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

Miguel Donderis de Vicente NIU: 1491198

June 28, 2022

Contents

CHAPTER 1

Introduction

CHAPTER 2

Methods

2.1 Principal Component Analysis (PCA)

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, analogous to the Principal Axis Theorem in mechanics, and developed by Harold Hotelling three 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.

no me termina de convencer esta introducción

The central idea of Principal Component Analysis 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 *Main Components*, 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 main components is reduced to solving an algebraic problem of calculating eigenvalues and eigenvectors for a positive and symmetric semi-definite matrix.

So as to do the analysis we will focus on the [5] paper.

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 the variances of these p random variables and the structure of the covariances or correlations between them are of interest. Unless p is very small, 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 correlations. Therefore, an alternative approach is necessary,

this introduction does neither convince me at all. I should revise it for sure

and the most optimal way is to look for a few variables that preserve most of the information given by these variances and covariances or correlations, what ara called the Principal Components. Although the analysis for finding these components called *Principal Component Analysis*, which we will refer to as PCA from now on, does not ignore covariances and correlations, it mainly concentrates on variances.

The first step for finding the Principal Components 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 $\boldsymbol{\alpha}_2\mathbf{x}$, uncorrelated with $\boldsymbol{\alpha}_1\mathbf{x}$, having maximum variance, and so on, so that at the kth step it is found a linear combination $\boldsymbol{\alpha}_k\mathbf{x}$ having 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 kth derived variable, $\boldsymbol{\alpha}_k\mathbf{x}$, is called the kth Principal Component, as stated before. It is clear that up to p PCs can be found, as much as the dimension of the original random vector, but it is hoped that most of the variation of $\mathbf{x}$ 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 ith and the jth elements of $\mathbf{x}$ when $i \neq j$ and the variance of the ith element when $i = j$. The more realistic case in which $\boldsymbol{\Sigma}$ is unknown follows by replacing it by a sample covariance matrix $\mathbf{S}$. Then, it turns out that the kth 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 kth largest eigenvalue λ_k. Furthermore, if $\boldsymbol{\alpha}_k$ is chosen to be normalized, then $\text{var}(\boldsymbol{z}_k) = \lambda_k$. The following algebraic derivation of the PCs is the standard one given in many multivariate analysis books, and it may be skipped by those familiarized with this method.

Consider first the expression $\boldsymbol{\alpha}_1^\top \mathbf{x}$, where $\boldsymbol{\alpha}_1$ is chosen to maximize $\text{var}(\boldsymbol{\alpha}_1^\top \mathbf{x}) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1$. We note that the maximum of this expression will not be achieved for a finite $\boldsymbol{\alpha}_1$, so a normalization constraint, $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$ must be imposed. The standard approach to maximize the given expression subject to the normalization constraint is to use the Lagrange multipliers technique. Hence, maximize $\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 - 1)$, where λ is the Lagrange multiplier. Differentiation 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. Thus, we note that λ is an eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ is its corresponding eigenvector. So as to decide to which of the eigenvectors of $\boldsymbol{\Sigma}$ does correspond, we recall that the quantinty to be maximized is

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

so λ must be as large as possible and therefore α_1 is the eigenvector corresponding to the largest eigenvalue of $\mathbf{\Sigma}$ and $\text{var}(\boldsymbol{\alpha}_1^\top \mathbf{x}) = \lambda_1$ the largest eigenvalue.

In general, the kth PC of $\mathbf{x}$ is $\boldsymbol{\alpha}_k^\top \mathbf{x}$ and $\text{var}(\boldsymbol{\alpha}_k^\top \mathbf{x}) = \lambda_k$, where λ_k is the kth largest eigenvalue of $\mathbf{\Sigma}$ and $\boldsymbol{\alpha}_k$ is its corresponding eigenvector. We now will derive this expression for the case $k = 2$, and the proof for $k \geq 3$ is slightly more complicated but similar in the procedure.

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

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

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

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

and, since the first two terms are zero because of the uncorrelation resctriction and $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$, we obtain that $\phi = 0$. Therefore, the equation simplified is $\mathbf{\Sigma} \boldsymbol{\alpha}_2 - \lambda \boldsymbol{\alpha}_2 = (\mathbf{\Sigma} - \lambda \mathbb{I}) \boldsymbol{\alpha}_2 = 0$, so λ is once more an eigenvalue of $\mathbf{\Sigma}$ and $\boldsymbol{\alpha}_2$ is its corresponding eigenvector. Again, the quantity to maximize is $\lambda = \boldsymbol{\alpha}_2^\top \mathbf{\Sigma} \boldsymbol{\alpha}_2$. Assuming that $\mathbf{\Sigma}$ does not have repeated eigenvalues, as if it dit it would follow that $\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_1$ violating the constraint of uncorrelation, we have that λ is the second largest eigenvalue of $\mathbf{\Sigma}$ 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 of coefficients $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, \ldots, \boldsymbol{\alpha}_p$ are the eigenvectors of $\mathbf{\Sigma}$ corresponding to $\lambda_3, \lambda_4, \ldots, \lambda_p$, the third, fourth until the smallest eigenvalue of $\mathbf{\Sigma}$, and furthermore $\text{var}(\boldsymbol{\alpha}_k^\top \mathbf{x}) = \lambda_k$, $k = 1, \ldots, p$.

It should be noted that sometimes the eigenvectors $\boldsymbol{\alpha}_k$ are referred to as 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 kth eigenvector $\boldsymbol{\alpha}_k$, which are also called Empirical Orthogonal Variables, and denote by *Empirical Orthogonal Functions*, or EOFs, the vectors $\boldsymbol{\alpha}_k$, which are also called vector of coefficients or loadings. Further discussion is provided in Appendix **??**.

2.1.2 Optimal properties of Population Principal Components

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

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

where $\mathbf{A}'$ is the orthogonal matrix whose kth column, α_k, is the kth eigenvector of $\mathbf{\Sigma}$. Hence, we note that the PCs are defined by an orthonormal linear transformation of $\mathbf{x}$. Furthermore, we have directly from the derivation in the previous section that

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

where $\mathbf{\Lambda}$ is the diagonal matrix whose kth diagonal element is λ_k, the kth largest eigenvalue of $\mathbf{\Sigma}$, such that $\lambda_k = \text{var}(\boldsymbol{\alpha}_k'\mathbf{x}) = \text{var}(z_k)$. 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 $\mathbf{A}'\mathbf{\Sigma A} = \mathbf{\Lambda}$ and $\mathbf{\Sigma} = \mathbf{A\Lambda A}'$. The orthonormal linear transformation (2.1) of $\mathbf{x}$, which defines the PCs $\mathbf{z}$, has a number of optimal properties which we will discuss.

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

$$\mathbf{y} = \mathbf{B}'\mathbf{x},$$

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

Proof. Let $\boldsymbol{\beta}_k$ be the kth column of $\mathbf{B}$. As the columns of $\mathbf{A}$ form a basis for a $p-$dimensional space, we have that $\boldsymbol{\beta}_k = \sum_{j=1}^{p} c_{jk}\boldsymbol{\alpha}_j$, $k = 1, \ldots, q$, where $c_{jk}, j = 1, \ldots, p, k = 1, \ldots, q$ are appropiately defined constants. Hence, we have that $\mathbf{B} = \mathbf{AC}$, where $\mathbf{C} \in \mathbb{R}^{p \times q}$ with (j,k)th element c_{jk}, and we also observe that

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

where we denote $\mathbf{c}_j'$ the jth row of $\mathbf{C}$ and we have used equation (2.2) and the fact that $\mathbf{\Lambda}$ is a diagonal matrix. Therefore,

