal dynamics of the different modes. Thus, as for the first precipitation EOF of MCA we noted that it is practically homogeneous, as also are the first precipitation and sea surface salinity modes of xMCA. However, we noted that the first salinity mode for MCA already has a complex structure, from which we can observe that has positive values where it rains the least and negative values where it rains the most, so we could expect some inverse relationship between precipitation and sea surface salinity. Regarding the PC time series, at first glance no pattern can be identified in the time evolution, beyond some abrupt variation around the year 2012 that may correspond to some peak in the precipitation time series. However, as we know, it does not rain homogeneously throughout the Iberian Peninsula, so it seems much more worthwhile analyzing changes in sea surface salinity with higher precipitation modes that distinguish between geographic areas of the Iberian Peninsula where it rains more or less. This is the case of the second mode of precipitation of MCA, where the region of Galicia differs from the rest. Regarding the first mode of sea surface salinity, the positive values have been accentuated now in the regions where it rains the most and negative in the regions where it rains the less, so the previous inverse possible relationship has disappeared. The time series allows us to see an important

change from the year 2014, which could correspond to the peak in precipitation after the year 2016. To analyze more in depth, a good idea would be to introduce a time lag to see the relationship between these changes. Finally, regarding the second mode of precipitation and salinity, we observe that once again a dipolar structure appears that differentiates between the east of the peninsula, where the most extreme phenomena tend to appear, and the west, where it usually rains more homogeneously. Regarding salinity, the negative values have been accentuated in the Mediterranean Sea while in the Atlantic Ocean they are either negative or close to zero. We could speculate again about some relationship between these two modes, since again we observe variations around the year 2014 that could correspond to some peak in the precipitation time series, but again to study it in more detail it would be necessary to introduce either a time lag or do an analysis with higher EOFs. Other option for the observed peaks are an increased variability of the sea surface salinity data. Since it is a reanalysis product, it may be affected by the sudden introduction of in-situ observations after the start of the ARGO programme in the year 2005. Therefore, the salinity data may have just very strong biases, as it has been studied in [14].

In summary, it stands out that there is a big difference between the different modes for both precipitation and sea surface salinity between MCA and xMCA, especially after seeing that in the case of PCA and xPCA they were very similar. Furthermore, we've seen that it's very difficult trying to detect covarying patterns by simply applying MCA or xMCA methods. In fact, there are a lot of variables that could interfere such as for example seasonal cycles, El Niño Southern Oscillation, or even climate change could be affecting both or at least one of these variables, so studying the effect of time lags or simplifying the dataset would be a good idea which could help identify some patterns.

# CHAPTER 6

# Conclusions

Throughout this thesis we have reviewed one of the most well known multivariate analysis methods, known as Principal Component Analysis, whose main objective is to reduce the dimensionality of a set of variables by finding a new set of variables maximizing the variance of the given data set, and we have also investigated a generalization of PCA, called Maximum Covariance Analysis, which maximizes the cross-covariance between two different sets of variables and which reduces to PCA in case both data sets are equal. The main goal of this work was to compare the results of the application of PCA to a recently developed extension, called Extreme Principal Component Analysis [6], allowing to work directly with extreme values. An additional contribution we have done in this work is to naturally extent [6] to the analysis for co-varying extremes, giving rise to what we called Extreme Maximum Covariance Analysis, allowing as to compare als between the standar and extreme MCA method.

Once studied the mathematical theory framework behind these four methods, we implemented a code in Python so as to carry out an analysis. For this, we considered two types of data: daily precipitation over the Iberian Peninsula and sea surface salinity ove the Atlantic Ocean and the Mediterranean sea. The comparison between PCA and xPCA leads us to partly unexpected results, because we obtain the same spatial decomposition, EOFs, for both matrices. However, looking more deeply we observe that the extreme PCA method actually detects extreme precipitation events more directly as values that stand out in the PC time series in five of the first six PCs, while in the case of PCA method we can only detect this extreme precipitation event in two of the first six PCs, for the particular case of the torrential rainfall that happened during the day 12/09/2019. Therefore, the conclusion we obtain from this first analysis is that the xPCA methods detects more directly and more convincingly the extreme events. Furthermore, as expected both methods reconstruct perfectly the original data.

