DIPLOMARBEIT

Robust Functional Principal Component Regression

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Statistik und Wirtschaftsmathematik

eingereicht von

Lukas Neubauer
Matrikelnummer: 01327001

ausgeführt am

Institut für Stochastik und Wirtschaftsmathematik

der Fakultät für Mathematik und Geoinformation der Technischen Universität Wien unter der Betreuung von

Univ.Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser

Wien, 23. August 2021

(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

Kurzfassung

In dieser Diplomarbeit wird das Thema der robusten funktionalen Regression mittels funktionaler Hauptkomponenten behandelt. Funktionale Daten basieren auf einem stochastischen Prozess, der nicht bekannt ist. Wir wollen eine skalare abhängige Variable auf solch einen Prozess regressieren. Wie auch in der multivariaten Statistik hat diese Regression mit Ausreißern zu kämpfen, weshalb ein robuster Zugang notwendig ist. Eine Technik um diese Regression durchzuführen, ist eine Hauptkomponentenanalyse des stochastischen Prozesses. In dieser Arbeit wird eine Einführung zu funktionaler Datenanalyse und funktionaler Hauptkomponenten, sowie zu robuster Statistik gegeben. Insgesamt werden 2 verschiedene Typen von Schätzmethoden verglichen. Die eine Methodik ist für reguläre, dicht beobachtete Daten gemacht ist, wobei die anderen Schätzer einen neuen Zugang für irreguläre, longitudinale Daten bieten. Alle Schätzer werden im Zuge einer Simulationsstudie auf 2 verschiedenen Modellen getestet. Weiters werden verschiedene Szenarien getestet wie reguläre oder irreguläre sowie dichte oder nicht dichte Daten. Alle Daten sind ohne Ausreißer verfügbar sowie auch mit verschiedenen Stufen an Kontamination. Die Resultate sind in manchen Fällen hinreichend aber vor allem im nicht dichten, irregulären kommt es zu schlechten Ergebnissen. Zum Abschluss werden die Schätzer auch an einem Echtdaten-Beispiel getestet. Dieses umfasst kanadische Wetterdaten und wir wollen den jährlichen Niederschlag mittels Temperaturkurven erklären. Alle Schätzer liefern vernünftige Ergebnisse wobei die für longitudinale Daten gemachten Methoden am besten funktionieren.

Abstract

This diploma thesis is about robust functional principal component regression. It is based on functional data where we observe underlying stochastic processes. In a regression setting we want to regress a scalar response onto such stochastic process. As in a multivariate setting this regression is sensitive to outliers in both response and explaining variable, and thus we want to robustify this regression. A common technique for such model is to use functional principal components of the corresponding process and use the resulting scores to explain the response. In this thesis we give a short overview of functional data analysis including functional principal components as well as a brief introduction to robust statistics. We compare two different types of estimators. One is made for regular, densely observed data whereas a new approach for irregular, longitudinal data is proposed. In a simulation study all estimators are applied to 2 models in various settings covering regular and irregular as well as dense or sparse data. The data is used in both clean and contaminated fashion. The results of this simulation study are partly satisfying, especially in regular settings. However, in very sparse, irregular settings the estimators are not as good. Finally, the estimators are applied to a real world example. In the Canadian Weather data we regress the annual precipitation onto temperature curves in various locations. All methods perform comparably while the newly proposed methods seem to work the best.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 23. August 2021

Lukas Neubauer

Contents

1. Introduction

In this diploma thesis the topic of functional data analysis is considered. This kind of data can be seen as a natural extension of multivariate data and is more and more common nowadays due to more and more data being available. In detail, this thesis is about robust functional regression using functional principal components. Such methods were already considered in the past. Febrero-Bande et al. (2017) reviewed this topic rather recently. Yao et al. (2005) proposed a similar approach to what this thesis is about. However, all these methods are not robust and quite sensitive to outliers just as in multivariate statistics. We focus on models where just a few observations per individual are available, i.e. we mainly look at longitudinal data where there are more individuals than corresponding observations per individual.

In Chapter 2 we give a brief introduction to functional data, and give a corresponding mathematical model for such data. Then the idea of functional principal components is annotated as well as various ways of defining them. This is very similar to the multivariate case of principal components. In general the functional principal components are defined using a Karhunen-Loéve expansion of the underlying stochastic process. Afterwards, different regression models regarding functional data are introduced. There are three main models which are the fully-functional model, the scalar-response model and the functional-response model. The scalar-response model can be seen as a simplification of the fully-functional model and is the model which is considered in this thesis. To estimate such model using functional principal components, we generally first need to estimate the mean function of the process, and the eigenfunctions and corresponding scores which are the coordinates of the process under the eigenbasis. Then the scores can be used to express the coefficient function as well as predict the response variable. Ramsay and Silverman (2005) and Horváth and Kokoszka (2012) give excellent summaries of functional data analysis, both on a formal and informal way.

After the introduction to functional data analysis, we give an overview of robust statistics

in Chapter 3. We introduce the most important properties of robust statistics such as breakdown point and statistical efficiency. Then we give the most important robust loss functions which will also be used in the following estimators. After that, we take a look at robust estimators in regression settings, namely regression estimators such as the M-estimator, S-estimator, or MM-estimator. Robust estimators of scale are also considered. Examples are the M-scale estimator and the median absolute deviation (MAD). A more detailed overview of robust statistics can be found in Maronna et al. (2006).

After having covered all necessary parts, the estimators used in this diploma thesis are given in Chapter 4. The first estimators are by Kalogridis and Van Aelst (2019), and are based on a robust projection pursuit approach to estimate the eigenfunctions. Then the corresponding scores are used in an MM-regression step to estimate the coefficient function's coefficients which are used to approximate the coefficient function with respect to the eigenbasis as well predict the response variable. The authors also propose a smoothing step onto the estimator to obtain a much smoother estimator for the coefficient function. Since this method is suited to only regular, and rather densely observed data, we adapted this procedure to also work on irregular data. The corresponding R source code was supplied by the authors. For comparison reasons, we also implement a non-robust estimator based on the same ideas.

In contrast to this method, we take the proposed estimators for the functional principal components of Boente and Salibián-Barrera (2021), and build a comparable yet more general regression framework around it. The authors use elliptically distributed stochastic processes to obtain the best linear approximation of the scores. This method uses conditional expectations to obtain estimators of the scores, in contrast to the above estimators which use a numerical integration approach. In terms of predicting we use the same approach as before, that is, we regress the estimated scores onto the response to obtain the coefficient function's coefficients. The authors also provided a non-robust alternative for their implementation. To emphasize this new robust regression approach, we consider 4 different estimators which use robust or non-robust functional principal component estimation as well as robust or non-robust regression of the scores onto the response. That way we can see which part (FPCA or regression) is more important to be robust in the estimation process. Based on the simulation study of the following chapter, it seems as if the regression part being robust is more influential than the FPCA part.

In Chapter 5 we apply all estimators in a simulation study which contains two different models, both based on the well-known Wiener process. We considered different settings for each model such as varying the number of observations per curve in a regular and also irregular setting. We test the estimators on both clean and contaminated data and measure the goodness in terms of relative mean squared error for estimation and mean squared prediction error. These errors are displayed using boxplots, and the corresponding estimated coefficient functions are also displayed using 20%-trimmed means. The fitted values are plotted against the true response values to see the effect of contamination as well. The results of the simulation study are partly good, especially in the regular setting. However, problems arise in the irregular setting with a very low number of observations per curve.

Last but not least, the methods are applied to a real data set Canadian Weather in Chapter 6. Using this data we want to explain the annual precipation by the corresponding temperature curves in various locations. We modified the data to obtain a longitudinal, and irregularly sampled dataset. All methods perform rather well with the methods based on the elliptical distribution performing the best.

All computations were done in `R` (R Core Team, 2020) and figures were produced using the package `ggplot2` (Wickham, 2016). The Canadian Weather data was obtained from the package `fda` (Ramsay et al., 2020). The `(R)FPCPR` estimators were based on private code provided by the authors while the `sparseFPCA` package (Boente, 2021) is publicly available at *https://github.com/msalibian/sparseFPCA*. All other code parts can be found at *https://github.com/neubluk/RFPCR*.

2. Functional Data Analysis

2.1. Motivation

Over the last years more and more data are accessible meaning many more observations are being recorded in a certain time period. This leads to functional observations rather than multivariate observations. Functional data analysis is based on underlying stochastic processes which are usually not known. The goal is to estimate properties of these processes to highlight interesting characteristics and study patterns and variations. Since these objects are naturally infinite dimensional, a lot of challenges arise. These also depend on how the functions are sampled. We might differentiate between regular data where realizations of a process have all been sampled at the same time points, or irregular data where every sample could have been sampled at totally different time points. For both cases the observed grid of time points can be dense or sparse.

One way to approach functional data, is to apply some sort of dimension reduction. That way both human and computer can better handle such data. This can be compared to dimension reduction known of multivariate statistics where it is used to represent to high dimensional data in just a few dimensions. That way functional data analysis can be seen a natural extension of multivariate data analysis.

A basic example of functional data is growth data. Assume the height is measured for a group of children resulting in a height function. We want to analyze the growing behaviour of a child, i.e. when does it grow the most or when does the growing start to slow down? However, these measurements are not done at the same days indicating irregular data. Functional data analysis can handle such data and can be used to find interesting patterns.

Such data can also be used in regression settings, i.e. we want to investigate the growing effect of children on their self-esteem, for example. For more details of functional data analysis, consult the most prominent work of Ramsay and Silverman (2005).

2.2. Model for Functional Data

To model functional data, a comprehensive mathematical model is needed. In particular, we will use fundamental concepts of functional analysis such as operators in Hilbert spaces, and basics about stochastic processes in such spaces. The following parts have been taken from Horváth and Kokoszka (2012), Part I, Chapter 2.