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

Now, $\mathbf{C} = \mathbf{A}'\mathbf{B}$, so $\mathbf{C}'\mathbf{C} = \mathbf{B}'\mathbf{A}\mathbf{A}'\mathbf{B} = \mathbf{B}'\mathbf{B} = \mathbb{I}_q$, because $\mathbf{A}$ is orthogonal and the columns of $\mathbf{B}$ are orthonormal. Hence, $\sum_{j=1}^{p}\sum_{k=1}^{q} c_{jk}^2 = q$, and the columns of $\mathbf{C}$ are also orthonormal and it can be thought as the first q columns of a orthogonal matrix $\mathbf{D} \in \mathbb{R}^{p\times p}$. But note that the rows of $\mathbf{D}$ are orthonormal, i.e. $\mathbf{d}_j'\mathbf{d}_j = 1$, $j = 1, \ldots, p$. As the rows of $\mathbf{C}$ consist of the first q elements of the rows of $\mathbf{D}$ it is clear 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, as we have seen $\sum_{k=1}^{q} c_{jk}^2 \leq 1$ is the coefficient of λ_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 $\mathbf{B}' = \mathbf{A}_q'$, then

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

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

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

$$\mathbf{y} = \mathbf{B}'\mathbf{x},$$

where $\mathbf{y} \in \mathbb{R}^q, \mathbf{B}' \in \mathbb{R}^{q\times p}$, and let $\mathbf{\Sigma}_y = \mathbf{B}'\mathbf{\Sigma}\mathbf{B}$ be the variance-covariance matrix for $\mathbf{y}$. Then $\mathrm{tr}(\mathbf{\Sigma}_y)$ 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, 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 statistical implications of Proposition 2.1.2 is that the last few PCs are not just unstructured left-overs after removing the important PCs, those which explain the greatest amount of variance. Because these last PCs have variances as small as possible they are useful in their own right, for example they can be used to detect 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.

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

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

Proof. From equation (2.2) we have that $\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}'$, 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 $\mathbf{A}\mathbf{\Lambda}\mathbf{A}' = \sum_{k=1} \lambda_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k'$, 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 can we 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 $\lambda_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^\top$ from each PC.

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

$$\mathbf{y} = \mathbf{B}'\mathbf{x},$$

where $\mathbf{y} \in \mathbb{R}^q, \mathbf{B}' \in \mathbb{R}^{q \times p}$*, and let* $\mathbf{\Sigma}_y = \mathbf{B}'\mathbf{\Sigma}\mathbf{B}$ *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, $\dim(S_k) = p - k + 1$, where $\dim(S_k)$ denotes the dimension of the vector space S_k. Furthermore, the kth eigenvalue of $\mathbf{\Sigma}$, λ_k, satisfies

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

Suppose that $\mu_1 > \mu_2 > \cdots \mu_q$ are the eigenvalues of $\mathbf{B}'\mathbf{\Sigma}\mathbf{B}$ and that $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_q$ are the corresponding eigenvectors. Let T_k be the the subspace of q-dimensional vectors orthogonal to $\boldsymbol{\gamma}_{k+1}, \ldots, \boldsymbol{\gamma}_q$, with $\dim(T_k) = k$. Then, for any non-zero vector $\boldsymbol{\gamma}$ in T_k it holds that $\boldsymbol{\gamma}'\mathbf{B}'\mathbf{\Sigma}\mathbf{B}\boldsymbol{\gamma} \geq \mu_k \|\boldsymbol{\gamma}\|$. Consider now the subspace $\tilde{S}_k$ of p-dimensional vectors of the form $\mathbf{\Sigma}\boldsymbol{\gamma}, \boldsymbol{\gamma} \in T_k$. Then we have that $\dim(\tilde{S}_k) = \dim(T_k) = k$, because $\mathbf{B}$ preserves lengths of vectors. From the Grassmann formula, 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 $\dim(S_k \cap \tilde{S}_k) \geq 1$. Therefore, there exists 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

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

Hence, the kth eigenvalue of $\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}$ is smaller than the kth 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 $\mathbf{B} = \mathbf{A}_q$. □

The importance of this result follows because 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. For multivariate normal distributed random vector $\mathbf{x}$, the first q PCs are q linear functions of $\mathbf{x}$ whose joint probability distribution has contours of fixed probability enclosing the maximum volume.

Proposition 2.1.5. *Suppose that we wish to predict each random variable x_j in $\mathbf{x}$ by a linear function of $\mathbf{y}$, where $\mathbf{y} = \mathbf{B}'\mathbf{x}$. If σ_j^2 is the residual variance in predicting x_j from $\mathbf{y}$, then $\sum_{j=1}^{p} \sigma_j^2$ is minimized if $\mathbf{B} = \mathbf{A}_q$.*

The statistical implication of this result 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. Althoigh stated as an algebraic property, it can be equally viewed geometrically. In fact, is the population equivalent of Proposition CITE G3. 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 2.1.5 was pointed out by Hotelling in [2], where ho noted that the first PCs derived from a correlation matrix is the linear function of $\mathbf{x}$ that has greater mean square correlation with the elements of $\mathbf{x}$ than any other function.

population equivalent of proposition G3

Proposition 2.1.6. *Consider the family of p-dimensional ellipsoids $\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} = \lambda$, where $\lambda \in \mathbb{N}$ is a constant. The PCs define the principal axes of these ellipsoids.*

Proof. The PCs are defined by the 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. □

This result is important above all if the random vector $\mathbf{x}$ has a multivariate normal distribution, as in this case the ellipsoids define contours of constant probability. The

interpretation of PCs as defining the principal axes of ellipsoids of constant density was also mentioned by Hotelling in [2].

I'm not really sure wether to include this proposition or not

Proposition 2.1.7. *Suppose that* $\mathbf{x}_1, \mathbf{x}_2$ *are two independent random vectors, both having the same probability distribution and both subjected to the same linear transformation* $\mathbf{y}_i = \mathbf{B}^\top \mathbf{x}_i$, $i = 1, 2$. *If* $\mathbf{B} \in \mathbb{R}^{p \times q}$ *is a matrix with orthonormal columns chosen to maximize* $\mathbb{E}[(\mathbf{y}_1 - \mathbf{y}_2)^\top (\mathbf{y}_1 \mathbf{y}_2)]$, *then* $\mathbf{B} = \mathbf{A}_q$.

Proof. First, note that $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Hence, $\mathbf{y}_1$ and $\mathbf{y}_2$ also have the same mean and covariance matrix, $\mathbf{B}^\top \mu$ and $\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B}$ respectively. Then,

$$\mathbb{E}[(\mathbf{y}_1 - \mathbf{y}_2)^\top (\mathbf{y}_1 - \mathbf{y}_2)] = \mathbb{E}[(\mathbf{y}_1 - \mathbf{B}^\top \boldsymbol{\mu})^\top (\mathbf{y}_1 - \mathbf{B}^\top \boldsymbol{\mu})] \mathbb{E}[(\mathbf{y}_2 - \mathbf{B}^\top \boldsymbol{\mu})^\top (\mathbf{y}_2 - \mathbf{B}^\top \boldsymbol{\mu})],$$

where the cross-products vanish because of the independence of the random vectors. Now, for $i = 1, 2$ we have that $\mathbb{E}[(\mathbf{y}_i - \mathbf{B}^\top \boldsymbol{\mu})^\top (\mathbf{y}_i - \mathbf{B}^\top \boldsymbol{\mu})] = \text{tr}(\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B}$, which is maximized when $\mathbf{B} = \mathbf{A}_q$ and as the present criterion has been shown to be $2\text{tr}(\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B}$ we have the desired result. □

This property says that the expected Euclidean distance in a q-dimensional subspace between two vectors of p random variables with the same distribution is made as large as possible if the subspace is defined by the first q PCs. Finally, this property can be of course reversed in the sense that $\mathbb{E}[(\mathbf{y}_1 - \mathbf{y}_2)^\top (\mathbf{y}_1 \mathbf{y}_2)]$ is minimized by taking $\mathbf{B} = \mathbf{A}_q^*$

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. Most of these properties are just the sample analogous of the population case, so they will be mentioned briefly, but in addition there are other properties that are relevant only to sample PCs, so these ones will be discussed more deeply.