Regarding the comparison between MCA and xMCA, the results obtained were not so remarkable, so we were not able to draw great conclusions from them. It is true that we can detect some inverse relationships between the sea surface salinity and the precipitation, but as data may be affected by strong bias because of the sudden introduction of in-situ observations after the start of the ARGO programme in the year 2005. Therefore, we leave the door open to introducing time lags in the temporal evolution of the sea surface salinity or just reducing the time of comparison between the two variables from 2005 until now. Furthermore, it would also be interesting studying the relationship between other type of data such as for example sea surface temperature and precipitation.

# APPENDIX A

## $\mathbb{X}^p$ is a vector space with an inner product

Here we will prove that $\mathbb{X}^p$ is a vector space with an inner product well defined. With this purpose, we first show that the 10 conditions for $\mathbb{X}^p$ begin a vector space are met. Remember that throughout all the proof the transformation $t$ is applied componentwise, i.e. for $\mathbf{x} \in \mathbb{X}^p$ we have that $\mathbf{y} = t(\mathbf{x}) \in \mathbb{R}^p$. Now, vector addition is closed

$$\mathbf{x}_1 \oplus \mathbf{x}_2 = t(t^{-1}(\mathbf{x}_1) + t^{-1}(\mathbf{x}_2)) = t(\mathbf{y}_1 + \mathbf{y}_2) \in \mathbb{X}^p,$$

vector addition is commutative

$$\mathbf{x}_2 \oplus \mathbf{x}_1 = t(t^{-1}(\mathbf{x}_2) + t^{-1}(\mathbf{x}_1)) = t(\mathbf{y}_2 + \mathbf{y}_1) = t(\mathbf{y}_1 + \mathbf{y}_2) = \mathbf{x}_1 \oplus \mathbf{x}_2 \in \mathbb{X}^p,$$

vector addition is associative

$$\begin{aligned}(\mathbf{x}_1 \oplus \mathbf{x}_2) \oplus \mathbf{x}_3 &= t(t^{-1}(t(t^{-1}(\mathbf{x}_1) + t^{-1}(\mathbf{x}_2)) + t^{-1}(\mathbf{x}_3)) = \\ &= t(t^{-1}(\mathbf{x}_1) + t^{-1}(\mathbf{x}_2) + t^{-1}(\mathbf{x}_3)) = \\ &= t(t^{-1}(\mathbf{x}_1) + t^{-1}(t(t^{-1}(\mathbf{x}_2) + t^{-1}(\mathbf{x}_3)))) = \\ &= \mathbf{x}_1 \oplus (\mathbf{x}_2 + \mathbf{x}_3),\end{aligned}$$

there exists $0 \in \mathbb{X}^p$, defined by $0 = t(0)$, then

$$\mathbf{x} \oplus 0 = t(t^{-1}(\mathbf{x}) + t^{-1}(t(0))) = t(t^{-1}(\mathbf{x} + 0) = t(t^1(\mathbf{x})) = \mathbf{x},$$

for every $\mathbf{x} \in \mathbb{X}^p$, there exists an element we denote by $-\mathbf{x}$ such that $\mathbf{x} \oplus -\mathbf{x} = 0$. Define $-\mathbf{x} = t(-t^{-1}(\mathbf{x}))$, then

$$\mathbf{x} \oplus -\mathbf{x} = t(t^{-1}(\mathbf{x}) + -t^{-1}(\mathbf{x})) = t(0) = 0,$$