2.2.1. L^2 Theory

Let $L^2[0,1]$ be the set of measurable functions $f : [0,1] \to \mathbb{R}$ such that $\int_0^1 f(x)^2 dx < \infty$. Such space is a separable Hilbert space with inner product $\langle f, g\rangle = \int_0^1 f(x)g(x)dx$, $f, g \in L^2[0,1]$, and tensor product $(f \otimes g)h = \langle g, h\rangle f$. A very important class of operators are the so-called *integral operators*. Let K be a symmetric, non-negative definite kernel, i.e.

$$K : [0,1] \times [0,1] \to \mathbb{R} \quad \text{such that} \quad K(x,y) = K(y,x),$$
$$\text{and} \quad \int_0^1 K(x,y)f(x)f(y)dxdy \geq 0,$$

for all $f \in L^2[0,1]$. This allows us to define an operator Φ given by

$$\Phi(f)(x) = \int_0^1 K(x,y)f(y)dy, \quad f \in L^2[0,1],$$

which is also symmetric and non-negative definite, that is,

$$\langle \Phi(f), g\rangle = \langle f, \Phi(g)\rangle, \quad f, g \in L^2[0,1], \quad \text{and}$$
$$\langle \Phi(f), f\rangle \geq 0, \quad f \in L^2[0,1].$$

We call Φ a *Hilbert-Schmidt operator* if and only if $\int_0^1 \int_0^1 K(x,y)^2 dxdy < \infty$. A Hilbert-Schmidt operator is always bounded, hence continuous and compact. Thus, by the *spectral theorem* there exists an orthonormal basis of eigenfunctions ϕ_i with corresponding real eigenvalues λ_i (some may be equal to zero) such that

$$\begin{aligned}\Phi(f) &= \sum_{i=1}^{\infty} \lambda_i \langle f, \phi_i \rangle \phi_i \\ &= \int_0^1 f(x) \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i dx \\ &= \int_0^1 f(x) K(\cdot, x) dx.\end{aligned}$$

This implies following result known by *Mercer's theorem*, namely

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y). \tag{2.1}$$

If we consider K to be continuous, then (2.1) holds for all $x, y \in [0, 1]$ and the series converges uniformly.

2.2.2. Probability on L^2

Consider $L^2[0, 1]$ equipped with the Borel σ-algebra and random elements $X = \{X(t) : t \in [0, 1]\}$. X is said to be integrable if and only if $\mathbb{E}[||X||] < \infty$, and square-integrable if and only if $\mathbb{E}[||X||^2] < \infty$. Such random elements allow us to define *mean function* and *covariance operator*, respectively, i.e.

$$\begin{aligned}\mu(t) &= \mathbb{E}[X(t)] \quad \text{and} \\ C &= \mathbb{E}[(X - \mu) \otimes (X - \mu)].\end{aligned}$$

This implies that for the covariance operator we obtain for $f \in L^2[0, 1]$

$$\begin{aligned}C(f)(t) &= \mathbb{E}[\langle X - \mu, f \rangle (X - \mu)(t)] \\ &= \int_0^1 \mathbb{E}[(X - \mu)(s)(X - \mu)(t)] f(s) ds \\ &= \int_0^1 c(s, t) f(s) ds.\end{aligned}$$

The *covariance function* $c = c(s, t)$ is naturally symmetric and non-negative definite. Since X is considered to be square-integrable we also have $\int_0^1 \int_0^1 c(s, t)^2 ds dt < \infty$. Thus, the covariance operator C is a Hilbert-Schmidt operator meaning there exists a count-

able basis of eigenfunctions (ϕ_i) with corresponding real eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots$.

Such process X now admits the so-called *Karhunen-Loève* representation. That is, we can write X as

$$X - \mu = \sum_{k=1}^{\infty} \xi_k \phi_k, \tag{2.2}$$

where the scores $\xi_k, k = 1, 2, \dots$ are uncorrelated random variables with mean 0 and variance λ_k. The scores are nothing but the coordinates of $X - \mu$ on the basis $\{\phi_k : k \geq 1\}$, i.e. $\xi_k = \langle X - \mu, \phi_k \rangle$. The convergence of this series is seen to be the L^2 sense. However, if the covariance function c is continuous then the series converges uniformly in t. The *Karhunen-Loève* expansion will be useful later when looking at principal component analysis. In case of a Gaussian process the scores will also be Gaussian random variables.

Another important fact is the functional *Central Limit* theorem. Consider a sequence of *iid*, mean zero, and square-integrable random elements $X_n,\ n = 1, \dots, N$ in $L^2[0, 1]$. Then the following holds.

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} X_n \xrightarrow{d} \sum_{k=1}^{\infty} Z_k \phi_k,$$

where $Z_k \sim N(0, \lambda_k)$ are independent. Thus, the CLT offers a limiting, Gaussian *Karhunen-Loève* expansion of the underlying process X.

2.2.3. Elliptical Processes

If we drop the need for square integrable processes, the family of *elliptically distributed processes* is quite important, especially in our upcoming estimation methods. Consider a stochastic process X which is not necessarily square integrable, i.e. X is a random element of an arbitrary seperable Hilbert space $\mathcal{H}$ (L^2 is just one example of such space). Given a location parameter $\mu \in \mathcal{H}$ and a self-adjoint, non-negative definite Hilbert-Schmidt operator C with kernel c known as the *scatter operator*, we say $X \sim \mathcal{E}(\mu, C, \phi)$ if and only if for any linear bounded operator $A : \mathcal{H} \to \mathbb{R}^d$, we have that $AX \sim \mathcal{E}_d(A\mu, ACA^*, \phi)$ which corresponds to a multivariate elliptical distribution[1] where A^* denotes the adjoint

[1] A d-dimensional random vector $X = (X_1, \dots, X_n)$ is elliptically distributed $\mathcal{E}_d(\mu, \Sigma, \varphi)$ if and only if the characteristic function of $X - \mu$ can be written as $\phi_{X-\mu}(t) = \varphi(t'\Sigma t)$ for all $t \in \mathbb{R}^d$.

operator of A, i.e. $\langle Ax, y\rangle = \langle x, A^*y\rangle$ for all $x, y \in \mathcal{H}$. One easy example of such process is a Gaussian process for which every finite-dimensional distribution is multivariate normal. The characteristic function $\varphi(x) = \exp(-a/2)$ satisfies above definition. Thus, the elliptical processes can be seen as a generalization of Gaussian processes similarly to the multivariate case.

A very nice property of elliptical processes is as follows. An integrable, elliptically distributed process X has mean function μ, and if the process is also square integrable, then its covariance operator is proportional to C. However, Bali and Boente (2009) showed that without loss of generality one can assume that the covariance operator is equal to C. This is essentially due to $\mathcal{E}(\mu, C, \varphi) = \mathcal{E}(\mu, C, \varphi_a)$ where $\varphi_a(x) = \varphi(x/a)$ and without loss of generality we can set $a = 1$. For more details about elliptical processes and their constructions, the reader may consult Boente et al. (2014). Note that the existence of the covariance operator does not imply the square integrability of the process. Hence, it is sufficient to just assume the existence of the covariance operator to reason the equality to the scatter operator.

Regarding the Karhunen-Loève expansion of (2.2) is is not clear whether such expansion exists for elliptically distributed processes (see Boente and Salibián-Barrera (2021), Prop 3.1). We may only look at processes with $\mu = 0$ since we can always center the process and the properties of the elliptical distribution still hold. We differ between 2 cases. First, consider the case of the scatter operator C having finite rank, i.e. a finite amount of non-zero eigenvalues $\lambda_1 \geq \ldots \lambda_q > 0$ with corresponding eigenfunctions $\phi_1, \ldots, \phi_q$. Let $A : \mathcal{H} \to \mathbb{R}^d$ be a linear operator such that $Af = (\langle f, \phi_1\rangle, \ldots, \langle f, \phi_q\rangle)'$. Since A is also bounded by q, the definition of the elliptical distribution applies and we obtain that $AX = (\xi_1, \ldots, \xi_q)' \sim \mathcal{E}_q(0, ACA^*, \varphi)$ where $ACA^* = \operatorname{diag}(\lambda_1, \ldots, \lambda_q)$. Because $\lambda_l = 0$ for $l > q$, we obtain that $X \overset{d}{=} \sum_{k=1}^{q} \xi_k \phi_k$. In the case of infinite rank but continuous kernel function c, it is shown in Boente et al. (2014), Prop. 2.1 that X can be expressed as $X \overset{d}{=} SV$ where V is a centered Gaussian process and S is a non-negative random variable independent of V. V naturally allows a normal Karhunen-Loève expansion such that $V = \sum_{k=1}^{\infty} \eta_k \phi_k$ where $\eta_k \sim N(0, \lambda_k)$. Since the covariance function of V is also c which is continuous the series converges uniformly and we can write $X(t) = \sum_{k=1}^{\infty} S\eta_k \phi_k(t) = \sum_{k=1}^{\infty} \xi_k \phi_k(t)$ where $(\xi_1, \ldots, \xi_q)' \sim \mathcal{E}_q(0, \operatorname{diag}(\lambda_1, \ldots, \lambda_q), \varphi)$. Hence, we can also expand an elliptically distributed process using the covariance opera-

tor's eigenfunctions.

2.3. Functional Principal Component Analysis

Before going into functional principal component analysis (PCA), we revise PCA in a multivariate setting. Its aim is to simplify complex relationships by reducing their dimensionality. However, we still want to preserve information which is considered to be the variance in this case.

2.3.1. Multivariate PCA

Consider a p-dimensional random variable $X = (X_1, \dots, X_p)$ with mean $\mu = \mathbb{E}[X]$ and covariance matrix $\Sigma = \mathbb{E}[XX'] - \mathbb{E}[X]\mathbb{E}[X']$. The goal is to find an orthogonal matrix Γ such that the variances of $Z = \Gamma(X - \mu)$ are maximized. This corresponds to minimal information loss. In the literature usually three ways to find such transformation are used. We call the matrix Γ the *loadings* matrix. More details may be found in Anderson (2003), Chapter 11 and Filzmoser (2020), Chapter 5.

Eigendecomposition of the covariance matrix Σ

The covariance matrix of $Z = (Z_1, \dots, Z_p)'$ is given by $\mathrm{Cov}(Z) = \Gamma\Sigma\Gamma'$. Maximizing the variances of $Z_1, \dots, Z_p$ subject to the orthogonality of Γ leads to $\Gamma\Sigma\Gamma' = A$ where $A = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are the eigenvalues of Σ and Γ contains the corresponding eigenvectors columnwise. Hence we obtain an eigendecomposition of Σ. That being said, the covariance of Z now reads $\mathrm{Cov}(Z) = A$, and in particular, $\mathrm{Var}(z_1) = \lambda_1 \geq \mathrm{Var}(z_2) = \lambda_2 \geq \cdots \geq \mathrm{Var}(z_p) = \lambda_p$. Indeed, the random variable Z has uncorrelated components (meaning that each component carries new information) and the components are ordered corresponding to their individual variance. This also allows us to define *variance explained* by the first K components, $\sum_{i=1}^{K} \lambda_i / \sum_{i=1}^{p} \lambda_i$.

