Ludwig-Maximilians-Universität München

Institut für Statistik

Functional Data

Functional Principal Component Analysis

Philipp Lintl

supervised by

Dr. Clara Happ

October 11, 2018

Contents

1 Introduction

The progress of technology and thus rise of data leads to scenarios, in which structures need to be revealed and complexity decreased. That is, when the so called principal component analysis is drawn on. An approach to motivate it, is to view it as a tool to achieve dimensionality reduction. This is especially desirable as with the examination of high dimensional data, problems arise. Multivariate data sets often contain a vast number of variables and thus lead to inconclusive or hardly interpretable Graphics. Secondly, further multivariate methods applicable to the statistical analysis tend to work less smoothly. This phenomenon is often referred to as the *curse of dimensionality*. For functional data, especially, it also serves as an explorational device. In this case, the data is being recorded continously over a period of time or discrete time points. Generally speaking, Functional data analysis provides statistical methodology suited for data that is in the form of functions. The high intrinsic dimensionality of these data poses challenges both for theory and computation. On the other hand, the high or infinite dimensional structure of the data is a rich source of information, which brings many opportunities. The functional principal component analysis identifies modes of variation in the data, and how many of them seem to be substantial and thus provides a first overview of the data and the underlying structures.[Ramsay and Silverman, 2005, p. 147; Jolliffe, 2002, p. 1; Everitt and Hothorn, 2011, p. 61]

The possibilities coming along with its application reach many further topics. It is essential to functional data analysis tasks, such as functional regression, functional classification, functional clustering or outlier detection. The procedure is firstly introduced for the multivariate data case and then adjusted for functional data.

2 Principal component analysis

In order to reduce the dimensionality while preserving the maximum amount of variation and thus information originally in the data set, multivariate *principal component analysis* is applied. It can be described as a statistical procedure that transforms a set of observations of possibly correlated variables into a new set of values

of linearly independent variables called **principal components**. These are ordered linear combinations of the original variables, so that the first few account for most of the variation in all the original variables. The best result would be to get a small number of new variables that replace an originally large number of variables, while providing a simpler basis for graphing or further multivariate methods.

2.1 For multivariate data

Let the obeserved data matrix be $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T, \quad \mathbf{X} \in \mathbb{R}^{Nxp}$ with N being the number of observations and p the number of variables.

Basic to the mathematical considerations are **centred** data, meaning $\tilde{x}_{ij} = x_{ij} - \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_{ij}$ is done as preparation for PCA. This way, maximizing the mean square corresponds to maximizing their variance, which is written as:

$$\widehat{Var}(\tilde{x}_j) = \frac{1}{N-1}\sum_{i=1}^{N} \tilde{x}_{ij}^2 \quad \text{and covariance} \quad \widehat{Cov}(\tilde{x}_j, \tilde{x}_k) = \frac{1}{N-1}\sum_{i=1}^{N} \tilde{x}_{ij}\tilde{x}_{ik} \quad (2.1)$$

Linear combinations are essential not only to other multivariate statistics, such as regression, but also drawn on in PCA.

The basic idea is to describe variation within a set of (correlated) variables $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)^T$ in terms of a new set of uncorrelated (orthogonal) variables $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_q)^T$. Every single one is a linear combination of the original Variables $\mathbf{x}$ and the weights $\boldsymbol{\phi}$ with the form as such:

$$\xi_1 = \phi_{11}x_1 + \phi_{12}x_2 + \ldots + \phi_{1p}x_p \quad (2.2)$$

$$\xi_i = \sum_{j=1}^{p} \phi_{ij}x_j = \boldsymbol{\phi}^T x_i, \quad i = 1, \ldots, N \quad (2.3)$$

$$\boldsymbol{\xi} = \boldsymbol{\phi}^T \boldsymbol{x} \quad (2.4)$$

The stepwise procedure tries to find **normalized weights** $\boldsymbol{\phi}$ that maximize variation in the new uncorrelated variables ξ_i.

1. First principal component: The weight vector $\boldsymbol{\phi}_1 = (\phi_{11}, \ldots, \phi_{p1})^T$ for which

the linear combination values

$$\xi_{i1} = \sum_j \phi_{j1} x_{ij} = \boldsymbol{\phi_1^T} \boldsymbol{x}_i, \quad i = 1, \ldots, N, j = 1, \ldots, p \tag{2.5}$$

are **maximized** in their variance, meaning $Var(\boldsymbol{\xi}_1) = \frac{1}{N-1} \sum_i {\xi_{i1}}^2 \rightarrow max$. As mentioned earlier this accounts for the variance, as the data is centred before applying PCA. $\boldsymbol{\phi_1}$ denotes the first principal component and as a result explains the largest mode of variability within the data. Randomly increasing the coefficients $(\phi_{11}, \phi_{12}, \ldots, \phi_{p1})$ can enlarge the variance of ξ_i. Therefore well-definity is required, which is fulfilled with placing a restriction on $\boldsymbol{\phi}_1$:

$$\|\boldsymbol{\phi_1}\|^2 = \langle \boldsymbol{\phi}_1, \boldsymbol{\phi}_1 \rangle = \sum_j \phi_{j1}^2 = 1 \tag{2.6}$$

However, the restriction does not uniquely determine the weights, as changes in signs of any vector $\boldsymbol{\phi}_m$ do not change the value of the variance it defines.

2. Further principal components: The steps described above are reiterated up to a maximum of the number of observed variables p. After m-1 steps and so many principal components found, the m-th principal component is found by computing a new weight $\boldsymbol{\phi}_m$ with components ϕ_{jm}. The corresponding so called *principal component scores* $\xi_{im} = \boldsymbol{\phi}_m^T \boldsymbol{x}_i$ are again **maximized** in their variance $V(\xi_{im})$. Additional to the first restriction $\|\boldsymbol{\phi_m}\|^2 = 1$, there are (m-1) additional constraints

$$\langle \boldsymbol{\phi}_m, \boldsymbol{\phi}_k \rangle = \sum_j \phi_{jk} \phi_{jm} = \boldsymbol{\phi}_k^T \boldsymbol{\phi}_m = 0, \quad k < m. \tag{2.7}$$

These are essential to finding $\boldsymbol{\phi}_m$, as they guarantee **orthogonality** between all identified principal components. This feature makes sure, that every identified pc explains something new. As every pc accounts for as much variation as possible at each step, the amount of variation measured declines. Often few principal components already explain a majority of variation and thus lead to less informative remaining principal components. That is why, usually a small number of pcs are chosen and analyzed. How to determine the right number

of pcs will be discussed later on. [Ramsay and Silverman, 2005, p.148/149]

Solving the optimizationproblem

In multivariate statistics, solving the optimizationproblem with respect to the constraints is defined as finding the eigenvalues and corresponding eigenvectors of the covariance or correlation matrix of the observations $\mathbf{X} \in \mathbb{R}^{Nxp}$. Let $\mathbf{V}$ be the pxp matrix representing the variance-covariance matrix $\mathbf{V} = (N-1)^{-1}\mathbf{X}^T\mathbf{X}$. The criterion for maximizing the variance of $\boldsymbol{\xi_i}$ and thus finding $\boldsymbol{\phi_i}$ can be represented as:

$$max\ V(\boldsymbol{\xi}) = max\ \frac{1}{N-1}\sum(\boldsymbol{\phi}^T\boldsymbol{x})^2 = max\ \frac{1}{N-1}\boldsymbol{\phi}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\phi} = max\ \boldsymbol{\phi}^T\boldsymbol{V}\boldsymbol{\phi} \quad (2.8)$$

As the means are subject to estimation, it can also be divided by N instead of N-1, though it does not effect the pc analysis greatly. The norming constraint $\|\boldsymbol{\phi}\|^2 = 1$ still holds.