scalar multiplication is closed

$$c \circ \mathbf{x} = t(ct^{-1}(\mathbf{x})) \in \mathbb{X}^p,$$

scalar multiplication is associative

$$c_1 \circ (c_2 \circ \mathbf{x}) = c_1 \circ (t(c_2 t^{-1}(\mathbf{x}))) = t(c_1 t^{-1}(t(c_2 t^{-1}(\mathbf{x})))) = t(c_1 c_2 t^{-1}(\mathbf{x})) = (c_1 c_2) \circ \mathbf{x},$$

there exists a multiplicative identity

$$1 \circ \mathbf{x} = t(1t^{-1}(\mathbf{x})) = \mathbf{x},$$

scalar multiplication is distributive over vector addition

$$\begin{aligned} c \circ (\mathbf{x}_1 \oplus \mathbf{x}_2) &= c \circ t(t^{-1}(\mathbf{x}_1) + t^{-1}(\mathbf{x}_2)) = t(ct^{-1}(t(t^{-1}(\mathbf{x}_1) + t^{-1}(\mathbf{x}_2)))) = \\ &= t(c(t^{-1}(\mathbf{x}_1) + t^{-1}(\mathbf{x}_2))) = t(ct^{-1}(\mathbf{x}_1) + ct^{-1}(\mathbf{x}_2)) = c \circ \mathbf{x}_1 \oplus c \circ \mathbf{x}_2, \end{aligned}$$

and finally multiplication is distributive over addition of scalars

$$(c_1 + c_2) \circ \mathbf{x} = t((c_1 + c_2)t^{-1}(\mathbf{x})) = t(c_1 t^{-1}(\mathbf{x}) + c_2 t^{-1}(\mathbf{x})) = c_1 \circ \mathbf{x} \oplus c_2 \circ \mathbf{x}.$$

Now, we show that

$$\langle \mathbf{x}, \mathbf{x} \rangle = \sum_{j=1}^{p} t^{-1}(x_j) t^{-1}(x_j) = \sum_{j=1}^{p} y_j^2 \tag{A.1}$$

is a well defined inner product. Now, to show positiviness we simply have to look at equation (A.1), because it is clear that $\langle \mathbf{x}, \mathbf{x} \rangle > 0$ if any $y_j \neq 0$ and this happend if and only if $\mathbf{x} \neq 0$, and therefore $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if all $y_j = 0$ and this happens if and only if $\mathbf{x} = 0$. Finally, we have that this inner product is symmetric

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = \sum_{j=1}^{p} t^{-1}(x_{1j}) t^{-1}(x_{2j}) = \sum_{j=1}^{p} t^{-1}(x_{2j}) t^{-1}(x_{1j}) = \langle \mathbf{x}_2, \mathbf{x}_1 \rangle,$$

is linear with respect addition

$$\begin{aligned} \langle \mathbf{x}_1 \oplus \mathbf{x}_2, \mathbf{x}_3 \rangle &= \sum_{j=1}^{p} t^{-1}(t(t^{-1}(x_{1j}) + t^{-1}(x_{2j}))) t^{-1}(x_{3j}) = \\ &= \sum_{j=1}^{p} (t^{-1}(x_{1j}) + t^{-1}(x_{2j})) t^{-1}(x_{3j}) = \\ &= \sum_{j=1}^{p} t^{-1}(x_{1j}) t^{-1}(x_{3j})) + \sum_{j=1}^{p} t^{-1}(x_{2j}) t^{-1}(x_{3j})) = \\ &= \langle \mathbf{x}_1, \mathbf{x}_3 \rangle + \langle \mathbf{x}_2, \mathbf{x}_3 \rangle, \end{aligned}$$

and respect multiplication by scalars

$$\begin{aligned}\langle c \circ \mathbf{x}_1, \mathbf{x}_2 \rangle &= \sum_{j=1}^{p} t^{-1}(t(ct^{-1}(x_{1j})))t^{-1}(x_{2j}) = \\ &= \sum_{j=1}^{p} ct^{-1}(x_{1j})t^{-1}(x_{2j}) = \\ &= c\sum_{j=1}^{p} t^{-1}(x_{1j})t^{-1}(x_{2j})) = c\langle \mathbf{x}_1, \mathbf{x}_2 \rangle.\end{aligned}$$