In case of observed data $X \in \mathbb{R}^{n \times p}$ we equivalently define $Z = (X - \mathbb{I}\hat{\mu}')\hat{\Gamma}'$ where $\mathbb{I}$ denotes the n-dimensional vector of ones. As before, the optimization problems yields $\hat{\Gamma}' S \hat{\Gamma} = \hat{A}$. Note that $\hat{\mu}$ and S are the standard estimators of the mean and covariance of X, respectively. The matrix Z is called the *scores* matrix and displays the original data X in a different coordinate system defined by $\hat{\Gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_K)$. One can now choose a certain amount of *principal components* $\hat{\gamma}_1, \dots, \hat{\gamma}_K$ to transform the data to a simpler representation. Thus,

the dimension of X has been reduced while retaining as much information (variance) as possible. The number $K \leq p$ can be found using the explained variance which must be bigger than a certain threshold. As p grows with n remaining in the same order of magnitude, the estimation of S becomes more and more infeasible, indicating a need for alternative methods.

One particular problem arises when having *flat* data as in $n << p$. Usual estimation techniques for the covariance matrix will fail due to singularity problems. For such data, a *singular value decomposition* of the data matrix X can help. For a centered data matrix X we can write $X = UDV'$ where $U \in \mathbb{R}^{n \times n}, V \in \mathbb{R}^{p \times p}$ are orthogonal matrices and $D \in \mathbb{R}^{n \times p}$ is diagonal with $d_{ii} \geq 0, i = 1, \ldots, \min(n, p)$. One can quickly see the relationship between this method and the eigendecomposition. Setting $\hat{\Gamma} = V, Z = UD$ yields $X = Z\hat{\Gamma}' = UDV'$, and $d_{ii}^2 \propto \hat{\lambda}_i$.

Projection Pursuit

Another way of defining PCA is to consider a different maximization problem. The maximal variance property suggests looking at the optimization problem

$$\max_{\gamma:||\gamma||=1} \mathrm{Var}(\gamma' x) = \max_{\gamma:||\gamma||=1} \gamma' \Sigma \gamma.$$

By linear algebra, the solution of this problem is obtained to be $\gamma = \gamma_1$ with optimal objective value of $\gamma_1' \Sigma \gamma_1 = \lambda_1$. Subsequent optimal projection directions can be found by constraining on orthogonality, i.e.

$$\max_{\gamma:||\gamma||=1, \gamma'\gamma_i = 0, i<k} \mathrm{Var}(\gamma' x) = \lambda_k,$$

attained at $\gamma = \gamma_k$. Therefore, this method yields the same solution as the previous one.

Minimizing Projection Residuals

A further possibility is to minimize the error between x and $\pi(x, L)$, the orthogonal projection of x into a subspace L of dimension m. Then the optimal subspace is spanned by the first m eigenvectors of Σ, denoted by L_0, i.e.

$$\mathbb{E}[||x - \pi(x, L_0)||^2] \leq \mathbb{E}[||x - \pi(x, L)||^2],$$

as long as $\lambda_m < \lambda_{m+1}$. In case of observed data, we can make use of the singular value decomposition and write

$$X = XVV' = XV_mV_m' + E,$$

such that XV_m are the first m principal components. Next, we want to minimize $||E||_F = ||X - XBB'||_F$ where $||\cdot||_F$ denotes the *Frobenius norm*, and the orthogonal matrix B satisfies $\text{rank}(B) \leq m$. This is equivalent since XBB' is a rank m approximation of X and one can show that

$$||X - XV_mV_m'||_F \leq ||X - XBB'||_F,$$

for any orthogonal, lower rank matrix B.

2.3.2. Functional PCA

We will see that the concepts of multivariate PCA can be easily transferred to a functional setting. Let's consider the expansion of the process X with covariance operator C and mean function μ in equation (2.2) which also holds if the process is not square integrable. In terms of independent trajectories we obtain

$$X_i(t) - \mu(t) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t),$$

where the scores ξ_{ik} are independent across i and uncorrelated across k. Since the truncated series converges in mean square, this expansion allows us to perform dimension reduction. If the first K terms for some large K provide a sufficiently good approximation to the finite sum, and hence for the trajectory, we may look at the approximated process in

$$X_i^{(K)}(t) - \mu(t) = \sum_{k=1}^{K} \xi_{ik}\phi_k(t). \tag{2.3}$$

That means the information of X_i is essentially contained in $(\xi_{i1}, \ldots, \xi_{iK})'$. Such expansion can also be done in different bases such as a Fourier basis. What makes the eigenbasis special, is that for fixed K the expansion using eigenbasis explains most variation of X in the L^2 sense, and is therefore optimal.

This leads us to the three equivalent ways of performing multivariate PCA. First, to obtain the eigencomponents ϕ_k and λ_k we can essentially do an **eigendecomposition of the covariance operator**, we solve following integral equations for λ and f such that

$$\int_0^1 c(s,t)f(s)ds = \lambda f(t), \tag{2.4}$$

or, equivalently,

$$C(f) = \lambda f,$$

where c is the covariance function corresponding to the covariance operator C. However, we can also look at a **projection-pursuit approach**. We want to maximize

$$Var(\langle X - \mu, \phi\rangle) = \langle C(\phi), \phi\rangle \quad \text{subject to } ||\phi|| = 1. \tag{2.5}$$

As in the finite case, the supremum is attained at $\phi = \phi_1$ with objective value $\langle C(\phi_1), \phi_1\rangle = \lambda_1$. Further eigencomponents are obtained by setting constraints of orthogonality on the previous eigenfunctions. Finally, we may also look at **minimizing the residuals** between X and its projection onto a subspace in terms of mean squared error. As denoted before, we obtain $\mathbb{E}[||X - \pi(X, L_0)||^2] \leq \mathbb{E}[||X - \pi(X, L)||^2]$.

For more information about FPCA see Horváth and Kokoszka (2012), Part I, Chapter 3 and Ramsay and Silverman (2005), Chapters 8,9.

2.4. Functional Regression

The common *multivariate linear model* is given by $y = X\beta + \epsilon$ with n-dimensional response vector $y = (y_1, \ldots, y_n)'$, regressor matrix $X = (X_1, \ldots, X_p)$ of dimension $n \times p$ where X_j denotes the j-th explaining variable and unknown p-dimensional coefficient vector β. The error term is denoted by $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$. This model is quite versatile but does have some limitations. In certain applications the regressor is of functional form. This is especially the case in growth data. Even though the observed data is still discrete, we might assume an underlying stochastic process from which we basically sample independent trajectories. Hence ordinary least squares methods are not approriate anymore.

We distinguish between three basic functional models. For simplicity we assume mean zero

for both the responses and regressors. Additionally, the error term ϵ_i is always assumed to be independent of the explanatory variables X_j. All following models have the same main issue. Given a finite number of observations we want to estimate a infinite dimensional function. Thus, we usually impose some constraints on the unknown coefficient function such as roughness penalties. Certain constraints will lead us to the use of principal components.

Fully-Functional Model

In such model, both the response and regressors are assumed to be functional. The model then is

$$Y(t) = \langle X, \beta(t, \cdot)\rangle + \epsilon(t), \tag{2.6}$$

where the unknown function β is a surface such that $\beta(t, s)$ reflects the effect of X at time s on the response Y at time t. At first, it is not clear how to estimate the coefficient surface β. However, using expansions of equation (2.2) for both X, Y leads us to the following representation of β. Assume expansions of

$$X(s) = \sum_{k=1}^{\infty} \xi_k \phi_k(s), \quad Y(t) = \sum_{k=1}^{\infty} \zeta_k \psi_k(t),$$

where the $\xi_k' s, \zeta_k' s$ are the FPCs of X and Y, respectively. Further, assume that $\int_0^1 \int_0^1 \beta(t, s)^2 ds dt < \infty$. Then we can write the coefficient surface as

$$\begin{aligned}
\beta(t, s) &= \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} b_{kl} \phi_l(t) \psi_k(s) \\
&= \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{\mathbb{E}[\xi_l \zeta_k]}{\mathbb{E}[\xi_l^2]} \phi_l(s) \psi_k(t) \\
&= \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{\mathrm{Cov}(\xi_l, \zeta_k)}{\lambda_l} \phi_l(s) \psi_k(t),
\end{aligned} \tag{2.7}$$

where the convergence is seen to be in L^2. Plugging in expression (2.7) into the model (2.6) yields following expression for the response,

$$Y(t) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{\mathrm{Cov}(\xi_l, \zeta_k)}{\lambda_l} \xi_l \psi_k(t). \tag{2.8}$$

This equivalent way of writing the response Y will turn out very useful in our further analysis.

To give some relation to the multivariate linear model where the coefficient vector is found using the normal equations $X'X\beta = X'y$, consider two bases $(\eta_k), (\theta_k)$ for expanding X and Y, respectively. Denote $Y(t) = (Y_1(t), \dots, Y_N(t))', X(t) = (X_1(t), \dots, X_N(t))'$ and $\epsilon(t) = (\epsilon_1(t), \dots, \epsilon_N(t))'$, then we can write the corresponding model for N observations as

$$Y(t) = \langle X, \beta(t, \cdot)\rangle + \epsilon(t), \tag{2.9}$$

analogously to the multivariate case. Now the idea is to expand β such that

$$\beta(t, s) = \sum_{k=1}^{K}\sum_{l=1}^{L} b_{kl}\eta_k(t)\theta_l(s),$$

i.e. we assume β can be expanded using a finite number of terms. The optimal coefficients (b_{kl}) are found by least squares estimation, i.e. we want to minimize

$$\sum_{i=1}^{N} ||Y_i - \langle X_i, \beta(t, \cdot)\rangle||^2,$$

where the inner product is computed componentwise. To derive the normal equations, we first express β as

$$\beta(t, s) = \eta(t)'B\theta(s), \tag{2.10}$$

where $\eta(t) = (\eta_1(t), \dots, \eta_K(t))', \theta(t) = (\theta_1(t), \dots, \theta_L(t))'$, and the $K \times L$ matrix $B = (b_{kl})$. Plugging expression (2.10) into the model (2.9) we obtain

$$Y(t) = \underbrace{\langle X, \eta'\rangle}_{=Z} B\theta(t) + \epsilon(t). \tag{2.11}$$

Letting $J = \langle \theta, \theta'\rangle$, equation (2.11) implies that

$$\langle Y, \theta'\rangle = ZBJ + \langle \epsilon, \theta'\rangle.$$

Ignoring the error terms and multiplying by Z' leads to the analog of the normal equations,

i.e.

$$Z'ZBJ = Z'\langle Y, \theta' \rangle, \tag{2.12}$$

which needs to be solved for B. This can be done using tricks of linear algebra. We write $\text{vec}(A)$ for a column vector consisting of the columns of the matrix A. Then $\text{vec}(CDE)$ can be written as $\text{vec}(CDE) = (E' \otimes C)\text{vec}(D)$. Hence we might write (2.12) as

$$(J' \otimes Z'Z)\text{vec}(B) = \text{vec}\left(Z'\langle Y, \theta' \rangle\right),$$

and, assuming the corresponding matrices are regular,

$$\text{vec}(B) = (J' \otimes Z'Z)^{-1}\text{vec}\left(Z'\langle Y, \theta' \rangle\right).$$

Regarding an example, consider the Canadian Weather data of Chapter 6. We could use a fully-functional model to explain the precipitation at time t using the temperature across the whole year.