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 $\tilde{z}_{i1} = \mathbf{a}_1^\top \mathbf{x}_i$, $i = 1, \ldots, n$ and choose the vector of coefficients $\mathbf{a}_1^\top$ to maximize the sample variance $\frac{1}{n-1} \sum_{i=1}^n (\tilde{z}_{i1} - \bar{z}_1)^2$ subject to the normalization constraint $\|\mathbf{a}_1\| = 1$. Next, let $\tilde{z}_{i2} = \mathbf{a}_2^\top \mathbf{x}_i$, $i = 1, \ldots, n$ and choose $\mathbf{a}_2^\top$ to maximize the sample variance of $\tilde{z}_{i2}$ subject to the normalization constraint $\|\mathbf{a}_2\| = 1$ and so that $\tilde{z}_{i2}$ is uncorrelated with $\tilde{z}_{i1}$. 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, $\tilde{\mathbf{z}}_k = \mathbf{a}_k^\top \mathbf{x} = (\tilde{z}_{ik}, \ldots, \tilde{z}_{in}$ is defined as the kth sample PC, $k = 1, \ldots, p$, and $\tilde{z}_{ik}$ is the score of the ith observation on the kth PC. If the derivation of section 2.1.1 is followed but with sample variances and

covariances replacing the population quantities, it turns out that the sample variance of the PC scores for the kth sample is l_k, the kth largest eigenvalue of the sample covariance $\mathbf{S}$ of $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, and $\mathbf{a}_k$ is the corresponding eigenvector for $k = 1, \ldots, p$.

Define the matrices $\tilde{\mathbf{X}}, \tilde{\mathbf{Z}} \in \mathbb{R}^{n \times p}$ to have (i,k)th elements equal to the value of the kth element $\tilde{x}_{ik}$ of $\mathbf{x}_i$ and $\tilde{z}_{ik}$ of $\tilde{\mathbf{z}}_i$. Then, $\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}$ are related by $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{p \times p}$ is the orthogonal matrix whose kth column is $\mathbf{a}_k$.

As the mean of $\mathbf{x}$ is usually unkown, the $(j,k)th$ element of $\mathbf{S}$ is $\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_{ij} - \bar{x}_j)(\tilde{x}_{ik} - \bar{x}_k)$, where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{ij}$, $j = 1, \ldots, p$. The matrix $\mathbf{S}$ can be therefore written as $\mathbf{S} = \frac{1}{n-1}\mathbf{S}^\top\mathbf{S}$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the matrix whose $(i,j)th$ element is $(\tilde{x}_{ij} - \bar{x}_j)$, the ith observation of the jth variable measured about its mean $\bar{x}_j$. Hence, a convenient notation to define the PC scores is $\mathbf{Z} = \mathbf{X}\mathbf{A}$, rather than the earlier definition. Now, the PC scores will have exactly the same variances and covariances as those given by $\tilde{\mathbf{Z}}$ but will have zero mean.

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 2.1.5 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 kth 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 2.1.5, 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 hold 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.

Proposition 2.1.8. *Suppose that* $\mathbf{X}$ *consists of* n *observations on* p *predictor variables* $\mathbf{x}$ *measured about their sample means, and that the corresponding regression equation is* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, *where* $\mathbf{y}$ *is the vector of* n *observations of the dependent variable, again measured about the sample mean. Suppose that* $\mathbf{X}$ *is transformed by the equation* $\mathbf{Z} = \mathbf{X}\mathbf{B}$, *where* $\mathbf{B} \in \mathbb{R}^{p \times p}$ *is a orthogonal matrix. The regression equation can then be rewritten as* $\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, *where* $\boldsymbol{\gamma} = \mathbf{B}^{-1}\boldsymbol{\beta}$. *The usual least square estimator for* $\boldsymbol{\gamma}$ *is* $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y}$. *Then the elements of* $\hat{\boldsymbol{\gamma}}$ *have successively the smallest possible variances if* $\mathbf{B} = \mathbf{A}$, *the matrix whose* k*th column is the* k*th eigenvector of* $\mathbf{S}$. *Hence,* $\mathbf{Z}$ *consists of values of the sample Principal Components for* $\mathbf{x}$.

unsure to include this proposition and its proof

Proof. The proof is a bit weird and this proposition is also a bit disconnected from the general overview of PCA, so by the moment I will not include it. □

This proposition implies that replacing the predictor variables in a regression anal-

ysis by their first few PCs is a good idea

However, most of the optimal properties of PCs specific to sample situation are geometric, or at least have an important geometric point of view. We now introduce and discuss some of the most relevant ones. 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 2.1.6 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 λ 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.

Regarding Proposition 2.1.7, it may also be carried over from populations to samples. Suppose that the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are transformed by the orthonormal linear transformation $\mathbf{y}_i = \mathbf{B}^\top \mathbf{x}_i$, $i = 1, 2, \ldots, n$, 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. Then,

$$\sum_{h=1}^{n}\sum_{i=1}^{n}(\mathbf{y}_h - \mathbf{y}_i)^\top(\mathbf{y}_h - \mathbf{y}_i)$$

is maximized when $\mathbf{B} = \mathbf{A}_q$. Conversely, the same quantity is minimized when $\mathbf{B} = \mathbf{A}_q^*$. This property, which for the population case seemed irrelevant or at least out of context, for the sample case is very important. This property means that if the n observations are projected onto a q-dimensional subspace, then the sum of squared Euclidean distances between all pair of observations in the subspace is maximized when the subspace is defined by the first q PCs, the PCs that account for the greatest part of the variance, while it is minimized when the subspace is defined by the last q PCs. The proof that this property holds is similar to the population case, so we will not repeat it.

The next property is the sample equivalent to Proposition 2.1.5, and both are concerned with least squares linear regression of each variable x_j on the q variables contained in $\mathbf{y}$.

Proposition 2.1.9. *Suppose that the observations $\mathbf{x}_1 l, \mathbf{x}_2, \ldots, \mathbf{x}_n$ 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. □

2.2 Extreme Principal Component Analysis (xPCA)

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, long-range dependence and domains of attraction. Roughly speaking, *regularly varying functions* are those functions which behave asymptotically like power functions, i.e. functions that are jointly heavy tailed. In this section we will consider only real functions of a real single variable and in the following sections we will considerate the multivariate case for which we will have to introduce concepts such as vague convergence.

Definition 2.2.1. *A measurable function $U : \mathbb{R}^+ \to \mathbb{R}^+$ is regularly varying at ∞ 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 ρ is called the exponent of variation.

A maximal domain of attraction

Suppose that $\{X_n, n \geq 1\}$ are independent identically distributed random variables with common distribution $F(x)$. The extreme is defined by the following expression $M_n = \bigvee_{i=1}^n X_i = \max\{X_1, \ldots, X_n\}$.

Vague convergence

2.2.2 Extension of regular variation

Resnick [6] provides a comprehensive treatment of multivariate regular variation. Roughly speaking, a random vector is said to be *multivariate regular-varying* if its joint tail decays like a power function, in other words, it is jointly heavy-tailed. Although regular variation can be defined on $\mathbb{R}^p$, during all this work we will consider it in the nonnegativa orthant, as done in [1], as this allows to focus our attention on the direction for which we want to assess risk.

Definition 2.2.2. *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 $n\mathbb{P}(b_n^{-1}\mathbf{X} \in \cdot) \xrightarrow{v} \nu_{\mathbb{X}}(\cdot)$ as $n \to \infty$, where $\xrightarrow{v}$ denotes vague convergence in $M_+([0,\infty))$, the space of nonnegative Radon measures on $[0,\infty)^p\backslash\{0\}$.*

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 $\mathbf{X}$. We then write $\mathbf{X} \in \mathrm{RV}^p_+(\alpha)$ so as to denote a regularly varying random vector $\mathbf{X}$ with tail index α. Furthermore, 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 cartesian, coodinates. Hence, given any norm $\|\cdot\|$, the unit sphere is defines as $\mathbb{S}_{p-1}^{+} = \{x \in \mathbb{R}_{+}^{p} : \|x\| = 1\}$. 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, $\nu_{\mathbf{X}}\{C(r,B)\} = r^{-\alpha}H_{\mathbf{X}}(B)$. The scale of $\mathbf{X}$ is related to $\{b_n\}$ and $\nu_{\mathbf{X}}$ or $H_{\mathbf{X}}$. If we now replace the sequence $\{b_n\}$ for $\{kb_n\}$, for some $k > 0$, then for any $r > 0$ and $B \subset \mathbb{S}_{p-1}^{+}$ such that $C(r,B)$ is a continuity set of ν_X, then $n\mathbb{P}\{(kb_n)^{-1}\mathbf{X} \in C(r,B)\} \to \nu_{\mathbf{X}}^{\text{alt}}\{C(r,B)\} = r^{-\alpha}H_{\mathbf{X}}^{\text{alt}}(B)$, where $\nu_{\mathbf{X}}^{\text{alt}} = k^{-\alpha}\nu_{\mathbf{X}}$ and $H_{\mathbf{X}}^{\text{alt}} = k^{-\alpha}H_{\mathbf{X}}$. We ackowledge this relationship between $\{b_n\}, \nu_{\mathbf{X}}$ and $H_{\mathbf{X}}$ by saying that $\mathbf{X}$ has limiting measure $\nu_{\mathbf{X}}$ and limiting angular measure $H_{\mathbf{X}}$ when normalized by the succession $\{b_n\}$.