# APPENDIX B

## Code

In this chapter we will introduce the code we have used for implementing the four methods we have studied. First of all, we must load the different packages that we are going to use.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import xarray as xr
import math
import cartopy.crs as crs
```

PCA and MCA methods are classes already defined in Python through the `scikit learn` library. However, we decided to implement the most basic functions by ourselves, such as standardizing the data, calculating the eigenvectors or EOFs through a SVD, calculating the PCs time series as the projection, the reconstruction of the data and the explained variance and cumulative variance. Next, we present the two Python classes `PCA` and `MCA` we defined.

```
class PCA:
    def __init__(self, X):
        self.sigma = np.std(X, axis=0)
        self.mu = np.mean(X, axis=0)
        self.X_std = (X - self.mu) / self.sigma

        self.cov_mat = self.X_std.T @ self.X_std / (self.X_std.shape[0] - 1)

        self.X_eofs, self.eigen_values, _ = np.linalg.svd(self.cov_mat)
```

```
        self.X_projected = np.dot(self.X_std, self.X_eofs)

    def eofs(self, n_X):
        return self.X_eofs[:, 0:n_X].T

    def projected(self, n_X):
        return self.X_projected[:, 0:n_X]

    def reconstructed(self, n_X):
        return np.dot(self.X_projected[:, 0:n_X], (self.X_eofs[:, 0:n_X]).T) *
            self.sigma + self.mu

    def explained_var(self):
        return self.eigen_values / np.sum(self.eigen_values)

    def explained_cumvar(self):
        return np.cumsum(self.eigen_values) / np.sum(self.eigen_values)

class MCA:
    def __init__(self, X, Y):
        self.sigma_X = np.std(X, axis=0)
        self.mu_X = np.mean(X, axis=0)
        self.X_std = (X - self.mu_X) / self.sigma_X

        self.sigma_Y = np.std(Y, axis=0)
        self.mu_Y = np.mean(Y, axis=0)
        self.Y_std = (Y - self.mu_Y) / self.sigma_Y

        self.cov_mat = self.X_std.T @ self.Y_std / (self.X_std.shape[0] - 1)

        self.X_eofs, self.eigen_values, self.Y_eofs =
            np.linalg.svd(self.cov_mat, full_matrices=False)
        self.Y_eofs = self.Y_eofs.T

        self.X_projected = np.dot(self.X_std, self.X_eofs)
        self.Y_projected = np.dot(self.Y_std, self.Y_eofs)

    def eofs(self, n_X, n_Y):
        return self.X_eofs[:, 0:n_X].T, self.Y_eofs[:, 0:n_Y].T

    def projected(self, n_X, n_Y):
        return self.X_projected[:, 0:n_X], self.Y_projected[:, 0:n_Y]

    def reconstructed(self, n_X):
        X_reconstructed = np.dot(self.X_projected[:, 0:n_X], self.X_eofs[:,
            0:n_X].T) * self.sigma_X[0:n_X] + self.mu_X[0:n_X]
        Y_reconstructed = np.dot(self.Y_projected[:, 0:n_Y], self.Y_eofs[:,
```