Scalar-Response Model

Assuming the response not to be functional, we can simplify the fully-functional model to

$$Y = \langle X, \beta \rangle + \epsilon. \tag{2.13}$$

As before, we obtain a way of expressing the coefficient function using equation (2.7) as

$$\beta(t) = \sum_{k=1}^{\infty} \frac{\text{Cov}(Y, \xi_k)}{\lambda_k} \phi_k(t), \tag{2.14}$$

where the convergence is again seen to be in L^2. The scalar-response model can then be written as

$$y = \sum_{k=1}^{\infty} \underbrace{\frac{\text{Cov}(\xi_k, y)}{\lambda_k}}_{=b_k} \xi_k.$$

A corresponding estimator is given by

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{b}_k \hat{\phi}_k(t), \tag{2.15}$$

and

$$\hat{y} = \sum_{k=1}^{K} \hat{b}_k \hat{\xi}_k. \tag{2.16}$$

The coefficients of β can be estimated by regressing y onto the estimated scores $\hat{\xi}_k$. In this thesis we will focus on scalar-response models. An example for such model can be seen in Chapter 6 where the annual cummulated precipitation is explained using the corresponding temperature curves.

Functional-Response Model

The simplest of such models is given by

$$Y(t) = X\beta(t) + \epsilon(t),$$

which is usually extended to the use of more parameter functions such as

$$Y(t) = \sum_{j=1}^{L} X_j \beta_j(t) + \epsilon(t).$$

More details about this kind of models can be found in Ramsay and Silverman (2005), Chapter 13. A slight adaption is to let $X = X(t)$ leading to **functional-concurrent models** which may be used to explain the precipitation at time t using only the temperatures at the same time in the Canadian Weather data.

3. Robust Statistics

In this chapter we take a look at robust statistics, and the methods used in the later estimators. In general, statistical methods require rather strict model assumptions. However, these are often not fully fulfilled which may lead to erroneous analyses. Robust methods usually focus on fitting a model only on a subset of the data where the requirements hold. That way one can still obtain sensible results which would be biased when using non-robust methods.

For further analysis, we need some basics of robust statistics. One essential definition is the one of the **breakdown point**. Consider an arbitrary estimator T, and data X with n observations. We write $\tilde{X}$ for the contaminated data set where $m < n$ observations are contaminated. The breakdown point of T is then defined to be

$$\epsilon_n^*(T, X) = \min \left\{ \frac{m}{n} : \sup_{\tilde{X}} ||T(X) - T(\tilde{X})|| \right\}.$$

By letting $n \to \infty$ we obtain the **asymptotic breakdown point** defined as $\epsilon^*(T, X) = \lim_{n\to\infty} \epsilon_n^*(T, X)$. The maximum asymptotic breakdown point in common regression scenarios is $\epsilon^* = 0.5$ meaning we can have half the data compromised and still obtain reasonable results. In case of ordinary least squares regression we have $\epsilon^* = 0$. The goal is to obtain an estimator with high breakdown point which also has additional desirable statistical properties.

A high breakdown point often goes hand in hand with the estimator's **statistical efficiency** defined to be

$$\text{eff}(T) = \frac{1/I(\theta)}{\text{Var}(T)} \leq 1,$$

where θ is the parameter estimated by T, $\text{Var}(T)$ is estimator's variance and $I(\theta)$ denotes its respective Fisher information. Since the Fisher information depends on a distribution

assumption of the population, one usually considers efficiency with respect to a Gaussian distribution. We want robust estimators to be efficient, i.e. having low variance, and still have a high breakdown point. Oftentimes robust methods will have lower efficiency compared to non-robust alternatives, yet they still outperform non-robust estimators.

Another quite important feature in statistics in general is **consistency** and **Fisher consistency**. An estimator T for θ is consistent if and only if T converges to θ in probability. Assuming F is the true distribution with parameter θ from which the observed data was sampled, we say T is Fisher-consistent if and only if $T(F) = \theta$. That is, the estimator gives the correct value of θ when using the true population as opposed to a finite sample.

3.1. Robust Loss Functions

Consider a generic model $y = f(\beta; x) + \epsilon$ for some function f. We usually want to minimize the squared residuals, that is,

$$\hat{\beta} = \text{argmin}_\beta \sum_i (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = f(\hat{\beta}; x)$ denote the fitted values. However, this procedure is sensitive to outliers meaning that unusual observations lead to a biased estimator. Hence, in robust statistics one replaces the unbounded squared loss function with a more general *ρ-function* which give large residuals lower weight in the estimation process. A ρ-function is characterised by certain properties. Commonly one assumes properties alike

- $\rho(0) = 0$,
- ρ is even, i.e. $\rho(x) = \rho(-x)$ for all $x \in \mathbb{R}$,
- ρ is nondecreasing in $|x|$, and
- ρ is increasing for $x > 0$ and $\rho(x) < \sup_x \rho(x)$.

Several of such functions are used in practice. Some of them are displayed below. One common property is that they have bounded derivatives.

- *Huber* family of functions given by

$$\rho_k(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq k, \\ k\left(|x| - \frac{1}{2}k\right) & \text{otherwise.} \end{cases} \tag{3.1}$$

- *Tukey's bisquare* family of functions given by

$$\rho_k(x) = \begin{cases} 1 - \left(1 - \left(\frac{x}{k}\right)^2\right)^3 & |x| \leq k, \\ 1 & \text{otherwise.} \end{cases} \tag{3.2}$$

While Huber functions are not bounded, Tukey bisquare functions are as seen in Figure 3.1. This might be an important difference when looking at the upcoming estimation methods.

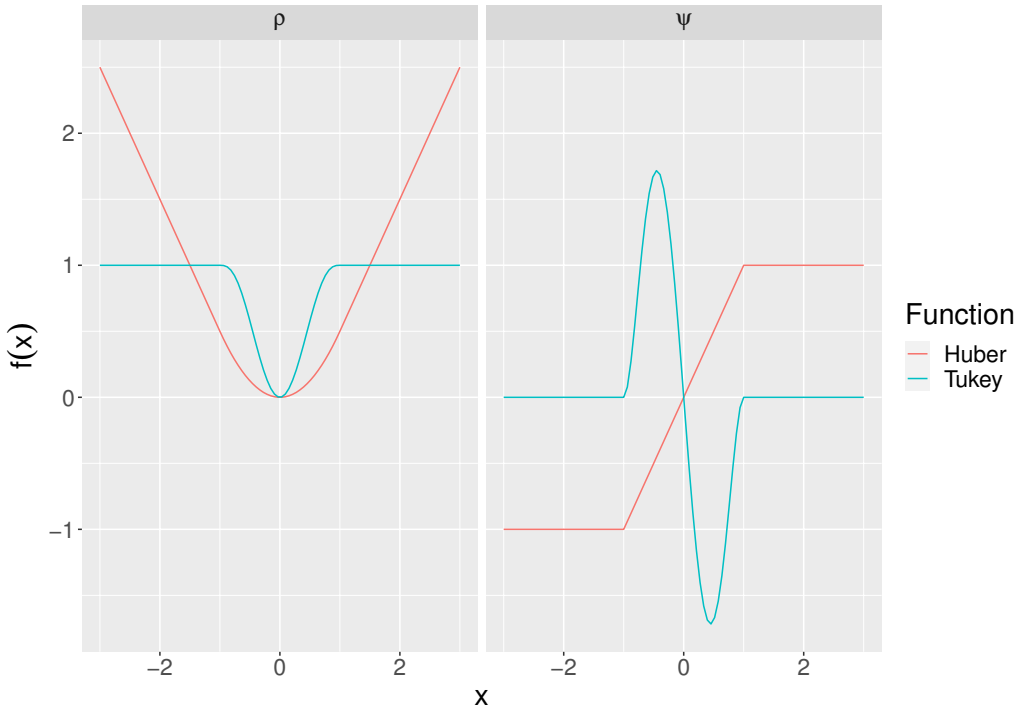

Figure 3.1.: Robust Loss Function denoted by ρ and their derivatives ψ for $k = 1$.

3.2. Robust Estimators of Regression

In this section we take a closer look at estimating the coefficient in a linear model $y = X\beta+\epsilon$. The **M-estimator** minimizes the corresponding sum of residuals via the ρ-function, i.e.

$$\hat{\beta} = \text{argmin}_\beta \sum_i \rho\Big(\frac{\overbrace{y_i - x_i'\beta}^{=r_i(\beta)}}{\hat{\sigma}}\Big), \tag{3.3}$$

where x_i denotes the i-th row (as a column vector) of the data matrix X. The residuals are scaled using a robust scale estimator to diminish scaling effects of the data. The minimization problem can be solved by taking derivatives and setting them to 0. The resulting equations are the so-called *M-estimating equations*

$$\sum_i \psi\left(\frac{r_i(\beta)}{\hat{\sigma}}\right) x_i = 0, \tag{3.4}$$

where $\psi = \rho'$. We might solve these equations by an *iterative reweighted least squares* procedure. For that we set $W(x) = \psi(x)/x$ and $w_i = W(r_i(\beta)/\hat{\sigma})$. Thus, (3.4) becomes

$$\sum_i w_i r_i(\beta) x_i = 0.$$

Assuming an initial robust estimator $\hat{\beta}_0$, one can then calculate the best estimator step-by-step until convergence. The choice of $\hat{\sigma}$ is also essential and is discussed in the next section.

Another way to obtain robust regression estimators is by considering a different minimization problem. Let $\hat{\sigma}$ be an arbitrary robust scale estimator. Then a robust estimator for β is obtained by

$$\hat{\beta} = \text{argmin}_\beta \hat{\sigma}(r(\beta)). \tag{3.5}$$

The simplest robust scale estimator is the median of the absolute residuals which leads to the *least median of squares* estimator. A very important estimator of this form is obtained by solving an equation of scale given by

$$\frac{1}{n}\sum_i \rho\left(\frac{r_i}{\hat{\sigma}}\right) = b. \tag{3.6}$$

First solving (3.6) with respect to $\hat{\sigma}$ gives us a robust scale estimator. The corresponding regression estimator using $\hat{\sigma}$ in (3.3) yields the **S-estimator**. This estimator by itself is not ideal since it does not achieve high efficiency. However, it is useful as our initial estimator in (3.3) leading to the **MM-estimator**.