2.2.3 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 so as to define a particular inner product space on its positive orthant, whose operations preserve regular variation.

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 $\mathbb{X}$, i.e. if $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{X}^p$, then $x_i \in \mathbb{X}$, $\forall x = 1, \ldots, p$. We denote by $t(y)$ the elementwise application of the transform t to the elements of a given vector $y \in \mathbb{R}^p$, in such a way that other functions operating on vectors will similarly be applied elementwise. We define the vector addition of two elements $x_1, x_2 \in \mathbb{X}^p$ by $x_1 \oplus x_2 = t\left(t^{-1}(x_1) + t^{-1}(x_2)\right)$. Given a real number $a \in \mathbb{R}$, we define the scalar multiplication by a vector $x \in \mathbb{X}^p$ by $a \circ x = t\left(at^{-1}(x)\right)$. We finally define the additive identity in $\mathbb{X}^p$ as $0_{\mathbb{X}^p} = t(0)$ and the additive inverse of any $x \in \mathbb{X}^p$ by $-x = t\left(-t^{-1}(x)\right)$. In the appendix section we show that the space $\mathbb{X}^p$ is in fact a vector space.

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

$$a_1 \circ x_1 \oplus a_2 \circ x_2 \oplus \cdots \oplus a_q \circ x_q = t\left(\sum_{j=1}^{q} a_j t^{-1}(x_j)\right).$$

We note that, as $\mathbb{X}^p$ is a $p-$dimensional vector space, any set of p vectors that are linearly independent in $\mathbb{X}^p$, namely vectors such that if $a_1 \circ x_1 \oplus \cdots \oplus a_q x_q = 0$ then $a_j = 0$, $j = 1, \ldots, q$, will form a basis for $\mathbb{X}^p$.

Let now be $A = (\tilde{a}_1, \ldots, \tilde{a}_q) \in \mathbb{R}^{p \times q}$ a matrix of real numbers. Then, for a vector $x \in \mathbb{X}^q$ we define the matrix multiplication by A as $A \circ x = a_1 \circ x_1 \oplus \cdots \oplus a_q \circ x_q = t\left(At^{-1}(x)\right) \in \mathbb{X}^p$. If we denote $\mathbb{I} \in \mathbb{R}^p$ the identity matrix, then $\mathbb{I} \circ x = t\left(It^{-1}(x)\right) = x$,

and we note that if $B \in \mathbb{R}^{p' \times p}$, then $B \circ A \circ x = B \circ t\left(At^{-1}(x)\right) = t\left(BAt^{-1}(x)\right) = BA \circ x \in \mathbb{X}^{p'}$. 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 $a_j \in \mathbb{R}$ and vectors $x_j \in \mathbb{X}^p \; \forall j = 1, \ldots, n$, the linear combination in this case can be reexpressed as

$$a_1 \circ x_1 \oplus \cdots \oplus a_q \circ x_q = Y \circ t(a),$$

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

Given two vectors $x_1, x_2 \in \mathbb{X}^p$, we define the scalar product in $\mathbb{X}^p$ by $\langle x_1, x_2 \rangle = \sum_{i=1}^{p} t^{-1}(x_{1i}) t^{-1}(x_{2i})$. In the Appendix section ? we prove that the conditions for being a well defined scalar product are met. We define the norm of a vector $x \in \mathbb{X}^p$ as $\|x\| = \langle x, x \rangle^{\frac{1}{2}}$ and say that two vectors $x_1, x_2 \in \mathbb{X}^p$ are orthogonal if $\langle x_1, x_2 \rangle = 0$, denoted by $x_1 \perp x_2$. Now, vectors $x_1, x_2 \in \mathbb{X}^p$ and their preimages, $y_1 = t^{-1}(x_1), y_2 = t^{-1}(x_2) \in \mathbb{R}^p$ share the same inner product value. Consequently, $\|x\| = \|y\|_2$, and $x_1 \perp x_2$ if and only if $y_1 \perp y_2 \in \mathbb{R}^p$.

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

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

where $y = t^{-1}(x)$. Hence, we see that 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 [4].

Proposition 2.2.1. *Let $S \in \mathbb{R}^{p \times p}$ be a diagonalizable matrix with real coefficients and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and $u_1, u_2, \ldots, u_p$ be the ordered eigenvalue/eigenvector pairs of the given matrix, and let $e_i = t(u_i)$, $i = 1, \ldots, p$. Then $\max_{x:\|x\|=1} Q(S, x) = \lambda_1$, occurring when $x = e_1$ and $\min_{x:\|x\|=1} Q(S, x) = \lambda_p$, occurring when $x = e_p$. Further,*

the sequence of vectors $x_1, \ldots, x_p$ where each x_k is such that $Q(S, x_k)$ is maximized subject to $x_k \perp x_i$, $i < k$ is given by the pairs $x_i = e_i, Q(S, e_i) = \lambda_i$, $i = 1, \ldots, p$.

In the following section we will apply the ideas of this section to the particular case of the transform $t : \mathbb{R} \to (0, \infty)$ defined by $t(y) = \log(1 + e^y)$. This bijection, known as the *softplus function* is mainly used in neural networks, and as it is continuous and infinitely differentiable it is used as a continuously differentiable approximation of the rectifier function in deep neural networks. The functions maps the elements of $\mathbb{R}$ to elements of its positive orthant $\mathbb{R}_+$ and its inverse is $t^{-1}(x) = \log(e^x - 1)$. Importantly for our purposes, it holds that $\lim_{y\to\infty} t(y)/y = \lim_{x\to\infty} t^{-1}(x)/x = 1$, that is, the transform and its inverse have a negligible effect on large values. For our purposes, we will extend t such that $t(-\infty) = 0$, $t^{-1}(0) = -\infty$ and $t(\infty) = t^{-1}(\infty) = \infty$. Then, we have that $t : \overline{\mathbb{R}}^p \to \overline{\mathbb{X}}^p$, where $\overline{\mathbb{R}}^p = [-\infty, \infty]^p$ and $\overline{\mathbb{X}}^p = [0, \infty]^p$. The additive zero vector in $\overline{\mathbb{X}}^p$ is a vector of p similar components that are equal to $t(0) = \log 2$.

2.2.4 Transformed linear operations on regularly-varying random vectors

We now consider the transformed linear operations defined in the previous section applied to regularly-varying random vectors with the transform $t(y) = \log(1 + e^y)$. For regular variation to be preserved, we need the following assumption on the lower tail of our random vectors.

Remark. Let $\mathbf{X} \in \mathrm{RV}_+^p$. Then

$$n\mathbb{P}\{X_i \leq e^{-kb_n}\} \to 0,\ k > 0, i = 1, \ldots, p, \tag{2.3}$$

as $n \to \infty$.

We note that, as e^{-kb_n} decreases to zero very rapidly, the remark does not seem very restrictive, but it precludes any of the marginals from having nonzero mass at zero. In fact, standard regularly varying distributions such as the Pareto or the Fréchet distributions meet this condition. Hence, we will assume that this holds for the random vectors we might consider in this study.