```
            0:n_Y].T) * self.sigma_Y[0:n_Y] + self.mu_Y[0:n_Y]
        return X_reconstructed, Y_reconstructed

    def explained_var(self):
        return self.eigen_values / np.sum(self.eigen_values)

    def explained_cumvar(self):
        return np.cumsum(self.eigen_values) / np.sum(self.eigen_values)
```

Regarding the xPCA method, we implemented the R functions built by Daniel Cooley and Yujing Jiang in [6]. The functions `take_month` and `take_month_year` are used to have access to the temporal data of the events. The preprocessing consists of applying functions `ma` for the moving average of the data and `ECDF_smoothing` and `to_alpha_2` so as to first get the interpolation of the cumulative distribution function and then transform each random variable to be regularly varying with tail index $\alpha = 2$. So as to estimate the TPDM, we implemented the functions `decls` for declustering the extreme events from the whole time series and finally the function `rw_Sigma` gives us the estimate of the TPDM. We also defined the transformation functions to the positive orthant and its inverse, `trans` and `inv_trans`, respectively. And finally `pc_one`, `pc` and `reconstruct` gives us the PC time series and the reconstruction of the data. In the case of MCA, the only difference comes from substituing one of the $x$ vectors in the function `rw_Sigma` by $y$ and avoid using the declustering `decls` function.

```
def take_month(time, Month):
    year = np.zeros(time.shape[0])
    month = np.zeros(time.shape[0])
    for i in range(time.shape[0]):
        ymd = time[i].rsplit("-", 2)
        year[i] = int(ymd[0])
        month[i] = int(ymd[1])

    return np.where(np.isin(month, Month))[0]

def take_month_year(time):
    year = np.zeros(time.shape[0])
    month = np.zeros(time.shape[0])
    for i in range(time.shape[0]):
        ymd = time[i].rsplit("-", 2)
        year[i] = int(ymd[0])
        month[i] = int(ymd[1])

    return month, year

def ma(x, k=3):
    res = np.zeros_like(x)
```

```
    aux = np.append(np.zeros((math.floor(k / 2), x.shape[1])) * np.nan, x,
        axis=0)
    for i in range(x.shape[0]):
        res[i, np.where(~np.isnan(x[i]))] = np.nanmean(aux[i:i+k,
            np.where(~np.isnan(x[i]))], axis=0)
        res[i, np.where(np.isnan(x[i]))] = np.nan

    return res

def ECDF_smoothing(x, k=3):
    each = np.zeros((x.shape[1], x.shape[0], k))
    for i in range(k):
        idx = np.arange(i, x.shape[0], k)
        for j in range(x.shape[1]):
            ni = np.sort(x[idx, j])[~np.isnan(np.sort(x[idx, j]))]
            if len(ni) > 0: each[j, :, i] = np.interp(x[:, j], ni,
                np.arange(1, ni.shape[0] + 1) / (ni.shape[0] + 1), left = 0,
                right = 1-1E-9)
            else: each[j, :, i] = np.nan

    return np.mean(each, axis = 2).T

def to_alpha_2(x):
    return np.sqrt(1 / - np.log(x))

def decls(x, th, k):
    id_big = np.where(x.T.flatten() > th)[0]
    start = id_big[np.append(np.zeros(1), np.where(np.diff(id_big) >= k)[0] +
        1).astype(int)]
    end = np.append(id_big[np.where(np.diff(id_big) >= k)[0]], id_big[-1])
    id_res = np.repeat(0, start.shape[0])
    for i in range(len(start)):
        temp = x.T.flatten()[start[i]:end[i] + 1]
        id_res[i] = np.where(temp == np.nanmax(temp))[0][0] + start[i]

    return id_res

def rw_Sigma(x, u=0.98):
    Sigma = np.zeros((x.shape[1], x.shape[1]))
    for i in range(x.shape[1]):
        if i % 100 == 0: print(i)
        for j in range(x.shape[1]):
            r = np.sqrt(x[:, i] ** 2 + x[:, j] ** 2)
            w1 = x[:, i] / r
            w2 = x[:, j] / r
            idx = decls(r, th = np.nanquantile(r, u), k = 3)
            Sigma[i, j] = np.nansum(w1[idx] * w2[idx]) / len(idx) * 2
```