Another very important class of robust location estimators is given by **local M-estimators**, especially **local linear smoothers**. To simplify concepts, we consider a linear model $y = a + bx + \epsilon$. In fact, the notation of neighbourhoods is more complicated in higher dimensions. Given x_0 in the range of observed x values, we define a local linear smoother by

$$\left(\hat{a}(x_0), \hat{b}(x_0)\right)' = \underset{a(x_0), b(x_0)}{\operatorname{argmin}} \sum_i w_i(x_0) \rho \left(\frac{y_i - a(x_0) - b(x_0) x_i}{\hat{\sigma}(x_0)} \right), \tag{3.7}$$

where $\hat{\sigma}(x)$ is a local robust estimator of scale, and the weights w_i are given by

$$w_i(x_0) = \frac{K\left(\frac{x_0 - x_i}{h}\right)}{\sum_j K\left(\frac{x_0 - x_j}{h}\right)}.$$

The function K is a kernel, i.e. a continuous, non-negative, symmetric, and integrable real-valued function, and h denotes the bandwidth parameter.

3.3. Robust Estimators of Scale

Next, we want to robustly estimate the scale in a linear model. One such estimator was already mentioned in (3.6). An estimator satisfying this equation is called **M-scale estimator**. The ρ-function is chosen to be bounded and b is a constant. This constant determines the estimator's breakdown point. In fact, we have that $\epsilon^* = \min(1, 1 - b)$. Thus, a natural choice of b is $b = 1/2$.

Another robust estimator of scale already mentioned is the **MAD** (median absolute deviation) for data x is given by

$$\text{MAD} = c \, \text{med}|x - \text{med}(x)|,$$

where the constant c is usually chosen such that the estimator is consistent for the parameter

of interest. This principle can also be applied locally, leading to the **local MAD**. Given a bandwidth h it is defined as

$$\mathrm{MAD}_h(x_0) = c \underset{|x-x_0|<h}{\mathrm{med}} \left| x - \underset{|x-x_0|<h}{\mathrm{med}}(x) \right|. \tag{3.8}$$

An alternative to the MAD which is more efficient is the **$\mathbf{Q_n}$-estimator** given by

$$Q_n = c\{|x_i - x_j| : i < j\}_{(k)}, \tag{3.9}$$

where c is a constant, and $k = \binom{h}{2}$ with h is roughly half the number of observations. That is, we take the k-th order statistic of these absolute distances.

A final robust scale estimator is the **$\boldsymbol{\tau}$-scale estimator** which offers both high efficiency and a high breakdown point. Consider the M-scale estimator as in (3.6) and denote it by s_n. The scale estimator τ_n is then defined by the solution of

$$\tau_n^2 = s_n^2 \frac{1}{n} \sum_i \rho_1 \left(\frac{r_i}{s_n} \right), \tag{3.10}$$

where ρ_1 is also a ρ-function. If we choose $\rho = \rho_1$ then $\tau_n = \sqrt{b} s_n$. Similarly to the S-estimator, the τ regression estimator is obtained by minimizing $\tau_n = \tau_n(r_1, \ldots, r_n)$ where $r_i = r_i(\beta)$. Yohai and Zamar (1988) showed that their τ-estimator behaves asymptotically as an M-estimator with ψ-function equal to a weighted average of the two ψ-functions corresponding to the τ-estimator. Under certain conditions the τ-estimates have breakdown point equal to 0.5, are consistent and are 95% efficient under a Normal distribution.

For more details about robust statistics, see the prominent book of Maronna et al. (2006), or Filzmoser (2020), Chapter 4.

4. The Estimators

This chapter goes into detail on two types of estimators used in the following model. As in Section 2.2.2 we consider stochastic processes $X = \{X(t) : t \in [0,1]\}$. The regression model is as in (2.13), i.e. $y = \langle X, \beta \rangle + \epsilon$. Given observations of y and realizations of X we want to estimate β, and predict the responses y. Specifically, we will differ between regularly and irregularly observed realizations of the process X. Simulation results can be seen in Chapter 5 and an application to real data in Chapter 6.

4.1. RobustFPCR

In this section we introduce estimators which are viable for data where all curves have been observed at the same time points. Hence the observations can be seen as a regressor matrix. The following estimators are based on the work of Kalogridis and Van Aelst (2019).

Preliminaries

The authors assume square integrable processes X and in order to ensure properties such as Fisher consistency and consistency in terms of L^2 of the following estimators, a few assumptions have to be made.

1. The process X allows a finite Karhunen-Loève expansion (2.2), i.e.

$$X - \mu = \sum_{k=1}^{K} \xi_k \phi_k.$$

2. The scores $\xi_1, \ldots, \xi_K$ are absolutely continuous and have joint density $g(\boldsymbol{x}) = h(||\boldsymbol{x}||_2)$ for some measurable, positive function h.

3. The cumulative distribution function of the errors ϵ is absolutely continuous and its density function is even, decreasing in $|x|$, and strictly decreasing in $|x|$ around 0.

4. The process X has finite fourth moments, i.e. $\mathbb{E}[||X||^4] < \infty$.

The first three assumptions imply the Fisher-consistency of both upcoming estimator of β. The fourth one is needed to obtain consistency of the S-estimator and M-scale estimator, and, as a consequence, also the L^2 consistency of both final estimators. For the smoothed estimator there is an additional assumption, namely the eigenfunction of the covariance operator C to satisfy $\int_0^1 (\phi_k''(t))^2 dt < \infty$. Having bounded second derivatives which also occur as penalty terms leads to asymptotic equality of the two estimators.

For these estimators we observe realizations of the process at the same time points, i.e. we can write the observations of X as a matrix such that

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \ddots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \dots & X_{Nn} \end{pmatrix}$$

where $X_{ij} = X_i(t_j)$ and X_i denotes the i-th independent realization of the process X.

Estimating the mean function

At first, a robust estimator of the mean function is needed. Here, the authors use a functional M-estimator (3.3) defined by

$$\hat{\mu} = \underset{y \in L^2[0,1]}{\operatorname{argmin}} \sum_{i=1}^{N} \rho\left(||X_i - y||\right).$$

As the loss function ρ the Huber function (3.1) with $k = 0.193$ is used. The absolute loss in L_2 leads to the functional median which serves as the initial estimator in the M-estimation process.

Estimating the eigenbasis

For estimating the functional principal components, a projection pursuit approach (2.5) is used. However, the sample variance usually used can not be utilized since it is not robust. It is replaced by the robust Q estimator of scale (3.9). Then on the population level the

eigenfunctions are defined to be

$$\phi_k(t) = \begin{cases} \underset{\{\phi \in L^2[0,1]:||\phi||=1\}}{\operatorname{argmax}} \quad Q(\langle \phi, X \rangle) & k = 1 \\ \underset{\{\phi \in L^2[0,1]:||\phi||=1, \langle \phi, \phi_j \rangle = 0, j<k\}}{\operatorname{argmax}} \quad Q(\langle \phi, X \rangle) & k > 1. \end{cases}$$

Estimating β

Next, the coefficient function β is estimated. Assuming a scalar-response model (2.13) without intercept, we estimate the scores to use in the finite expansion (2.3). That is, we compute

$$\hat{\xi}_{ik} = \langle X_i - \hat{\mu}, \hat{\phi}_k \rangle, \quad i = 1, \ldots, N, \; k = 1, \ldots, K.$$

In practice, these integrals are approximated numerically. Using the estimated scores we form a $(N+1) \times K$ predictor matrix for an MM-estimation as follows.

$$Z = \begin{pmatrix} 1 & \hat{\xi}_{11} & \ldots & \hat{\xi}_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \hat{\xi}_{N1} & \ldots & \hat{\xi}_{NK} \end{pmatrix}$$

The MM-estimator is now the solution to (3.3), that is,

$$\begin{aligned} (\langle \beta, \mathbb{E}[X] \rangle, \langle \beta, \phi_1 \rangle, \ldots, \langle \beta, \phi_K \rangle)' &\approx \left(\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_K \right) \\ &= \underset{b=(b_0,\ldots,b_K)'}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \rho_1 \left(\frac{y_i - z_i' b}{\hat{\sigma}_N} \right). \end{aligned}$$

The corresponding M-scale estimator $\hat{\sigma}_N$, and the inital S-estimator $b^{(\text{in})}$ (3.6) are found simulatenously using

$$\frac{1}{N} \sum_{i=1}^{N} \rho_0 \left(\frac{y_i - z_i' b^{(\text{in})}}{\hat{\sigma}_N} \right) = b, \quad \text{and } b^{(\text{in})} = \underset{b}{\operatorname{argmin}} \, \hat{\sigma}(r(b)), \tag{4.1}$$

where $\rho_1 \leq \rho_0$ and $b = 0.5 \sup_x \rho_0(x)$ to ensure a maximal breakdown point of 0.5 of both robust estimators. The ρ-functions used here are Tukey bisquare functions (3.2) with corresponding $k's$ of $k_0 = 1.548$ and $k_1 = 4.685$ which leads to $\rho_1(x) \leq \rho_0(x)$ for all $x \in \mathbb{R}$.

Finally, an estimate for β is achieved as in (2.15), and corresponding fitted values are computed as in (2.16). Namely, this yields

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{b}_k \hat{\phi}_k(t), \quad \text{and} \quad \hat{y}_i = \hat{b}_0 + \sum_{k=1}^{K} \hat{b}_k \hat{\xi}_{ik}. \tag{4.2}$$

In further analysis we denote this method of estimation and prediction by `RFPCR`.