The following propositions show that regular variation is preserved by the transformed linear operations. In order to proof them we will have to introduce a pair of lemmas, as well as the definition of regular variation on $\mathbb{R}^p$. We recall that we denota a regularly-varying random vector $\mathbf{Y} \in \mathbb{R}^p$ by $\mathbf{Y} \in \mathrm{RV}^p(\alpha)$. Regular variation on $\mathbb{R}^p$ is analogous to that on $\mathbb{R}_+^p$, with the logical changes that convergence to the limiting measure $\nu_{\mathbf{Y}}(\cdot)$ is in the space $M_+\left(\bar{\mathbb{R}}^p \backslash \{0\}\right)$ and the angular mesure H_Y is on $\mathbb{S}_{p-1} = \{y \in \mathbb{R}^p : \|y\| = 1\}$. Henceforth, we denote $y^{(0)} = \max(y, 0)$ applied componentwise to $y \in \mathbb{R}^p$.

Proposition 2.2.2. *Suppose that $\mathbf{X}_1, \mathbf{X}_2 \in \mathrm{RV}_+^p(\alpha)$ are independent random vectors, $n\mathbb{P}(b_n^{-1}\mathbf{X}_1 \in \cdot) \xrightarrow{v} \nu_{\mathbf{X}_1}(\cdot)$ and $n\mathbb{P}(b_n^{-1}\mathbf{X}_2 \in \cdot) \xrightarrow{v} \nu_{\mathbf{X}_2}(\cdot)$. Then, $\mathbf{X}_1 \oplus \mathbf{X}_2 =$*

$t\left(t^{-1}(\mathbf{X}_1)+t^{-1}(\mathbf{X}_2)\right) \in \mathrm{RV}_+^p(\alpha)$ *and* $n\mathbb{P}(b_n^{-1}(\mathbf{X}_1 \oplus \mathbf{X}_2) \in \cdot) \xrightarrow{v} \nu_{\mathbf{X}_1}(\cdot)+\nu_{\mathbf{X}_2}(\cdot)$.

Proposition 2.2.3. *Let* $\mathbf{X} \in \mathrm{RV}_+^p(\alpha)$ *be such that* $n\mathbb{P}(b_n^{-1}\mathbf{X} \in \cdot) \xrightarrow{v} \nu_{\mathbf{X}}(\cdot)$. *Then, for* $a \in \mathbb{R}, n\mathbb{P}(b_n^{-1}(a \circ \mathbf{X}) \in \cdot) \xrightarrow{v} a^\alpha \nu_{\mathbf{X}}(\cdot)$ *if* $a > 0$, *and* $n\mathbb{P}(b_n^{-1}(a \circ \mathbf{X}) \in \cdot) \xrightarrow{v} 0$ *if* $a \leq 0$.

We note that condition 2.3 is necessary because if $\mathbf{X}$ were to have enough mass near 0 and $a < 0$, then $t\left(at^{-1}(\mathbf{X})\right)$ could interfere with the regular variation in the upper tail. Furthermore, we obtain a first outcome from these two propositions, which is a method for constructing a regularly-varying random vector by applying a matrix A to a vector of independent regularly-varying random variables.

Corollary 2.2.0.1. *Let* $A = (\tilde{a}_1, \ldots, \tilde{a}_q) \in \mathbb{R}^{p\times q}$ *be a* $p\times q$ *matrix where* $\max_{i=1,\ldots,p} a_{ij} > 0\ \forall j = 1, \ldots, q$, *and let* $Z = (Z_1, \ldots, Z_q)^T$ *be a vector of independent and identically distributed regularly varying* α *random variables with* $\{b_n\} \in \mathbb{R}$ *a succession such that* $n\mathbb{P}(Z_j > b_n z) \to z^{-\alpha}$, $j = 1, \ldots, q$ *and* $n\mathbb{P}(Z_j \leq e^{-kb_n}) \to 0$, $\forall k > 0, j = 1, \ldots, q$. *Then* $A \circ Z \in \mathrm{RV}_+^p(\alpha)$ *and, when normalized by* $\{b_n\}$, *has angular measure* $H_{A\circ Z}(\cdot) = \sum_{j=1}^q \|a_j^{(0)}\|^\alpha \delta_{a_j^{(0)}/\|a_j^{(0)}\|}(\cdot)$, *where* δ *is the Dirac mass function.*

Introduction of their paper ot xPCA

We propose a method for analyzing extremal behaviour through the lens of a most efficient basis of vectors. The method is analogous to principal component analysis, but is based of methods from extreme value analysis. Specifically, rather than decomposing a covariance or correlation matrix, we obtain the basis of vectors by performing an eigendecomposition of a matrix that describes pairwise dependence matrix. We finally apply the method to contiguous United States.

The foundations of this method is the framework of *multivariate regular variation.*

where does this expression come from?

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

The tail dependence of a pair of random variables is a measure of their comovements in the tails of the distributions. The concept is used in extreme value theory. Random variables that appear to exhibit no correlation can show tail dependence in extreme deviations. For instance, it is a stylized fact of stock returns that they commonly exhibit tail dependence.[1]

2.2.5 Tail pairwise dependence matrix

We assume $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ is a $p-$dimensional regularly varying random vector with index $\alpha = 2$ and angular measure H. The tail pairwise dependence matrix (TPDM), which we denote by $\mathbf{\Sigma}_\mathbf{X}$, is the matrix whose $(i,j)-$th element is

$$\sigma_{ij} = \int_{\mathbb{S}} \omega_i \omega_j dH(\omega), \quad i,j = 1, \ldots, p.$$

Although if focuses on extremal dependence due to its reliance on the angular measure H in its definition, the construction of the TPDM is similar to that of the standard covariance matrix, and therefore it has similar properties. Thus, if X_i has scale b, then the $i-$th diagonal element σ_{ii} is b^2, and $\sigma_{ij} = 0$ if and only if X_i and X_j are asymptotically independent. A fact which is important to PCA is the fact that $\mathbf{\Sigma_x}$ is symmetric and positive definite, and therefore its eigenvectors are real and the eigenvalues are positive. If the marginal distributions are transformed to have a common scale of one, then the TPDM is like the correlation matrix with diagonal entries of one. Another quite important property not shared by the correlation or covariance matrix is that the TPDM is completely positive, namely there exists a nonnegative matrix $B \in \mathbb{R}^{p+q}$, with $q \geq p$ such that $\mathbf{\Sigma_x} = BB^T$. Although in this study is not used, it is also known that completely positive yields a construction method for generating a random vector with a given TPDM (Cooley and Thibaud 2019).

why does this happen?

So as to estimate the TPDM, let $\mathbf{x}_t, t = 1, \ldots, n$, be the transformed observations for all stations on day t. The elements of the matrix are estimated using pairs of $\mathbf{x}_t$'s elements. Define the radial component $r_{t,ij} = \sqrt{x_{t,i}^2 + x_{t,j}^2}$, and let $(\omega_{t,i}, \omega_{t,j}) = \frac{(x_{t,i}, x_{t,j})}{r_{t,ij}}$. We therefore estimate σ_{ij} as

$$\hat{\sigma}_{ij} = \frac{2}{n_{ij,\text{exc}}} \sum_{t=1}^{n} \omega_{t,i}\omega_{t,j}\mathbb{I}(r_{t,ij} > r_{0,ij}), \tag{2.4}$$

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. We note that the indicator function $\mathbb{I}$ 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 $\hat{\mathbf{\Sigma}}_y$ is not guaranteed to be positive definite, so one would have to find the nearest positive definite matrix and perform all the corresponding analysis with it.

2.2.6 PCA decomposition for extremes

Critical to essential PCA is the fact that the eigenvectors of the covariance matrix form an orthonormal basis for the $p-$dimensional reals, and this basis is ordered in importance by the eigenvalues of the covariance matrix, what yields the amount of variance explained by each eigenvector. Critical to our method will be therefore obtaining an ordered orthonormal basis for the $p-$dimensional positive orthant. So as to find this basis, one must first have a vector space. [1] create a vector space for the $p-$dimensional positive orthant by applying the transformation $\mathbf{x} = \tau(y) = \log(1 + e^y)$ componentwise to the vector $\mathbf{y} \in \mathbb{R}^p$. 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. 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.

Moreover, [1] show that applying this transformation to the eigenvectors of the TPDM yields an orthonormal basis for the positive orthant. This basis is ordered by eigenvalues that yield the scale explained by each eigenvector

2.2.7 Extremal dependence measure

This is the case for which σ_{ij} was defined for the bivariate case and referred to as extremal dependence measure.