```
    return Sigma

def trans(x):
    v = np.log(1 + np.exp(x))
    v[~np.isfinite(v)] = x[~np.isfinite(v)]
    v[x < -20] = np.exp(x[x < -20])

    return v

def inv_trans(v):
    x = np.log(np.exp(v) - 1)
    idx = np.logical_and(~np.isfinite(x), ~np.isnan(x), v > 1)
    x[idx] = v[idx]

    return x

def pc_one(i, U, invX):
    idxna = np.isnan(invX[i])

    return np.matmul(U[~idxna].T, invX[i, ~idxna])

def pc(U, invX):
    res = np.zeros_like(invX)
    for i in range(invX.shape[0]):
        res[i] = pc_one(i, U, invX)

    return res

def reconstruct(V, U, k, day):

    return trans(np.matmul(U[:, 0:k], V[day, 0:k]))
```

For more information about how these different methods are applied, you can access the Github repository I have created.

# Bibliography


[1] C. S. Bretherton, C. Smith, and John M. Wallace. "An Intercomparison of Methods for Finding Coupled Patterns in Climate Data". In: *Journal of Climate* 5.6 (June 1992), pp. 541–560. DOI: https://doi.org/10.1175/1520-0442(1992)005<0541:AIOMFF>2.0.CO;2.

[2] D. Cooley and E. Thibaud. "Decompositions of dependence for high-dimensional extremes". In: *Biometrika* 106.3 (June 2019), pp. 587–604. ISSN: 0006-3444. DOI: https://doi.org/10.1093/biomet/asz028.

[3] Wolfgang Karl Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Springer Berlin, Heidelberg, 2015. DOI: https://doi.org/10.1007/978-3-662-45171-7.

[4] H. Hotelling. *Analysis of a complex of statistical variables into principal components*. Vol. 24. 6. Journal of Educational Psychology, 1933, pp. 417–441. DOI: https://doi.org/10.1037/h0071325.

[5] H. Hotelling. "Relations Between Two Sets of Variates". In: *Biometrika* 26.3/4 (1936), pp. 321–377. DOI: https://doi.org/10.2307/2333955.

[6] Y. Jiang, D. Cooley, and M.F. Wehner. "Principal Component Analysis for Extremes and Application to U.S. Precipitation". In: *Journal of Climate* 33.15 (2020), pp. 6441–6451. DOI: https://doi.org/10.1175/JCLI-D-19-0413.1.

[7] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007, p. 80. ISBN: 9780131877153. URL: https://books.google.es/books?id=gFWcQgAACAAJ.

[8] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer New York, NY, 2002. ISBN: 978-0-387-95442-4. DOI: https://doi.org/10.1007/b98835.

[9] M. Larsson and S. Resnick. "Extremal dependence measure and extremogram: The regularly varying case". In: *Extremes* 15 (2011), pp. 1–26. DOI: 10.1007/s10687-011-0135-9.

[10] K. Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: https://doi.org/10.1080/14786440109462720.

[11] Sidney I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 1987. ISBN: 978-0-387-75952-4. DOI: https://doi.org/10.1007/978-0-387-75953-1.

[12] Sidney I. Resnick. *Heavy-Tail Phenomena. Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 2007. ISBN: 978-0-387-24272-9. DOI: https://doi.org/10.1007/978-0-387-45024-7.

[13] Michael B. Richman. "Rotation of principal components". In: *Journal of Climatology* 6.3 (1986), pp. 293–335. DOI: https://doi.org/10.1002/joc.3370060305.

[14] S. Sivareddy and et al. "The pre-Argo ocean reanalyses may be seriously affected by the spatial coverage of moored buoys". In: *Sci. Rep.* 7 (2017). DOI: 10.1038/srep46685.

[15] D. Wilks. *Statistical Methods in the Atmospheric Sciences*. Fourth Edition. Elsevier, 2019. ISBN: 978-0-12-815823-4. DOI: https://doi.org/10.1016/B978-0-12-815823-4.09987-9.