The optimization problem turns into an eigenvalue problem. [Ramsay and Silverman, 2005, A.5] In particular, it is the task of finding the largest eigenvalue λ, which solves the *eigenvector problem*:

$$\boldsymbol{V}\boldsymbol{\phi} = \lambda\boldsymbol{\phi} \quad (2.9)$$

In this equation, principal components $\boldsymbol{\phi}$ now act as *eigenvectors*. The solution yields eigenvalue-eigenvector pairs (λ_j, ϕ_j), whereas all eigenvectors $\boldsymbol{\phi_j}$ are orthogonal. Before applying PCA, the mean of each column of $\mathbf{X}$ is usually subtracted of every respective observation. So, one row can be represented as the linearcombination of all other row vectors resulting in the rank of $\mathbf{X}$ not exceeding N-1. Thus, the pxp Matrix $\mathbf{V}$ has at most $min\{p, N-1\}$ nonzero *eigenvalues* λ_j. As mentioned earlier, the amount of variation explained by a found pc decreases in every step. So, in step m, the m-th largest eigenvalue λ_m represents the m-th largest explained variance: $V(\boldsymbol{\xi_m}) = \lambda_m$. [Ramsay and Silverman, 2005, p. 152/153]

2.2 For functional data

In the functional data case, a random sample of independent real valued functions $x_1(t), \ldots, x_N(t)$ on an Interval $\mathcal{T} = [0, T]$ is looked at. Individually they can be viewed as realizations of a one-dimensional stochastic process $X = X(t)$ with mean function $\mu(t) = \mathbb{E}(X(t))$ and the emprical covariance function $\hat{\gamma}(s,t) = Cov(x(s), x(t))$. Those functions live in the $L^2(\mathcal{T})$ space, which is the case, if a stochastic process $X(t)$ satisfies $\mathbb{E}(\int_I X^2(t)dt) < \infty$. [Wang et al., 2016, p. 2; Cuevas, 2014, p. 5]

For functional data, the principal component analysis works as described in the previous section. Variable values x_{ij} along with the discrete index j are replaced by function values $x_i(t)$ and the respective continuous index t. In the multivariate case, the inner product $\langle \boldsymbol{\phi}, \mathbf{x} \rangle = \boldsymbol{\phi}^T \mathbf{x} = \sum_j \phi_j x_j$ was used in order to combine the weight vector $\boldsymbol{\phi}$ with a data vector $\mathbf{X}$. Whereas in the functional case, there are **weight functions** $\phi(s)$ and **function values** $x(t)$ due to the inner product for L_2, defined as such: $\langle \phi, x \rangle = \int_{\mathcal{T}} \phi(t)x(t)$. The summations over the discrete index j are substituted by integrations over the continuous index t. In the context of principal components, this leads to **principal component functions** $\phi_j(t)$ and principal component scores

$$\xi_i = \langle \phi, x_i \rangle = \int \phi(t) x_i(t) dt. \tag{2.10}$$

1. Equivalently to the multivariate case, the first functional principal component $\phi_1(t)$ is found by **maximizing** the variance

$$Var(\xi_1) = \frac{1}{N-1} \sum_{i=1}^{N} \xi_{i1}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\int \phi_1 x_i \right)^2. \tag{2.11}$$

   Well-definition is again acquired by restricting the principal component function:

$$\|\phi_1\|^2 = \langle \phi_1, \phi_1 \rangle = \int \phi_1^2 = \int \phi_1(t)^2 dt = 1. \tag{2.12}$$

2. Further principal components ϕ_m, similar to the multivariate case, have to

fulfill orthogonality constraint(s)

$$\langle \phi_k, \phi_m \rangle = \int \phi_k(t)\phi_m(t)dt = 0, \quad k < m. \tag{2.13}$$

[Ramsay and Silverman, 2005, p. 149/150]

In terms of dimensionality reduction, FPCA can also be defined with the help of a basis expansion. We can assume, that each function x_i has a basis expansion, which underlies the so called *Karhunen-Loeve expansion*. It's theoretical properties can be found in 4.1.1.

This is applied to centred functional data, leading to the following Karhunen-Loeve representation of each observed function.

$$x_i(t) = \sum_{j=1}^{\infty} \xi_{ij}\phi_j(t) \tag{2.14}$$

A reduction in dimensionality is accomplished, when the first k terms approximate the infinite sum well. The information included within x_i thus can be represented in the k-dimensional vector $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{ik})$. In the following, the approximated process

$$x_i(t) \approx \sum_{j=1}^{k} \xi_{ij}\phi_j(t) \tag{2.15}$$

is worked with. The funtional principal components express this underlying stochastic process as a sequence of uncorrelated variables, which are then reduced to a finite vector. This vector of random scores then becomes subject to multivariate data analysis tools allowing the conversion from infinite dimensional functional data to a finite dimensional vector of random scores. [Wang et al., 2016, p. 8/9; Ramsay and Silverman, 2005, p. 151/152]

So to conclude, Functional PCA, as well as multivariate PCA, if similarly notated, underly the same formal steps. Generally speaking, it is the search for a set of mutually orthogonal and normalized weight functions ϕ_m, which solve the eigenvalueproblem of the covariance operator Γ, which is the equivalent to the covariance matrix in the multivariate case for functional data.

Computational methods for actually finding these principal component functions

ϕ_m, will now be introduced. In order to deal with the computational difficulties of the integrations in (2.10 -2.13) two approaches are considered.

2.2.1 Solving the optimizationproblem by discretization

Again, we assume the data to be centred before applying PCA. This leads to a definition of the covariance function $\hat{\gamma}(s,t)$ as follows:

$$\hat{\gamma}(s,t) = \frac{1}{N-1}\sum_{i=1}^{N} x_i(s)x_i(t). \tag{2.16}$$

This amounts to principal component functions $\phi_j(s)$, which fulfil the equation:

$$\int \hat{\gamma}(s,t)\phi(t)dt = \lambda\phi(s) \tag{2.17}$$

The left side in (2.17) represents an *integral transform* Γ of ϕ defined as such:

$$\mathbf{\Gamma}\phi = \int \hat{\gamma}(\cdot,t)\phi(t)dt \tag{2.18}$$

The so called *covariance operator* $\mathbf{\Gamma}$ allows to describe the eigenequation as:

$$\mathbf{\Gamma}\phi = \lambda\phi \tag{2.19}$$

Meaning, that ϕ turned from eigenvector to **eigenfunction**. As this notation resembles the multivariate eigenequation (2.9), it also holds $V(\xi_m) = \lambda_m$. Though, (2.9) seems equivalent to (2.19), there is a difference in the maximum number of different eigenvalue-eigenfunction pairs. In particular concerning the maximum number of eigenfunctions. Which, in this case corresponds to the number of function values $x_i(t)$, thus infinity. [Ramsay and Silverman, 2005, p. 153/154]

Using a **discrete grid** $t_1, \ldots, t_l$ for N functions $x_i(t)$ deals with this problem. It reduces the maximum number of eigenfunctions ϕ to $\min(N, l)$.