2.3 Maximum Covariance Analysis (MCA)

2.4 Extreme Maximum Covariance Analysis (xMCA)

CHAPTER 3

Applications to climate data

3.1 Data

Once we have introduced and explained the theoretical mathematical framework of our work, which has consisted in explaining the derivation of the Principal Components and some optimal algebraic and geometric properties, explaining the Extreme Principal Component Analysis technique developed in [1] and finally introducing the Maximum Covariance Analysis, both for standard and extreme cases as well, is time to explain the applications we have studied. In this section we will review the different type of data we have used for our study and the preprocessing we have applied, depending on the goal of the study.

3.1.1 Precipitation

Inspired by the study Daniel Cooley and Yujing Jiang carried out in [3], the first goal of our study was to reproduce the Extreme Principal Component Analysis technique which we have explained in section ?. In this study 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)-daily dataset. Following the same procedure of the authors, we limit our investigation to the months of August, September and October, as they correspond to the height of hurricane activity in the Atlantic Ocean, although it is important to remark that we study all kind of extreme precipitation events during these three months regardless of whether these events could have been associated with a hurricane event. This dataset contains data from 1140 stations and 6164 days in the analyzed dataset.

However, our work did not simply reduced to recreating the results in [3] using the

mentioned data, but we decided to extend the analysis to different regions, in particular from the Iberian Peninsula. With this purpose we used the E-OBS dataset, which is a daily gridded land-only observational dataset over Europe whose station data are sourced directly from the European National Meteorological and Hydrological Services (NMHSs) or other data holding institutions 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 in a square meter. This dataset also contains data from wind speed, temperature, humidity, etc., which could be interesting for other studies. In this case, we limit our investigation to the months of July, August, September, October and November, as they correspond to the months of maximum precipitation in the east coast of the Iberian Peninsula and the height of Cold Drop phenomena, when most extreme events over the iberian peninsula take place. The dataset consists of a daily time series of 25567 timepoints and a resolution of 464 and 201 longitude and latitude gridpoints, respectively.

3.1.2 Sea surface salinity

With the aim of extending the work of Daniel Cooley and Yujing Jiang as mentioned, we have introduced the mathematical framework of Maximum Covariance Analysis and Extreme Maximum Covariance Analysis, as an analogous to Extreme Principal Component Analysis, with the goal of finding patterns in different time series which explain a maximum fraction of the covariance between two variables. Hence, one of the consired variables will be precipitation over the Iberian Peninsula, data which we have explained previously, while the second one will be sea surface salinity. One could be tempted to use sea surface temperature, for example, as the relation with extreme precipitation events is clear, but in our case we focus on sea surface salinity as it is a parameter which has not been studied that deeply. 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 from the years 1993-2019, 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.

3.2 Precipitation extremes

3.2.1 Preprocessing

The regular variation framework described requires that each of the random variables is heavy-tailed with a common tail index α. However, it is unsual for the 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. As Cooley and Jiang do in [3], we choose to analyze the 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, then we define

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

to be the moving average data, where the superscript simply denotes that $x_{t,i}^{\text{orig}}$ is in the original scale before further transformations of the marginal distributions. One may note that selecting a 3-day moving average avoids some of the problems of a single extreme precipitation event being partially recorded over two separate days and it could aso help other problems such that a extreme event is recorded in different close station superscript simply denotes that $x_{t,i}^{\text{orig}}$ is in the original scale before further transformations of the marginal distributions. One may note that selecting a 3-day moving average avoids some of the problems of a single extreme precipitation event being partially recorded over two separate days and it could aso help other problems such that a extreme event is recorded in different close stations. It is important to remark that taking a 3-day moving average does not include dependence in the $x_{t,i}^{\text{orig}}$, an important fact for the further analysis. Furthermore, we take the 3-day moving average as it was the original length of extreme events Cooley and Thibaud wished to explore so we replicate it, 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 index $\alpha = 2$. Therefore, to ensure that this is true for all data, we perform the transformation

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

should I explain in the appendix how the smoothing of the distribution works?

where $G(x) = e^{-x^{-2}}$ is the cumulative distribution function of a Frechét random variable with scale 1 and $\alpha = 2$ and $\hat{F}_i$ is the estimated marginal cumulative distribution function of the data from station i. Therefore, this transforms the moving averaged data value $x_{t,i}^{\text{orig}}$ to a variable with Frechét distribution $x_{t,i}$ having the same cumulative probability as that associated with $x_{t,i}^{\text{orig}}$ within its cumulative distribution function. The simplest method for obtaining $\hat{F}_i$ is to use a simple rank transform. However, due to the dependence induced by the 3-day moving average, this kind of trasformation would ignore this dependence, so as performed in [3] we take the average of three linearly interpolated cdf estimates from three subsequences of $x_{t,i}^{\text{orig}}$, which has been shown in [3] to better retain the clustering generated by the moving average. Finally, the transformed data $\mathbf{x}_t$, $t = 1, \ldots, 6164$ number of stations, are treated as independent

and identically distributed.

3.3 Identifying covarying extremes in precipitation and sea surface salinity

3.3.1 Preprocessing

CHAPTER 4

Results

4.1 Comparison of PCA and xPCA for precipitation

Once the PCA and xPCA methods have been introduced, as well as the different type of preprocessing of the data, in this section we are going to make a comparison of both methods, which is the essential objective of this thesis. With this purpose, we will focus on three regions that we have already mentioned. In the first place, the Iberian Peninsula, then we will focus on a particular region, the Valencian Community, Spain. Furthermore, we will limit our investigation to the months from July to November, as it corresponds to the Cold Drop season in the east coast of Spain that is when most extreme precipitation events take place. The third region will be finally the United States, taking advantage of the fact that we have the results of [3], so we will try to recreate them and compare with those obtained through the PCA standar method.

We remark that the eigenvectors $\mathbf{u}_i$, $i = 1, \ldots, p$ obtained through the standard decomposition of eigenvectors and eigenvalues of the estimate of the TPDM, $\hat{\mathbf{\Sigma}}_x$ are transformed to $\mathbf{e}_i = t(\mathbf{u}_i)$, which form and ordered orthonormal basis of the positive orthant $\mathbb{R}^p_+$, while no transformation is needed for the eigenvectors obtained using the standad PCA. Throughout this section we will focus on the first six eigenvectors of each basis, $\mathbf{e}_i$, $i = 1, \ldots, 6$. As in standard PCA, due the orthogonality of the eigenvectors their interpretation, that is the direction of maximum variation of data after accounting for the information contained in the previous eigenvectors, is harder as i increases so choosing a low number of eigenvectors which account for a great part of the variance is a right decision.

In the comparison, we will look to each of the first six eigenvectors, both for PCA and xPCA. To facilitate comparison we include the first PCA mode, or eigenvector, with the corresponding first xPCA mode, the second PCA mode with the corresponding second

xPCA mode, and so on until the sixth mode. Finally, we include a reconstruction of the precipitation data for both cases.

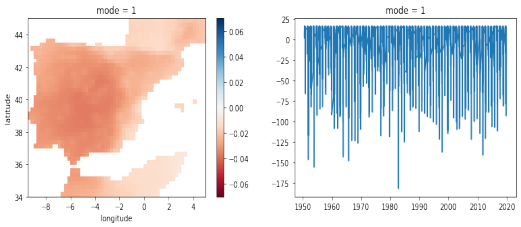
4.1.1 Iberian Peninsula

The first region of interest is the Iberian Peninsula, which includes both countries of Portugal and Spain. As stated, we will concentrate our attention on detecting extreme events in the mediterranean coast of Spain. The first basis vectors for both cases are completely negative, which as PCs are defined up to a sign, in fact we can consider they're completeyly positive. This is due to the fact that both the covariance matrix **S** and the TPDM are completely positive. Another noticeable feature is that there is little variation amonth the values for the different grid points. As we see from both time series, the first coefficient is quite homogeneous with a value around 100 excluding some deviations. Taken by itself, this tells us that precipitation could be quite homoegenous in all the Iberian Peninsula. However, we know this is not true, and in fact when combined with further 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.