Smoothing $\hat{\beta}$

However, the authors argue that the resulting β coefficient function can be quite unsmooth in certain cases. Therefore, they proposed a regularization penalty on the second derivatives of the eigenfunctions to overcome this problem. That is, we consider the M-estimation step and add penalties of high second derivatives, i.e.

$$\left(\hat{b}_0^{(p)}, \hat{b}_1^{(p)}, \ldots, \hat{b}_K^{(p)}\right) = \underset{b=(b_0,\ldots,b_K)'}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \rho_1 \left(\frac{y_i - z_i' b}{\hat{\sigma}_n} \right) + \lambda(b_1, \ldots, b_K) A (b_1, \ldots, b_K)',$$

where $A_{ij} = \langle \phi_i'', \phi_j'' \rangle$ denotes the matrix of second derivatives of the eigenfunctions. However, instead of actually minimizing this objective function, we write the solution motivated by the shrunk ridge regression estimator. In ridge regression a penalty on the coefficient estimator is added to ordinary least squares regression in terms of its Euclidean norm in order to penalize large coefficient estimates[1]. Similarly, we obtain with $\tilde{Z}$ equal to Z without the first column

$$\left(\hat{b}_1^{(p)}, \ldots, \hat{b}_K^{(p)}\right)' = \left(\tilde{Z}' \hat{W} \tilde{Z} + \lambda A\right)^{-1} \tilde{Z}' \hat{W} \tilde{Z} \left(\hat{b}_1, \ldots, \hat{b}_K\right)',$$

and $\hat{b}_0^{(p)} = \hat{b}_0$. The weight matrix is a diagonal matrix with $\hat{w}_i = \psi_1(r_i/\hat{s}_N)/r_i/\hat{s}_N$ which are the robustness weights of the MM-estimator. The resulting estimators for β and y can analogously be computed using (2.15),(2.16). The resulting estimator is now called `RFPCPR`.

Selection of K and λ.

The selection of K is done as proposed in Section 4.3. The optimal choice of λ in the penalized case is more difficult. Essentially the authors base their choice of λ on a mixed normal linear model $y|u$ with variance σ^2 and u with variance σ^2/λ. Weighing the corresponding

[1] For centered data (X,y) the ridge estimator is given by $\hat{\beta}_{\text{Ridge}} = (X'X + \lambda I)^{-1} X'X \hat{\beta}_{\text{OLS}}$ where $\lambda \geq 0$ controls the level of penalization. Obviously, for $\lambda = 0$ we have $\beta_{\text{Ridge}} = \beta_{\text{OLS}}$.

likelihoods of the normal distributions with weights obtained from the MM-estimator $\hat{b}$ leads to a robust way of estimating the model's variances. An estimator for λ can then be obtained by dividing the two quantities. For a more detailed explanation see Kalogridis and Van Aelst (2019).

Non-Robust Alternative

To actually see the effect of the robust methods, we also introduce a non-robust alternative which follows the main steps of the estimators described above.

First, the estimation of the mean function can be easily implemented non-robustly by computing the arithmetic mean of observations at time $t_j, j = 1, \ldots, n$, i.e. one has

$$\hat{\mu}(t_j) = \frac{1}{N} \sum_{i=1}^{N} X_{ij}.$$

The following step of performing the projection pursuit (2.5) non-robustly can be executed by maximizing a non-robust measure of variance, i.e. the standard deviation. Finally, the non-robust regression step is done using ordinary least squares. The smoothing step of the estimated coefficient function is quite similar but no weighing is added. In further analysis we will only consider the smoothed estimator denoted by `FPCPR`.

Adaption for Irregular Data

To be able to compare all methods also for irregular data, we propose an adaption step for this kind of data to make it suitable for the regular data method. Given irregular observations for individual i, $X_{i1}, \ldots X_{in_i}$ where n_i is the number of values for this observation, we linearly interpolate the observed data points, and extrapolate in a constant manner. That way we can construct regular observations on which the method above can be applied to. However, depending on the nature of the process this method may turn out to give very unreasonable results, especially if n_i is very small.

4.2. sparseFPCR

These estimators follow a different and rather new approach as the previously discussed `RobustFPCR` methods and can also handle irregularly observed data meaning each curve may have been observed at different time points. Thus, no regressor matrix is available and

basic methods do not work anymore. However, the sparse methods can still be applied to regular data. The following method makes use of a generalization of normally distributed processes, and finds an elegant way to estimate the scores. For detailed discussions and proofs, see Boente and Salibián-Barrera (2021).

Preliminaries

The authors consider stochastic processes X on the interval $[0,1]$ and allow for atypical observations that non square integrable processes are also included (see Section 2.2.3). That is, we let $X \sim \mathcal{E}(\mu, C, \varphi)$ with $\mu \in L^2[0,1]$ and self-adjoint, non-negative definite Hilbert-Schmidt operator C on $L^2[0,1]$ with continuous kernel function c. Boente and Salibián-Barrera (2021), Prop. 3.1 shows that the scores $\xi_k = \langle X - \mu, \phi_k \rangle$ conditioned on a finite set of evaluations of X is again elliptically distributed. Specifically, one obtains for $X_m = (X(t_1), \ldots, X(t_m))'$ the conditional mean of

$$\mathbb{E}[\xi_k | X_m] = \lambda_k \phi_k' \Sigma_{X_m}^{-1} (X_m - \mu_m), \tag{4.3}$$

where $\phi_k = (\phi_k(t_1), \ldots, \phi_k(t_m))'$ and $\mu_m = (\mu(t_1), \ldots, \mu(t_m))'$. The (l,j)-th element of $\Sigma_{\boldsymbol{X}_m}$ is equal to $c(t_l, t_j)$. This expression will turn out quite useful when estimating the scores in the scalar-response model.

Another useful part of this proposition is the conditional mean of $X(t_0)|X(s_0)$, $0 \leq t_0 \neq s_0 \leq 1$ given by

$$\mathbb{E}[X(t_0)|X(s_0)] = \mu(t_0) + \frac{c(t_0, s_0)}{c(s_0, s_0)}(X(s_0) - \mu(s_0)), \tag{4.4}$$

which is very similar to the expression found in multivariate normal models.

In contrast to the estimators in Section 4.1 we do not need to observe every X_i at the same time points. Each curve is rather observed at independent random time points t_{ij}, $1 \leq j \leq n_i$. The number of observations per curve n_i are assumed to be independent random variables as well. The corresponding observed model then reads

$$X_{ij} = X_i(t_{ij}) = \mu(t_{ij}) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t_{ij}), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, N.$$

Estimating the mean function

As done previously, the first step is estimating the mean function of the process. The authors use a local M-estimator as in (3.7). Specifically, for each $t_0 \in [0,1]$ we have

$$\left(\hat{\beta}_0(t_0), \hat{\beta}_1(t_0)\right)' = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{n_i} w_{ij}(t_0) \rho_1 \left(\frac{X_{ij} - \beta_0 - \beta_1(t_0 - t_{ij})}{\hat{\sigma}(t_0)} \right),$$

with weights $w_{ij}(t_0)$ given by

$$w_{ij}(t_0) = \frac{K\left(\frac{t_{ij}-t_0}{h}\right)}{\sum_{k=1}^{N} \sum_{l=1}^{n_k} K\left(\frac{t_{kl}-t_0}{h}\right)}.$$

As a consistent robust estimator of scale the local MAD (3.8) is chosen, i.e.

$$\hat{\sigma}(t_0) = c \underset{|t_{ij}-t_0|\leq h}{\operatorname{med}} \left| X_{ij} - \underset{|t_{ij}-t_0|\leq h}{\operatorname{med}} X_{ij} \right|,$$

where $c^{-1} = \Phi_{N(0,1)}(3/4)$ to ensure Fisher consistency under the normal model. The authors use a Huber ρ-function (3.1) with parameter $k = 1.345$.

The resulting robust estimator for the mean function is then given by $\hat{\mu}(t_0) = \hat{\beta}_0(t_0)$.

Estimating the eigenbasis

Instead of using a projection-pursuit approach, the authors estimate the covariance operator via an eigendecomposition (2.4). First, the diagonal elements are estimated. An M-scale (3.6) is used whereby the ρ-function satisfies $\sup_x \rho_2(x) = 1$ and is bounded. Then $\hat{c}(t_0, t_0)$ is the solution to

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} w_{ij}(t_0) \rho_2 \left(\frac{X_{ij} - \hat{\mu}(t_{ij})}{\hat{c}(t_0, t_0)} \right) = b, \tag{4.5}$$

where $b = 1/2$ to ensure consistency and maximum breakdown point in case of the Gaussian model. The ρ-function used here is a Tukey bisquare function (3.2) with $k = 1.548$ which naturally satisfies the condition of boundedness.

Estimating the off-diagonal elements is more difficult. However, (4.4) will turn out to

be useful here, namely that the slope of the conditional expectation is proportional to the sought $\hat{c}(t_0, s_0)$. This equation suggests to first center the observations to obtain $\tilde{X}_{ij} = X_{ij} - \hat{\mu}(t_{ij})$. As before, a local M-estimation is done yielding an estimator for the slope $a(t_0, s_0) = c(t_0, s_0)/c(s_0, s_0)$. Namely, we have

$$\hat{a}(t_0, s_0) = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j \neq l} \rho \left(\frac{\tilde{X}_{ij} - a\tilde{X}_{il}}{\hat{s}(t_0, s_0)} \right) K \left(\frac{t_{ij} - t_0}{h} \right) K \left(\frac{t_{il} - s_0}{h} \right),$$

where the robust scale estimator $\hat{s}$ is obtained by a local MAD of the residuals. The authors propose to use a local median $\tilde{a}(t_0, s_0)$ of the slopes $Z_{ijl} = \tilde{X}_{ij}/\tilde{X}_{il}$ as the initial robust estimator to calculate residuals from. Thus, we have residuals of $r_i = \tilde{X}_{ij} - \tilde{a}(t_0, s_0)\tilde{X}_{il}$ where $\tilde{a}(t_0, s_0) = \operatorname{med}_{|t_{ij}-t_0|<h, |t_{il}-s_0|<h} Z_{ijl}$. Then the local MAD of the residuals is given by

$$\hat{s}(t_0, s_0) = \underset{|t_{ij}-t_0|<h, |t_{il}-s_0|<h}{\operatorname{med}} \left| r_i(t_0, s_0) - \underset{|t_{ij}-t_0|<h, |t_{il}-s_0|<h}{\operatorname{med}} r_i(t_0, s_0) \right|.$$

Having estimated this slope we can now compute the off-diagonal elements by $\tilde{c}(t_0, s_0) = \hat{a}(t_0, s_0)\hat{c}(s_0, s_0)$ where $\hat{c}(s_0, s_0)$ is the solution to (4.5). However, there is no guarantee of $\hat{c}$ being symmetric and smooth. Hence, the authors also implement a two-dimensional smoothing step to obtain $\tilde{\tilde{c}}(t_0, s_0), \tilde{\tilde{c}}(s_0, t_0)$. The final estimation for the covariance operator is given by

$$\hat{c}(t_0, s_0) = \frac{\tilde{\tilde{c}}(t_0, s_0) + \tilde{\tilde{c}}(s_0, t_0)}{2}.$$

This estimator can now be used to calculate approximate eigenfunctions and eigenvalues which are needed in the next step. The ρ-function in this step corresponds also to a Tukey bisquare function, however, with a higher tuning constant in $k = 3.444$ to achieve higher efficiency which is desired in two-dimensional problems.