As mentioned, the observed functions x_i spanning the Interval $\mathcal{T}$ are discretizised to a fine grid of l equally spaced values s_j. The resulting data matrix $\mathbf{X} \in \mathbb{R}^{Nxl}$ is then suitable for multivariate PCA.

This procedure yields l eigenvalues and l eigenvectors $\mathbf{u}$ in respect to the new co-

variance matrix Σ.

$$\mathbf{\Sigma u} = \lambda \mathbf{u} \tag{2.20}$$

If l is much larger than N, it is more handy to work with the Singular value Decomposition $\mathbf{UDW}^T$ of $\mathbf{X}$ than the lxl matrix $\mathbf{\Sigma}$ in order to solve (2.20). The singular values of $\mathbf{D}$ are denoted by its diagonal elements $d_1, \ldots, d_q$, which fulfil $d_1 \geq \ldots \geq d_q \geq 0$. Here, q represents the maximum number of linearly independent rows or columns of $\mathbf{X}$, also referred to as rank($\mathbf{X}$). In case of a square and symmetric matrix, which is satisfied by $\mathbf{\Sigma}$, the diagonal elements d_i include all nonnegative eigenvalues of $\mathbf{X}$.

This effects the covariance matrix $\mathbf{\Sigma}$ as follows:

$$N\mathbf{\Sigma} = \mathbf{X}^T\mathbf{X} = (\mathbf{WD}^T\mathbf{U}^T)(\mathbf{UDW}^T) = \mathbf{WD}^2\mathbf{W}^T \tag{2.21}$$

Due to (2.21), the eigenvalues of $\mathbf{\Sigma}$ are given by the squares of the singular values of $\mathbf{X}$, $diag(\mathbf{D}^2)$. Corresponding Eigenvectors can be found in the columns of $\mathbf{W}$.

In order to transform these vectors back to functions, a vector $\tilde{\boldsymbol{\phi}}$ of l values $\phi(s_j)$ is defined. Then, for each s_j the following holds approximately:

$$\mathbf{\Gamma}\phi(s_j) = \int \gamma(s_j, s)\phi(s)ds \approx \frac{T}{l}\sum \gamma(s_j, s_l)\tilde{\boldsymbol{\phi}}_l \tag{2.22}$$

In which T represents the length of the interval $\mathcal{T}$ and $\hat{\gamma}(s_j, s_k)$ denotes the elements of the sample variance-covariance matrix $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$. With the help of (2.22), the functional eigenequation $\mathbf{\Gamma}\phi = \lambda\phi$ is approximated by the discrete form

$$\frac{T}{l}\mathbf{\Sigma}\tilde{\boldsymbol{\phi}} = \lambda\tilde{\boldsymbol{\phi}} \tag{2.23}$$

The solutions of (2.23) coincide with the solutions of (2.20). The normalization constraint $\int \phi(s)^2 ds = 1$ is discretely approximated by $\frac{T}{l}\|\tilde{\boldsymbol{\phi}}\|^2 = 1$. For the discrete approximation, $\tilde{\boldsymbol{\phi}} = \frac{T}{l}^{-\frac{1}{2}}\boldsymbol{u}$ is defined, assuming, that $\mathbf{u}$ is a normalized eigenvector of $\mathbf{\Sigma}$.

Concluding, the application of any conventional interpolation method amounts to converting the discrete values $\tilde{\boldsymbol{\phi}}$ into an approximate eigenfunction ϕ. The choice

of interpolation method usually does not essentially impact the results, if the discretization values s_j are closely spaced. [Ramsay and Silverman, 2005, p. 161]

2.2.2 Defining an optimal empirical orthonormal basis

As mentioned above, FPCA can also be seen as the search for a set of k orthonormal functions ϕ_m.

$$\hat{x}_i(t) \approx \sum_{j=1}^{k} \xi_{ij}\phi_j(t), \tag{2.24}$$

whereas ξ_{ij} is the principal component score $\int x_i\phi_j$. In order to evaluate the fit, the integrated squared error is defined as a fitting criterion, which is to be **minimized**:

$$PCASSE = \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2 = \sum_{i=1}^{N} \langle x_i - \hat{x}_i, x_i - \hat{x}_i \rangle \tag{2.25}$$

$$= \sum_{i=1}^{N} \int [x_i(s) - \hat{x}_i(s)]^2 ds = \int [x(s) + \hat{x}(s)]^2 ds \tag{2.26}$$

It can be shown, that the set of principal component functions, which maximize the variance components as defined above, solve this minimization problem and so called **global error**, as well. That is why the basisfunctions ϕ_m are often referred to as *empirical orthonormal functions*, as they are set by the data they are used to expand. [Ramsay and Silverman, 2005, p. 152]

Other than the described discretization approach, each function x_i can be represented as a linear combination of known basis functions ϕ_m, in order to formulate the functional eigenproblem (2.19) in matrix notation.[Ramsay and Silverman, 2005, p. 162] The exact stepwise procedure can be found in [Ramsay and Silverman, 2005, p. 162-164] and [Benko, 2004, p. 31-33].

2.3 Choice of number of Principal Components

After defining the procedures of Principal Component analysis, the question arises, how many principal components should be chosen. On one hand, it is aimed to reduce the number of variables to a much smaller number of relevant principal components. On the other hand, the amount of variation explained by the principal components ought to be rather large in contrast to the amount of variance within

the data. Is the latter rather small, a lot of information is lost about the observed data.

A rather intuitive approach is to choose as many principal components k as needed, in order to explain a certain percentage t% of total variation, which is usually set to be between 70% or 100%. As $V(\xi_j) = \lambda_j$ holds, the total amount of variance within the data can be written as $\sum_j^\infty \lambda_j = w$. Let λ_j be the variance of the jth principal component and w denote the total amount of variation within the data. If all pc's are chosen, the explained variance would equal the total variance within the data. So, the first k principal components account for

$$t_k = 100 \frac{\sum_{j=1}^k \lambda_j}{w} \tag{2.27}$$

of the total variation. The number of principal components required to fulfill a set cutoff value t% is defined as such:

$$k^* = \min \left\{ k | \sum_{j=1}^k \lambda_j \geq \frac{t \cdot w}{100} \right\} \tag{2.28}$$

[Jolliffe, 2002, p. 111-113]

There is a way to visually get a notion of how many principal components are necessary, as well. This is done with the help of the so called *Scree-plot*, which plots the Index of the principal components against the eigenvalues λ_j and will be applied later on. Other criteria to choose a suitable number of principal components for multivariate data can be found in [Jolliffe, 2002] and adjusted for functional data. As the interpretation of the obtained principal component functions can raise difficulties, some visual examinations are looked into in the applied section.