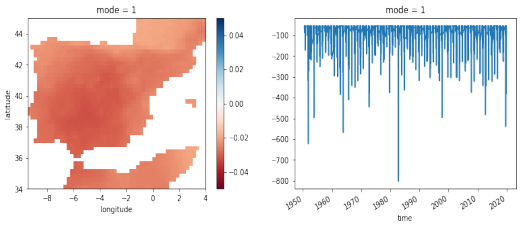
Regarding the second and third modes, we appreciate that now a dipole structure arises, which is natural in PCA decomposition. This dipole structure seems to be stronger in the xPCA decomposition, as the gradient of values is higher. The second mode of PCA shows large negative values for the southeast 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 mode shows the opposite trend, with high low negative values on the north east coast of Spain and low positive values of the south west region of the Iberian Peninsula. This last mode seems less interesting for our purposes, as it gives the same values to the Cantabrian coast and the Catalan coast, and we know that extreme events are normally not happening at the same time in both places. Regarding xPCA, the mode shows homogenous large negative values in the whole east coast, and negative positive values for the western half of the Iberian Peninsula. As for the third mode, we can see that it shows very high positive values in the northern half of the peninsula and very high negative values in the southern half of the peninsula. Therefore, if combined with negative coefficients both second modes will allocate normal and extreme precipitation in the east coast.

Fourth, fifth and sixth mode show a tripole structure, again what is expected in PCA decomposition. We again observed that the gradient of values in xPCA decomposition is higher that in standard PCA decomposition, what makes sense as the TPDM decomposition summarizes better extreme behaviour. Furthermore, we note that these three modes have a quite similar structure, contrary to what we have seen in the second and third mode. The fourth mode shows a transition from positive values in the north west coast, negative values in the east coast and again positive in the Algerian coast, the fifth mode show the opposite behaviour, going from the south west coast to the north

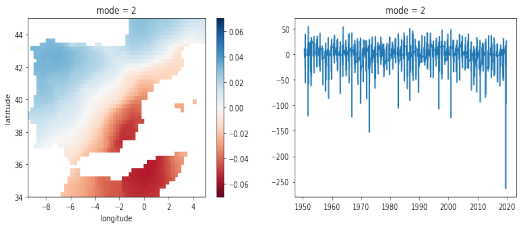
east coast, and finally the sixth mode shows a gradient between the west, center and the east coast. This last mode is the most interesting, as it shows a homogeneous behaviour along the whole mediterranean coast, although it seems to be correlated to the west coast. However, as the number of PCs increases, the global interpreation becomes more difficult, so it is a better idea on focusing on the trends.



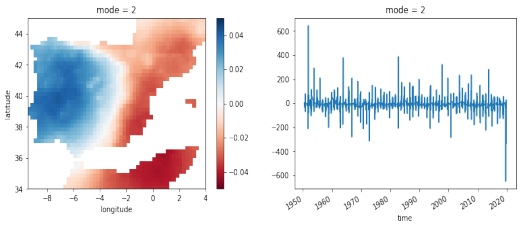
(a) First PCA mode



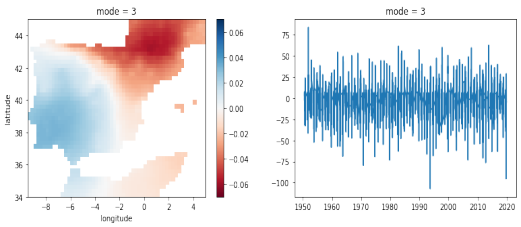
(b) First xPCA mode



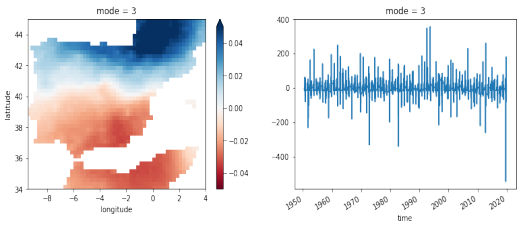
(a) Second PCA mode



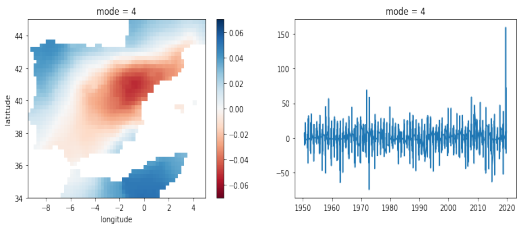
(b) Second xPCA mode



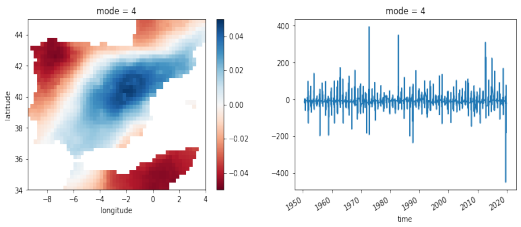
(a) Third PCA mode



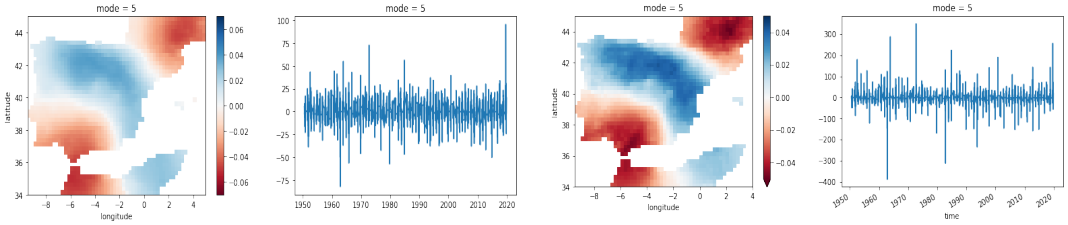
(b) Third xPCA mode



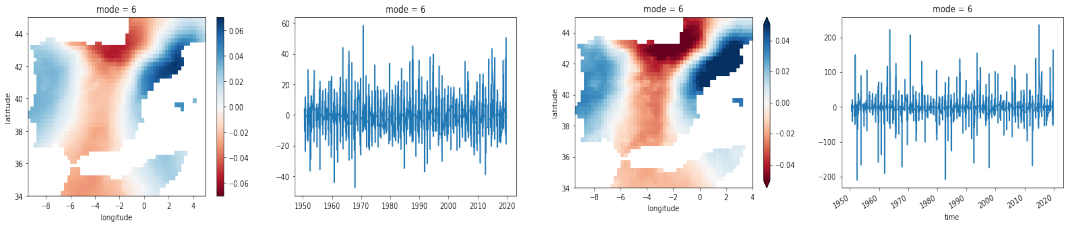
(a) Fourth PCA mode



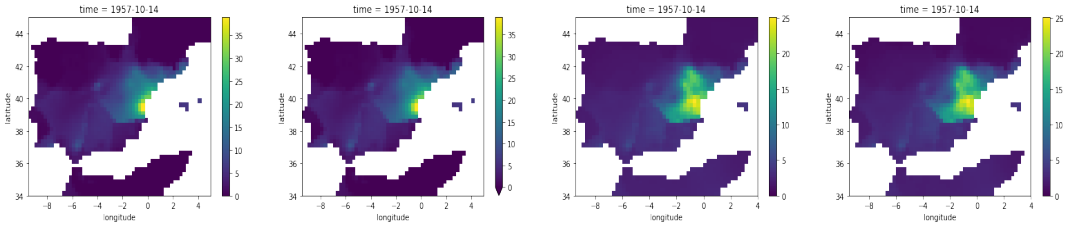
(b) Fourth xPCA mode



(a) Fifth PCA mode (b) Fifth xPCA mode



(a) Sixth PCA mode (b) Sixth xPCA mode



(a) Reconstruction of precipitation using xPCA (b) Reconstruction of precipitation using xPCA

4.1.2 Valencian Country

Now that we have investigated the behaviour of the PCs in the whole Iberian Peninsula both for the standard and the extreme PCA methods, we will make a zoom and focus on a small region, the Valencian Country. Recall that we decided to investigate extreme events during the Cold Drop season of Spain and those happen mainly in the eastern third of the Iberian Peninsula, so here is the motivation to choose this region. The interpretation of the PCs will be very similar to the one made in the previous case.