Estimating β

We use the same estimation step as in (4.1) with the difference of how the scores are estimated. Since we assume the process to be elliptically distributed we saw in (4.3) how the best linear predictor of the scores ξ_k looks like. Having estimated all unknown quantities

we obtain a robust estimator for the scores given by

$$\hat{\xi}_{ik} = \hat{\lambda}_k \hat{\phi}'_{ik} \left(\hat{\Sigma}_i + \delta I_{n_i} \right)^{-1} (X_m - \hat{\mu}_i),$$

where $\hat{\boldsymbol{\phi}}_{ik} = (\hat{\phi}_k(t_{i1}), \ldots, \hat{\phi}_k(t_{in_i}))$, and $\hat{\mu}_i = (\hat{\mu}(t_{i1}), \ldots, \hat{\mu}(t_{in_i}))$. The matrix $\hat{\Sigma}_i$ contains the entries of $\hat{c}$ belonging to observations i, i.e. the (j,l)-th element is equal to $\hat{c}(t_{ij}, t_{il})$. The parameter $\delta > 0$ is used to ensure regularity of the matrix to be inverted and is usually set to a small positive number. Given the estimated scores we can now estimate the response and the coefficient function as in (4.2). The method is denoted as `S-ROB-ROB` in further analysis.

Non-Robust Alternative

As before, we also introduce some non-robust alternative methods to compare the robust methods' performances. The algorithm described above essentially contains two parts: the FPCA part and the regression step. The FPCA step can be easily done in a non-robust fashion by setting the corresponding ρ-functions to $x \mapsto x^2$. As discussed previously, the regression step using ordinary least squares is an easy way to make this step non-robust.

To see more clearly the effects of robustness in the two steps, we will differ between `S-LS-LS` where both steps are carried out non-robustly, `S-ROB-LS` where the FPCA step is realized robustly as described above, and `S-LS-ROB` which only uses a robust regression step.

4.3. General Tuning

The regression based on functional principal components heavily depends on the choice of the number of components K. One way to choose a sensible K is by using cross-validation. To robustify this approach for the use in completely robust methods we use a robust measure of the CV errors, namely the τ-scale estimator (3.10) as in Kalogridis and Van Aelst (2019). As for cross-validation, we use a leave-one-out (LOO) approach since the MM-estimator is a linear estimator with suited weight matrix. This is useful because the LOO residuals can be calculated without needing to fit any more models using all but one observation. Let $\hat{y}_{-i}$ denote the fitted values obtained when leaving out (X_i, y_i) and r_{-i} the respective residuals. Then these residuals can be approximated[2] by

[2]The residuals can only be approximated since the weights depend on the response y.

$$r_{-i} = y_i - \hat{y}_{-i} \approx \frac{r_i}{1 - h_{ii}}, \tag{4.6}$$

where h_{ii} are the diagonal elements of the hat matrix $H = Z(Z'WZ)^{-1}Z'W$ as in $\hat{y} = Z\hat{\beta} = Z(Z'WZ)^{-1}Z'WY = HY$ where Z is the matrix consisting of a column of ones and the scores (see 4.1). We compute $\tau(r_{-1}, \ldots, r_{-N})$ for $K = 1, \ldots, K_{\max}$. In contrast to Kalogridis and Van Aelst (2019) we do not choose K which minimizes the CV errors because we found that it tends to overfit on the number of components actually needed. We rather use the *one standard error rule* and select the minimal K that is one standard error away from the global minimum. This is common technique also used in penalized regression settings like LASSO, see Hastie et al. (2009).

In case of non-robust regression we calculate a usual MSE of the LOO residuals. This does not require an approximation of the residuals since (4.6) holds strictly. As before we use the one standard error rule to obtain the optimal number of components.

5. Simulation Study

In a simulation study we want to evaluate the methods' performances in terms of prediction and estimation. We focus on a setting where data is not often observed. Moreover, we differ between regularly observed data where all individuals have been observed at the same time points, and irregularly observed data meaning individuals may not be observed at the same time points. Following scenarios for the model $y = \langle X, \beta \rangle + \sigma\varepsilon$, $\varepsilon \sim N(0, 1)$ are considered. These have also been used in Kalogridis and Van Aelst (2019) and Febrero-Bande et al. (2017). For both scenarios we take X to be the Wiener process which has a Karhunen-Loéve representation given below.

- Model 1: $X(t) = \sum_{k=1}^{\infty} \xi_k \phi_k(t)$ where $\xi_k \sim N(0, \lambda_k)$ are independent with eigenfunctions $\phi_k(t) = \sqrt{2}\sin((k - 0.5)\pi t)$ and eigenvalues $\lambda_i = k/((k - 0.5)t)^2$. As for β we take $\beta(t) = 2\sin(0.5\pi t) + 4\sin(1.5\pi t) + 5\sin(2.5\pi t)$. This β can be written as the linear combination of the first three eigenfunctions. Note that for numerical aspects we consider only the first 50 parts of the sum.

- Model 2: We take the same explaining process in X but for β we choose $\beta(t) = \log(1.5t^2 + 10) + \cos(4\pi t)$. This coefficient function can only be expressed using an infinite number of eigenfunctions.

For each model we consider various settings. We apply contamination to the observed data. For that we select $100\epsilon\%$ of the training observations and multiply the explaining curve by 2 and the corresponding responses by 3 to create significant outliers in X and y. We choose $\epsilon \in \{0, 0.1, 0.2\}$. Figure 5.1 shows trajectories of the Wiener process with contaminated curves depicted in red. The signal-to-noise ratio σ was set to $\sigma = 0.1$.

As for observing the data, we randomly sample $N = 200$ curves in the interval $[0, 1]$. The first 100 observations are for training the models, while the remaining 100 clean observations are solely for evaluation purposes. These are observed at time points given as follows.

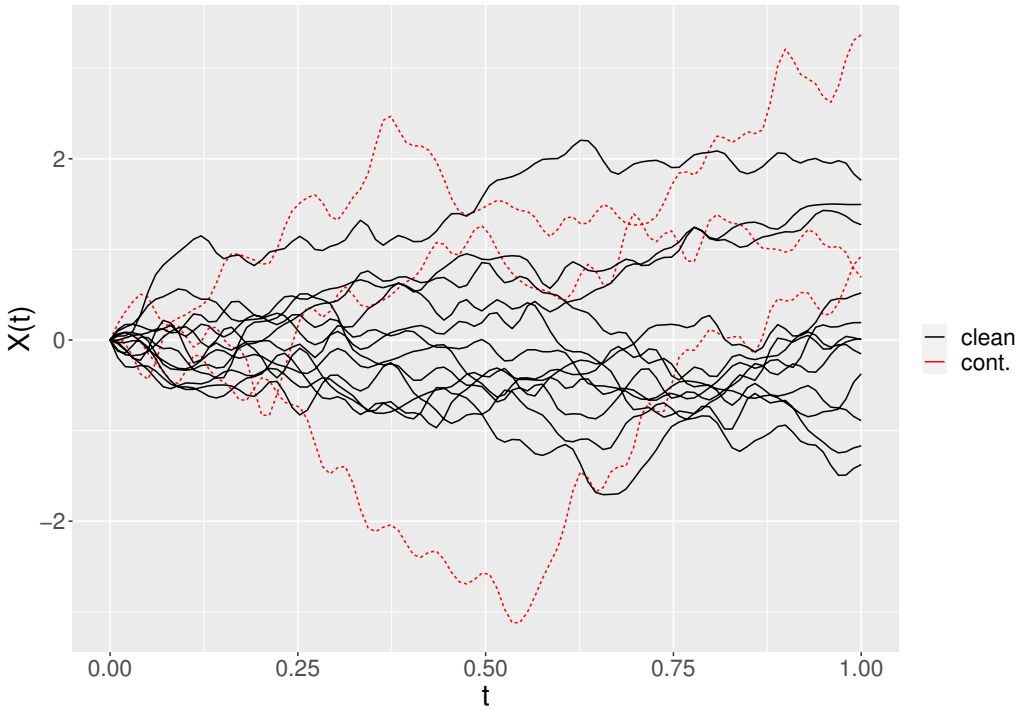

Figure 5.1.: Clean and Contaminated Curves of the Wiener Process

- Regularly observed: We observe at equally spaced time points $t_1, \dots t_n$, $n \in \{20, 50, 100\}$ for each curve X_i, $i = 1, \dots, N$.
- Irregularly observed: We sample $3, 4, 5$ or $10, 11, \dots, 20$ time points from $t_1, \dots, t_{100}$ for each curve X_i. Further, we also differ the method of sampling here. While in non-aggressive sampling we do not care if observations are sampled very close to each other, in aggressive sampling we want the sparse observations to be rather far apart. Both sampling types represent common use-cases.

Last but not least, we set the maximum number of components to be considered in the estimation to be $K = 8$. This should work rather well in the first model since we only need 3 eigenfunctions for reconstructing β but gives an unavoidable error in Model 2. In total $m = 30$ replications were run. This rather low number is due to runtime aspects.

Following measures are introduced for evaluation. To assess the quality of the fits in terms

of estimation, we introduce a relative mean squared error, that is,

$$\text{Rel. MSE} = \frac{\int_0^1 (\hat{\beta}(t) - \beta(t))^2 dt}{\int_0^1 \beta(t)^2 dt} \approx \frac{||\hat{\beta} - \beta||_2^2}{||\beta||_2^2}.$$

The quality of predictions are evaluated by a mean squared prediction error, that is,

$$\text{MSPE} = \sum_{i=N/2+1}^{N} (\hat{y}_i - y_i)^2.$$

5.1. Discussion of Model 1

For the first model where the coefficient function can be expressed using just 3 basis functions, we fix the number of components to be used at 3 to remove the effect of the choice of components. However, in both regular and irregular case, Tables A.1 and A.2 in the appendix show the mean number of components chosen with corresponding standard deviations in parenthesis. These numbers suggest that all methods estimate the correct number of components quite accurately. Hence, we only report the cross-validated results in further analysis. For all following boxplots, outliers (about 10% of the data) are not displayed because some of them are extreme and would distort the overall picture.

Regular Case

At first, we look at the estimation aspect of the simulations, i.e. we compare the estimated coefficient functions with the true one. Figure 5.1.1 shows boxplots of the relative MSE on a log-scale depending on the contamination level (x-axis) and the number of observations per curve (y-axis).