2.4 Regularized principal components analysis

The obtained weight functions estimated solely by decomposition of the covariance operator often are very variable or rough. In order to improve processability of those rough functions, smoothing approaches are drawn on. As for instance, data is pre-processed to obtain functional observations or regressionanalysis requires a certain degree of smoothness, in order to perform well. Now, the focus lies on integrating

smoothing in the functional principal component procedure, as smooth results are more favorable for either methods that work with FPCA results, or the interpretation. Alternatively, the smoothing could also be done before applying PCA, e.g. in the form of smoothing-Splines oder polynomial smoothing. [Ramsay and Silverman, 2005, p. 173]

The basic incorporated idea deals with a roughnes penalty on the principal component functions ϕ, which is represented by:

$$PEN_2(\phi) = \|D^2\phi\|^2 = \int \phi''(t)^2 dt \tag{2.29}$$

The integrated second derivative corresponds to the roughness, as it measures the curvature of a function. Before, the principal components were obtained by maximizing the variance $V(\xi_j) = \frac{1}{N-1}\sum_i \xi_{ij}^2 = \frac{1}{N-1}\sum_i (\int \phi_j x_i)^2$ with respect to $\|\phi_j\|^2 = 1$. The solution of the eigenfunction equation 2.17 solves this optimizationproblem. Now, it is aimed to additionally account for the roughness of the principal components and prevent $PEN_2(\phi_j)$ from getting to large. Therefore, a smoothing parameter $\lambda \geq 0$ is introduced. The so called **penalized sample variance**

$$PCAPSV(\phi) = \frac{var(\int \phi x_i)}{\|\phi\|^2 + \lambda PEN_2(\phi)} \tag{2.30}$$

serves as a new criterion, whereas λ regulates the importance of the roughness penalty term. Its impact shall be demonstrated by two extrem values:

- $\lambda \to 0 : PCAPSV(\phi) = \frac{var(\int \phi x_i)}{1}$, results in the same principal component as in the unpenalized case.
- $\lambda \to \infty$, yields a constant $\phi = a$ in the periodic case and $\phi = a + bt$ in the nonperiodic case, as the penalized variance term is 0. Thus the principal component turns into a constant or a straight line with slope.

The j-th principal component is obtained by maximizing the penalized variance $PCAPSV(\phi_j)$ in respect to two constraints. Firstly, $\|\phi_j\|^2 = 1$ again provides

well-definity. Secondly, a modified version

$$\int \phi_j(t)\phi_k(t)dt + \int D^2\phi_j(t)D^2\phi_k(t)dt = 0, \quad k = 1, \ldots, j-1 \tag{2.31}$$

of the orthogonality constraint is used. [Silverman, 1996, p. 3-5; Ramsay and Silverman, 2005, p. 177/178]

Finding the regularized PCA in practice

An easy approach to finding the smooth principal components, is the idea of basis expansion. Looking at non periodic functions $x_i(t)$, B-splines or orthogonal polynomials up to some degree would be suitable basis functions $\gamma(t)$. Let $\mathbf{c}_i$ resemble the vector of coefficients of the data function $x_i(t)$ in the basis $\{\gamma_v\}$. Any nonperiodic function can now be expanded as a series with coefficients $c_i = \int x\gamma_i$, yielding

$$x_i(t) = \sum_i c_i\gamma_i(t) = \mathbf{c}^T\gamma(t). \tag{2.32}$$

Furthermore,

$$\phi_m(t) = \mathbf{y}_m^T\gamma(t) \tag{2.33}$$

with the vector of coefficients of any potential pc holds. This means, that a possible principal component can be represented by the same basis functions $\gamma(t)$ as $x_i(t)$. $\mathbf{V}$ denotes the covariance matrix of the coefficient vectors $\mathbf{c}_i$, whereas $\mathbf{J}$ is defined as the matrix $\int \gamma\gamma^T$ with elements $\int \gamma_j\gamma_k$. Also let $\mathbf{K}$ be the matrix with elements $\int D^2\gamma_j D^2\gamma_k$. Now, the penalized sample variance is represented by

$$PCAPSV = \frac{\mathbf{y}^T\mathbf{JVJy}}{\mathbf{y}^T\mathbf{Jy} + \lambda y^T\mathbf{Ky}} \tag{2.34}$$

which is equivalent to the eigenproblem

$$\mathbf{JVJy} = \rho(\mathbf{J} + \lambda\mathbf{K})\mathbf{y}. \tag{2.35}$$

An approach to solve this eigenproblem, in order to then obtain the principal component functions ϕ is given in [Benko, 2004, p. 34/35; Ramsay and Silverman, 2005,

p. 180-182]. An explanation of the exact terms behind the fraction pf PCAPSV is given in Appendix 4.1.3.

3 Application to Airquality data

3.1 Description

Observations

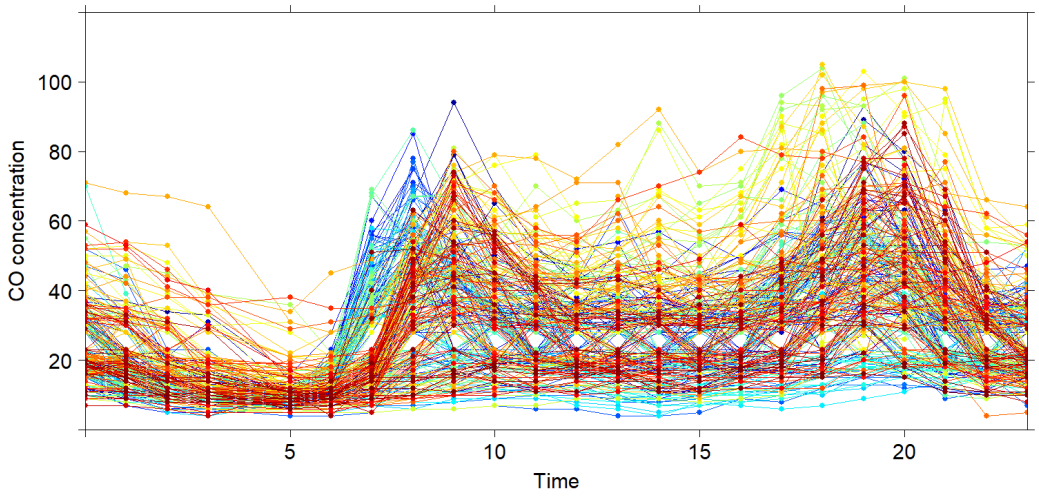
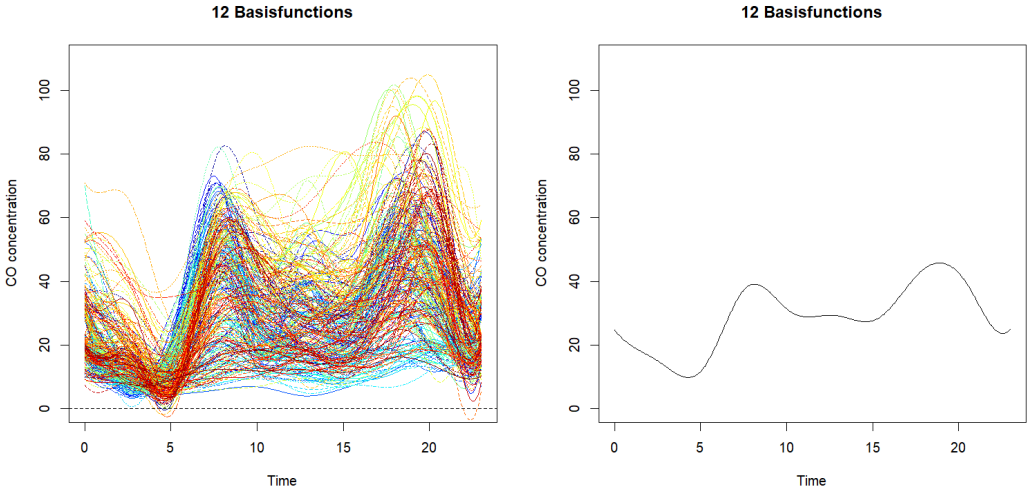