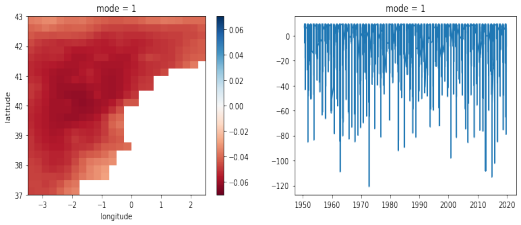
The first eigenvector for both cases again shows a completely negative homogeneous state along all the region for both cases, which is equivalent to a complete positive behaviour, due to the complete positiviness of both standard and extreme covariance matrices. Again, if considered alone the result of this mode 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 in these regions.

Regarding the second and third mode, we see that they're surprisingly very similar in both cases. The dipole structure that we mentioned in the analysis of the PCs of the Iberian Peninsula again arises in this case. We see that the second mode distinguishes

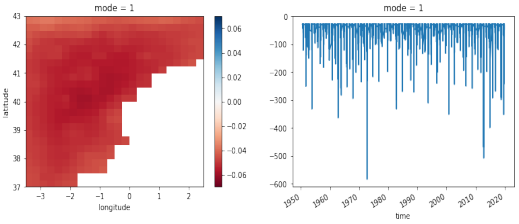
between the north and south regions while the third mode focuses on the interior of the Valencian Country and the mediterranean coast. Furthemore, something remarkable is that the different of gradient of values that we appreciated in the Iberian Peninsula now has disappeared and they're very similar. We note that third mode will therefore have a higher importance in our case, as it clearly differentiates a homogeneous behaviour on the east coast.

So as to the fourth, fifth and sixth mode, we see the tripole structure we also talked about in the previous section. If we pay attention to the fourth and fifth mode, we can again see that they're similar, except for a sign in the fifth mode, and that the gradient of values is the same. However, the sixth mode seems quite different. In the case of the PCA, it does not seem to be very important as there is no homogeneous behaviour in the east coast, what can be seen in the sixth xPCA mode.

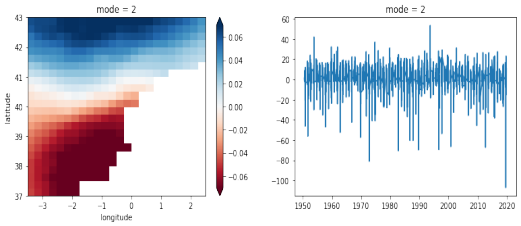
We finally include a reconstruction of precipitation events during the day 14/10/1957, where an important flood took place in Valencia. One can note that the reconstructions are different, but this is again due to the fact that the preprocessing of the data is different, so the PCA reconstructs the image for the moving averaged data and the xPCA for the regularly varying transformed data.



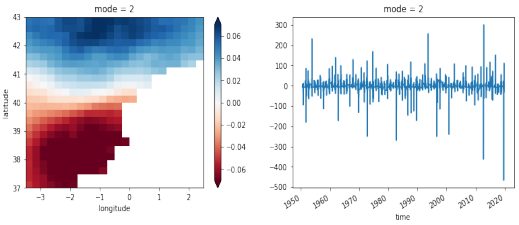
(a) First PCA mode



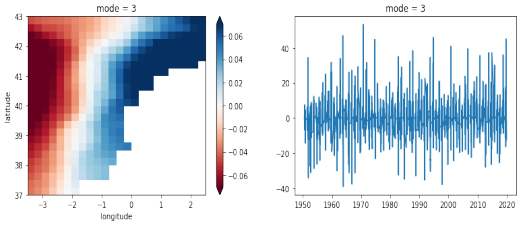
(b) First xPCA mode



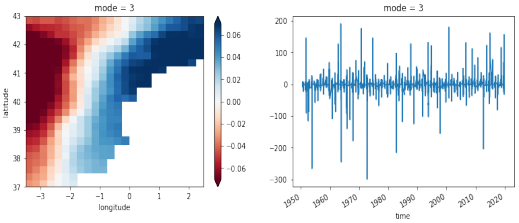
(a) Second PCA mode



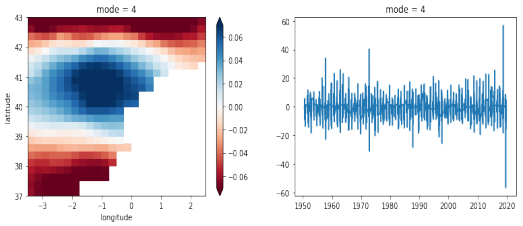
(b) Second xPCA mode



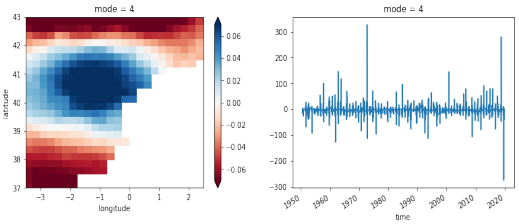
(a) Third PCA mode



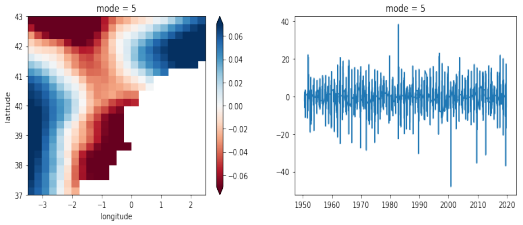
(b) Third xPCA mode



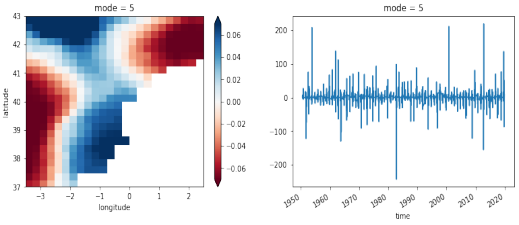
(a) Fourth PCA mode



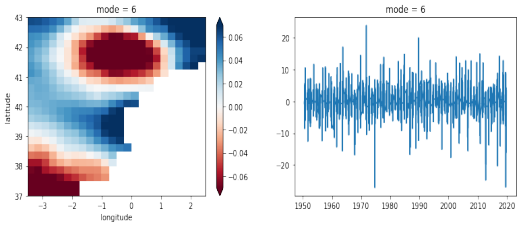
(b) Fourth xPCA mode



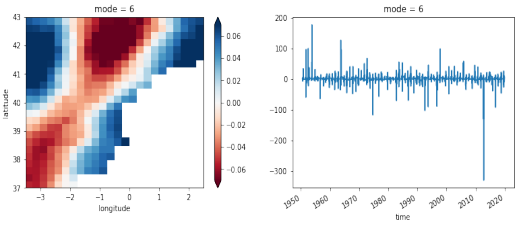
(a) Fifth PCA mode



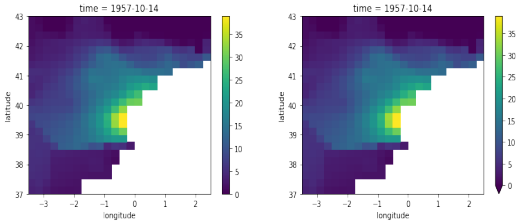
(b) Fifth xPCA mode



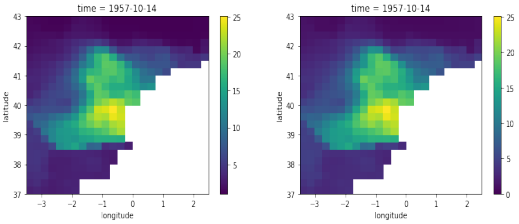
(a) Sixth PCA mode



(b) Sixth xPCA mode



(a) Reconstruction of precipitation using PCA



(b) Reconstruction of precipitation using xPCA

4.1.3 North America

4.2 Comparison of MCA and xMCA for precipitation and sea surface salinity

4.2.1 Iberian Peninsula and North Atlantic ocean

CHAPTER 5

Discussion

CHAPTER 6

Outlook

APPENDIX A

$\mathbb{X}^p$ is a vector space

Here I will explain that the space that Cooley and Thibaud describe is a vector space.

APPENDIX B

Code

Bibliography

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.