Figure 1: 282 observations (days) represented by lines

The data were collected between 03.2004 and 04.2005 and represent hourly average values of CO concentration $[mg/m^3]$ in an Italian city. I chose an observation to be a 24 hour period, thus representing a day. As many days lacked the fifth value, only the remaining 23 hours were chosen, leading to 282 observations rather than 147. As seen in Figure 1, many observations follow similar curves. Meaning, that after decreasing up until about 7 o clock, the concentration rises to a first peak at approximately 9 o clock. Afterwards it decreases again until about 4 o clock, which is followed by a second peak at around 8 pm. Especially in the time frame between 6 am and 8 pm, the observations seem to vary the most. In order to identify modes of variation and thus get a first view on the structures within the days, FPCA is applied. [**?**, data]

Characteristics

After fitting the data according to 12 B-Spline basisfunctions of order 4, the fitted observations and the resulting meanfunction can be seen in Figure 2. 12 basis functions were chosen, as roughness was to be avoided. In Figure 9 it can be seen, that 12 basisfunctions represent the data in a suitable way, without loosing too much information.

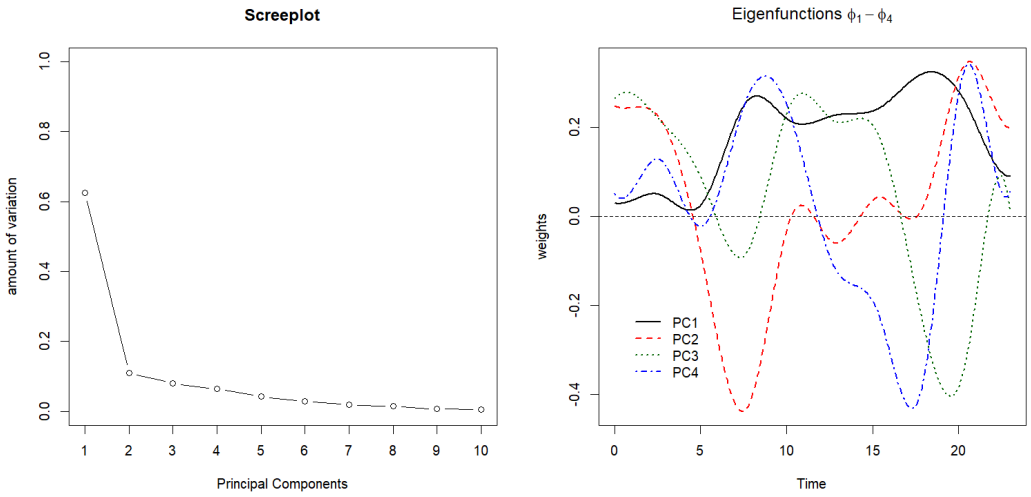


(a) 282 fitted observations (b) mean function of fitted data

Figure 2

3.2 FPCA

Now, the illustrated FPCA procedure is applied and, as mentioned before, visualized. At first, the appropriate number of principal components shall be found using the amount of variation explained by each principal component in the form of the scree plot (Figure 3a). According to it, only the first principal component would be enough, as it accounts for the largest amount of variation (64%). However, four pc's are chosen, in order to explain the graphical methods in a more advanced way. Figure 3b shows the 4 obtained eigenfunctions. The first eigenfunction more or less follows the form of the meanfunction (Figure 2a). This implies, that the first weightfunction is positive and that greatest variability between the days can be found by weighting the hours between 6 am and 9 pm and especially the two peaks



(a) Screeplot (b) eigenfunctions $(\phi_1, \ldots, \phi_4)$

Figure 3

in a stronger way. So the CO concentration is most variable in that period and especially at the two peaks. As the second component fulfills orthogonality to the first component, it's mode of variation is not as important as the first one, which can be seen in the amount of variation explained (11.2%). The second weight function and thus second mode of variation weighs the first and last 5 hours positively with a minimum around hour 7. So, the observations, that score low on this pc, will have greater concentration in the first peak and smaller at night and in the evening.

A further approach to visualize the results of FPCA, is to plot the principal components as perturbations of the mean. Meaning, that the solid line represents the overall smoothed meanfunction $\hat{\mu}(t)$, as seen in Figure 2b. The cross curve marks the effect of adding a suitable multiple of the weight function ϕ_m to the mean, whereas the minus curve represents a suitable subtraction. This is done by

$$\hat{\mu}(t) \pm c\phi_m(t),$$

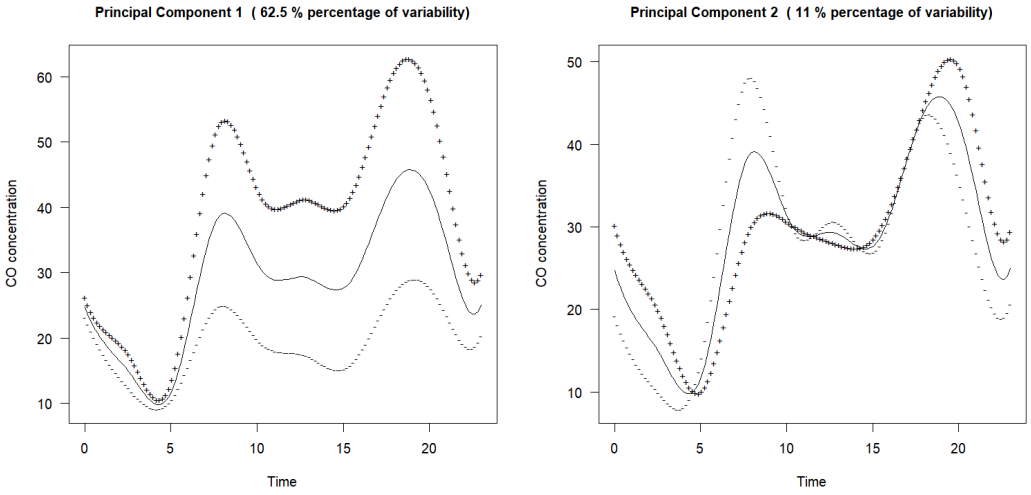
whereas there are many options to choose the constant c. In fact, one choice for c works in terms of the Karhunen-Loeve decomposition for uncentred data, $x(t) \approx \hat{\mu}(t) + \sum_{j=1}^{k} \xi_j \phi_j(t)$. For instance, when examining the effect of the first mode of variation, a new function $\tilde{x}$ with a score vector $\tilde{\boldsymbol{\xi}} = (\pm c, 0, ..., 0)$, is looked at. Then,

the new function

$$\tilde{x}(t) = \hat{\mu}(t) + \sum_{j=1}^{k} \tilde{\xi}_j \phi_j(t) = \hat{\mu}(t) \pm c\phi_1(t)$$

resembles the effect of the first principal components on the outcome compared to the overall mean. In the specific data case, the addition of c illustrates the CO concentration of an observation with a high first principle component score. Subtracting c demonstrates the concentration progression of a low scorer on the first pc. [Ramsay and Silverman, 2005, p. 154/155]

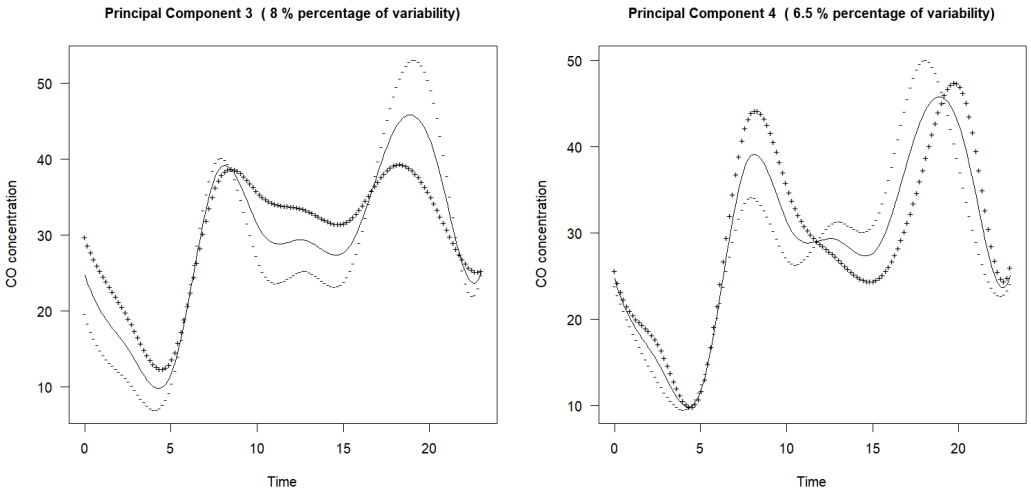
The first (Figure 4a) quantifies the general level of CO concentration throughout



(a) First principal component (b) 2nd pc

Figure 4: First two principal components

the day. A high scorer on this component would show especially above-average concentration in the hours from around 8 am to 8 pm. The CO concentration among all days thus varies the most in this time frame. The second component (Figure 4b) indicates a mode of variability corresponding to below average concentration on the first peak and higher than average in the first and last 5 hours. High scorers are observations with high CO concentration in the first 5 hours, low air pollution around hour 7, who then increase in CO concentration in the afternoon until midnight. On the other hand those with large negative scores tend to be more polluted in the morning and then later state lower concentration than average. Interpretations followed the examples in [Ramsay and Silverman, 2002, p. 27/28].



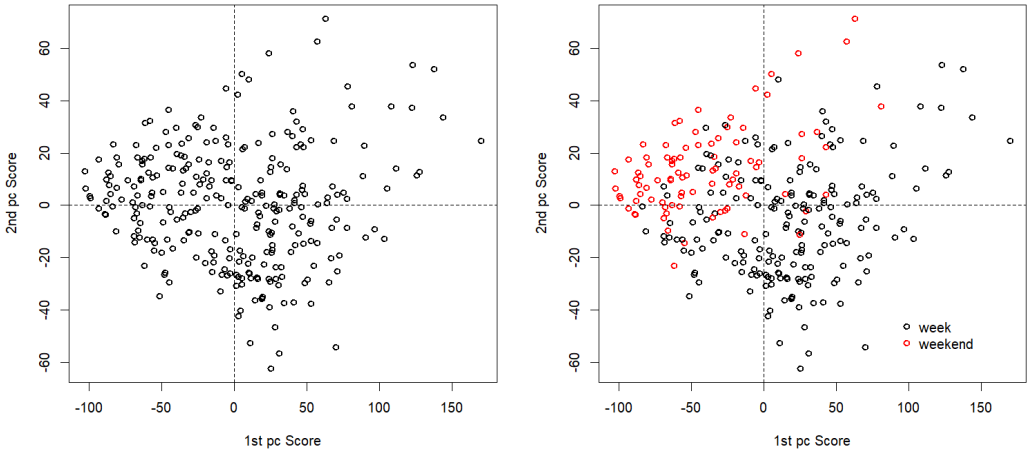
(a) 3rd pc (b) 4th pc

Figure 5

Further principal components usually are harder to interpret. In this case, the third pc could be described as quite opposite to the second, resembling variability especially around the second peak, as seen in 5a.

Another way to interpret the results of FPCA is to examine the principal component scores ξ_{im} of each curve on each component. So for the Airquality data, Figure 6 displays each observation with corresponding scores on the first and second principal component. Figure 6a does not allow a lot of interpretations. Furthermore, separating the obervations into working days and weekends, opens up another possible interpretation. As seen in Appendix (Figure 8), weekends tend not to share the bimodal structure of the overall mean. They seem to be less polluted throughout the day and especially beyond average at the first peak, but tend to show higher than average concentration at night. Bearing in mind the interpretation of the first two components, low general pollution and higher than average pollution at night should yield in a low first and a high second score. That is, why separating the observations into week and weekend days, shows a clear difference in CO concentration. So, weekend observations almost entirely score low on the first principal component, which coincides with the notion of them being less polluted throughout the day and especially between 8 am and 8 pm. Additionally, they mostly score high on the second pc, meaning that they are more polluted in the first and last 5 hours

and less than average polluted at the first peak.



(a) 1st vs 2nd score (b) 1st vs 2nd score with week / weekend

Figure 6

3.3 Regularized FPCA

Now, the smoothing of FPCA via roughness penalty shall also be applied, in order to explain the effect of the smoothing parameter λ. Therefore, the data were fitted according to 20 basisfunctions and thus made rougher (Appendix Figure 10a). However, the approach pursued by Ramsay and Silverman is based on basis representation and thus already posing a certain degree of smoothing on the obervations. Additional to smoothing the data before applying pca, Regularized FPCA serves as a second approach. The obtained eigenfunctions were maximized with respect to the penalized sample variance, which introduced a roughness penalty. In order to demonstrate the effect of λ, it is chosen to be 0.01, 0.5, 1, 10 and 100000. As explained in Section 2.4, setting $\lambda = 0$ (upper left corner) yields the **unpenalized** eigenfunctions as seen in Appendix Figure 10b. The second case shows especially slight differences in the first few hours. It can be seen, that the first 4 lambdas consecutively smooth the eigenfunctions a little bit more each time. $\lambda = 10$ effects the eigenfunctions even stronger and the last case resembles lines, meaning the most extreme case of smoothing (Figure 7, bottom right plot). It can be seen, that the λ combining interpretability and representing information the most optimal way, will

be betweend 0 and 1. To calculate the optimal degree of smoothing, usually cross validation is applied. In this case, the obtained eigenfunctions for $\lambda = 1$ are very similar to the eigenfunctions obtained by fitting the data according to 12 B-spline basisfunctions.

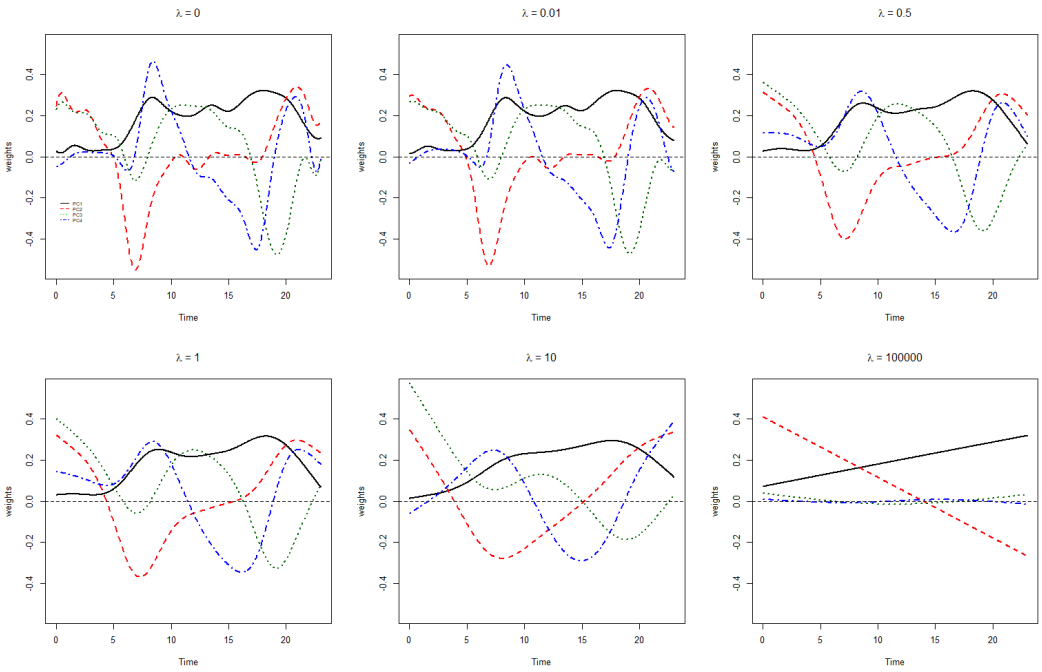


Figure 7: Effect of smmothing parameter λ

References

M. Benko. Functional principal components analysis, implementation and applications. Master's thesis, Humboldt-Universität zu Berlin, 2004.

A. Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 2014.

B. Everitt and T. Hothorn. *An Introduction to Applied Multivariate Analysis with R.* Springer, 2011.

I. Jolliffe. *Principal Component Analysis.* Springer, 2 edition, 2002.

J. Ramsay and B. Silverman. *Applied Functional Data Analysis: Methods and Case Studies.* Springer, 2002.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer, 2005.

B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, 24(1):1–24, 02 1996.

Z. Szabó. Functional data analysis (lecture 6). Regularized functional PCA, Nov. 2016. URL `http://www.cmap.polytechnique.fr/ zoltan.szabo/teaching/FDA/lecture6/lecture6.pdf`.

S. D. Vito. Air quality data set. The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. URL `https://archive.ics.uci.edu/ml/datasets/Air+Quality`.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 2016.

4 Appendix

4.1 Theoretical quantities

4.1.1 Karhunen-Loeve expansion

As mentioned above, we assume the underlying stochastic process $X = X(t)$, $t \in [0, \mathcal{T}]$, to fulfil $\mathbb{E}(X(t)) = 0$ and $\mathbb{E}(X^2(t)) < \infty \; \forall t \in [0, \mathcal{T}]$. Additionally assuming the covariance function $\gamma(s,t)$ to be continuous, then $X(t)$ can be expressed as such

$$X(t) = \sum_{j=1}^{\infty} Z_j e_j(t), \tag{4.1}$$

where convergence is in respect to L^2 and $e_{j\,j\in\mathbb{N}}$ is an orthonormal basis of $L^2[0, \mathcal{T}]$ given by the eigenfunctions of the covariance operator Γ. Other properties are the covariance function $\gamma(s,t)$ with corresponding eigenvalues λ_i satisfying $\lambda_k e_j(t) = \int \gamma(s,t) e_i ds$. $Z_j = \int X(t) e_j(t) dt$ is a sequence of orthogonal random variables with $\mathbb{E}(Z_j) = 0, \mathbb{E}(Z_k^2) = \lambda_k$. [Cuevas, 2014, p. 5]

4.1.2 Singular Value Decomposition:

Let $\mathbf{X}$ be a Nxk matrix. The **singular value decomposition** (SVD) of $\mathbf{X}$ shows as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{W}^T \tag{4.2}$$

- $\mathbf{U}$ is Nxq and orthogonal: $\mathbf{U}^T\mathbf{U} = \mathbf{I}_q$;
- $\mathbf{D}$ is a qxq diagonal matrix with strictly positive diagonalelements;
- $\mathbf{W}$ is nxq and also orthogonal: $\mathbf{W}^T\mathbf{W} = \mathbf{I}_q$

4.1.3 Finding the regularized PCA in practice 1

With assumptions: $\phi(t) = \mathbf{y}^T\gamma(t), \quad x_i(t) = \mathbf{c}_i^T\gamma(t), \quad var(\int \phi x_i) = var(\langle \phi, x \rangle)$

1. Numerator of PCAPSV:

$$\langle \phi, x \rangle = \int \phi(t) x_i(t) dt = \int \mathbf{y}^T \phi(t) \phi^T(t) \mathbf{c}_i dt = \mathbf{y}^T \underbrace{\left[\int \phi(t) \phi(t)^T dt \right]}_{=:\mathbf{J}} \mathbf{c}_i$$

$$var(\langle \phi, x \rangle) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{y^T J c_i})(\mathbf{c_i^T J y}) \right) = \mathbf{y^T J} \underbrace{\left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{c_i c_i^T} \right)}_{=:\mathbf{V}} \mathbf{J y}$$

2. Denominator: With $\phi(t) = \mathbf{y}^T \gamma(t) = \sum_j y_j \gamma_j(t)$ it yields

$$\rightarrow \|\phi\|^2 = \int \mathbf{y}^T \gamma(t) \gamma(t)^T \mathbf{y} dt = \mathbf{y}^T \underbrace{\int \gamma(t) \gamma(t)^T \mathbf{y}}_{=\mathbf{J}},$$

$$\rightarrow PEN_2(\phi) = \|D^2 \phi\|^2 \|D^2 \sum_j \mathbf{y}_j \gamma_j\|^2 = \left\langle \sum_i \mathbf{y}_i D^2 \gamma_i, \sum_j \mathbf{y}_j D^2 \gamma_j \right\rangle$$

$$\sum_{ij} \mathbf{y}_i \mathbf{y}_j \underbrace{\langle D^2 \gamma_i, D^2 \gamma_j \rangle}_{=:K_{ij}} = \mathbf{y}^T \mathbf{K y}$$

$\Rightarrow$ In total:

$$PCAPSV(\phi) = \frac{var\langle \phi, x \rangle}{\|\phi\|^2 + \lambda PEN_2(\phi)} = \frac{\mathbf{y}^T \mathbf{J V J y}}{\mathbf{y}^T \mathbf{J y} + \lambda \mathbf{y}^T \mathbf{K y}}$$

[Szabó, 2016]

4.2 Application to Airquality Data

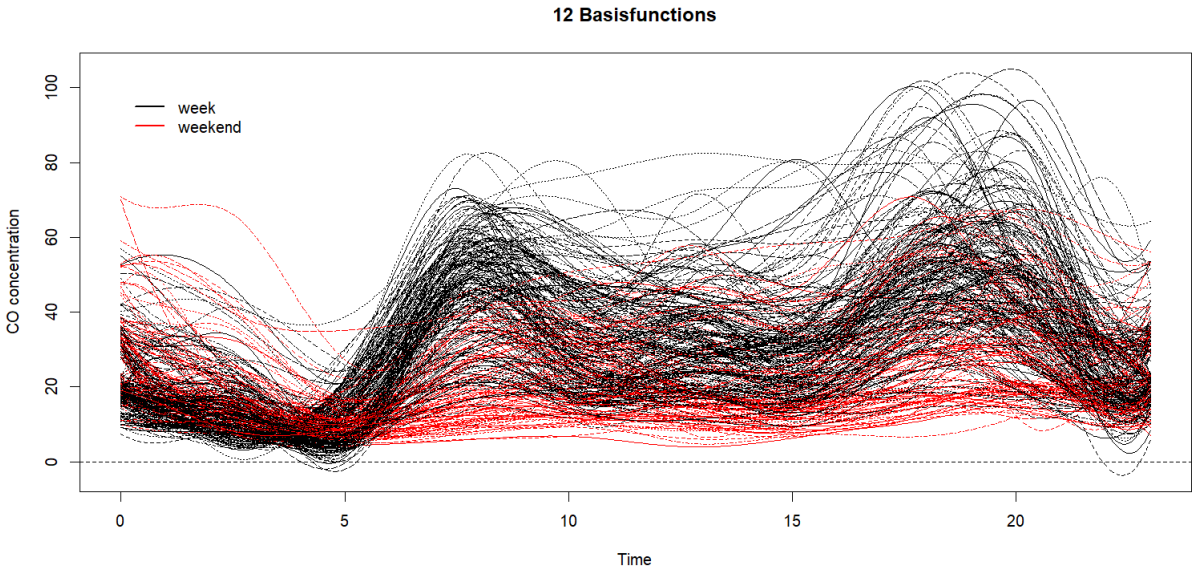
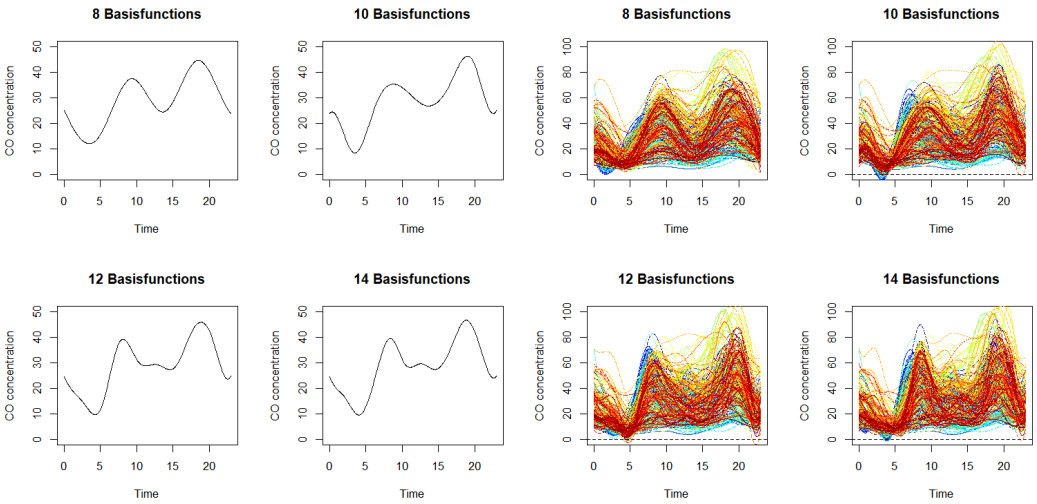


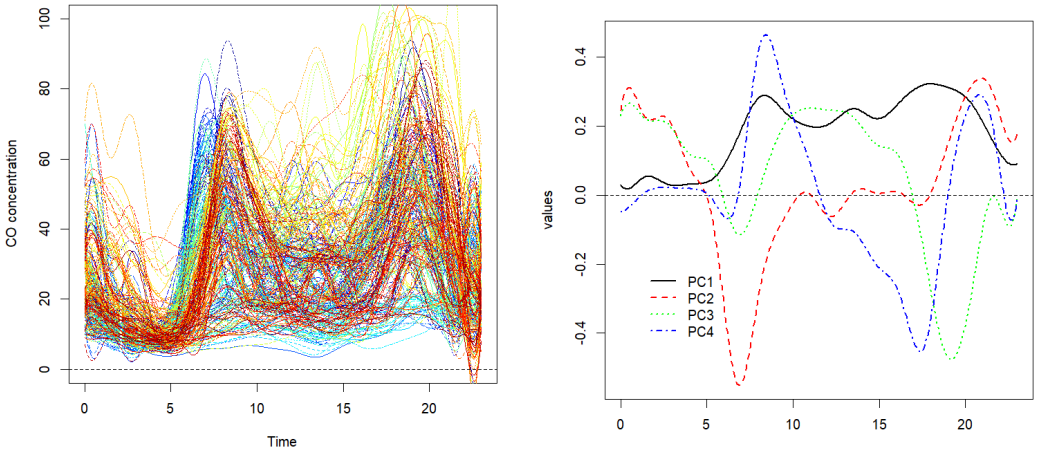
Figure 8:
Fitted data with respect to 12 B-spline basisfunctions separated for weekend and working days



(a) comparison of means for 8,10,12 and 14 basisfunctions

(b) comparison of means for 8,10,12 and 14 basisfunctions

Figure 9



(a) Fitted data with respect to 20 B-spline basisfunctions

(b) corresponding $\phi_1, \ldots, \phi_4$

Figure 10