What is seen immediately, is the effect of contamination on the non-robust methods (`FPCPR`, `S-LS-LS`, and partly `S-ROB-LS`). As expected, these methods result in larger errors once contamination is added to the model. Another effect is also visible, namely the number of observations per curve. While in a low-observed setting ($n = 20$) the methods `FPCPR` and `RFPCPR` perform worse than the other methods based on conditional expectations but once this number is increased, we see the opposite trend in the S-methods performing worse compared to the methods based on numerical integration. This is due to numerical integration which works better once the approximation grid is dense enough. In terms of the S-methods, we observe that the robust regression part seems to be more important than

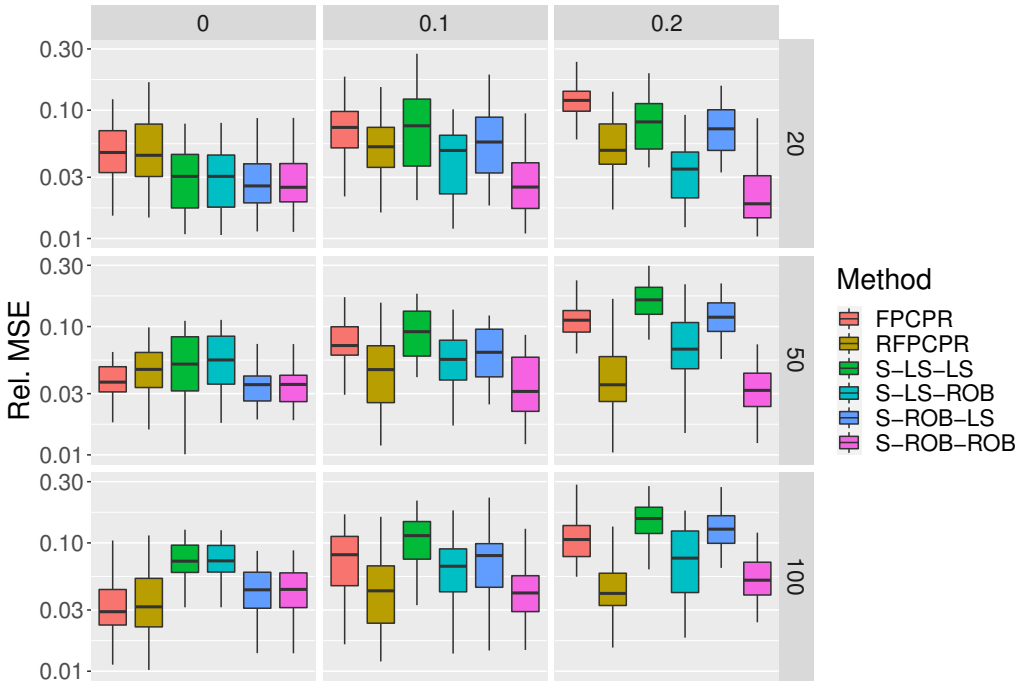

Figure 5.1.1.: Model 1 - Regular: Boxplots of relative MSEs

estimating the functional components robustly (see `S-LS-ROB`, `S-ROB-LS`, `S-ROB-ROB`). A reason might be that the robust regression can handle very well possible outliers in the scores created by the FPCA part of the algorithm. Overall the boxplots suggests the best method to be `S-ROB-ROB` for a lower number of observations, while for observations observed at a higher number of time points, `RFPCPR` seems to be outperforming.

Figure 5.1.2 shows the estimated coefficient functions in terms of a 20%-trimmed mean over all simulations. While for no contamination all methods yield a useful approximation, there are significant differences when contamination is present. The methods based on robust regression (`RFPCPR`, `S-LS-ROB`, `S-ROB-ROB`) cope well with the unusual observations but the remaining methods seem to have the same problem. When comparing to the true coefficient function, these non-robust methods can not approximate well the first peak whereas the errors are less crucial on the second peak.

Boxplots displaying MSPEs of all settings in Model 1 can be seen in Figure 5.1.3. We do see that in terms of prediction the S-methods outperform the methods `FPCPR` and `RFPCPR`,

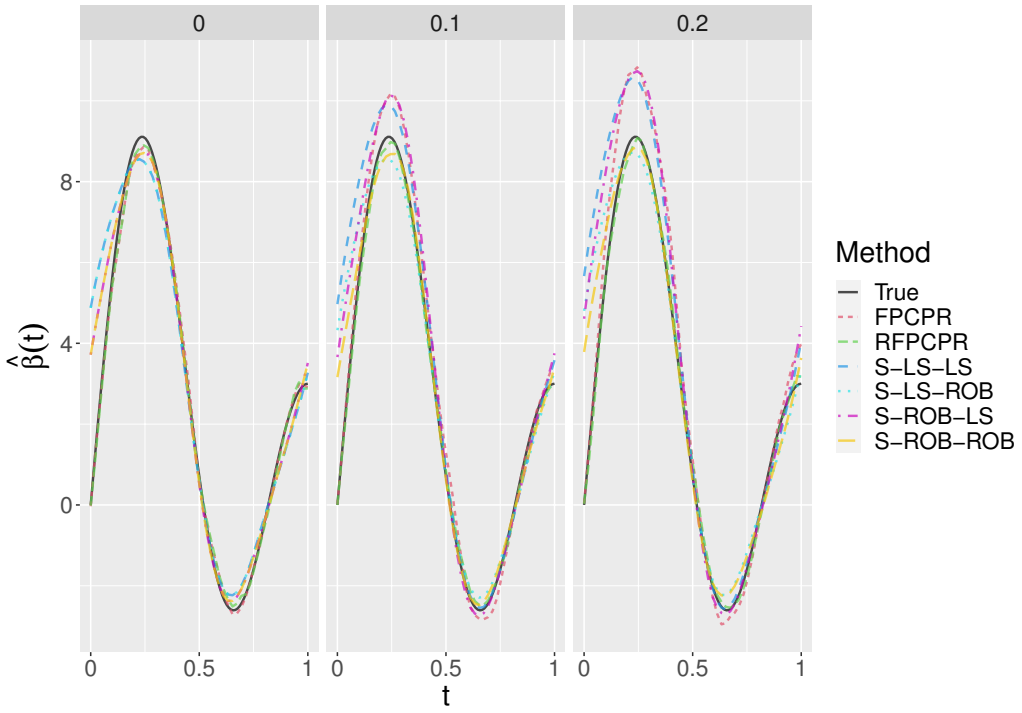

Figure 5.1.2.: Model 1 - Regular: Estimated Coefficient Functions

also in the cases of a high number of observations per curve. As before, the general effect of contamination is very well observable. Whilst in the no-contamination settings all methods seem to perform quite equally, the difference becomes more and more extreme as the contamination level is increased. Overall, the best method seems to be `S-ROB-ROB` again because `RFPCPR` does not seem to profit a lot from the higher number of observations. It is to be noted that `S-LS-ROB` appears to perform quite equally compared to `S-ROB-ROB` with minor distinctions especially in the non-contaminated cases.

Finally, a plot showing fitted values vs. true values can be seen in Figure 5.1.4. A very common effect of bad leverage points can be seen here, namely the tilting of the line. The robust methods show a stable behaviour that most points lie more or less on the identity line. This is true for all contamination levels. For the three methods without a robust regression step (indicating by red), the true values are systematically not estimated correctly anymore once contamination is added.

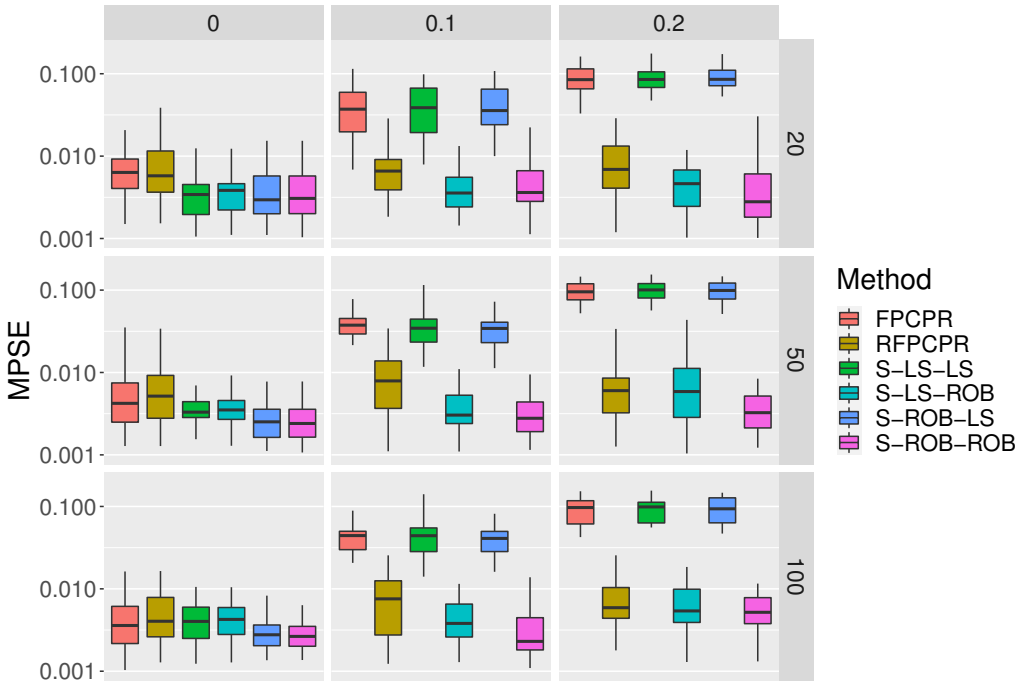

Figure 5.1.3.: Model 1 - Regular: Boxplots of MSPEs

Irregular Case

In the **non-aggressive** setting, the estimation errors are seen in the boxplots in Figure 5.1.5. In the case where 3 to 5 observations are available per curve, the estimation errors are rather high compared to the regular case. This may be due to having too little information to be able to estimate well the coefficient function. The effect of contamination is still present, however it is not as significant as before. `RFPCPR` does not seem to improve the results by a vast amount compared to its non-robust counterpart `FPCPR`. This effect is more visible when looking at the S-methods. Here a clear trend is visible in the robust methods outperforming their non-robust alternatives. Still, all S-methods show high variability indicating a rather unstable estimation process. In terms of the median estimates the S-methods perform worse than the numerical integration methods. However, a lot of estimates show a lower error. A reason might be the inter- and extrapolation needed for methods `(R)FPCPR` which makes these methods more stable but does not recreate the true trajectory well. Overall, one can argue that this setting is not really suited for any estimator used in this analysis, and different approaches should be looked at.

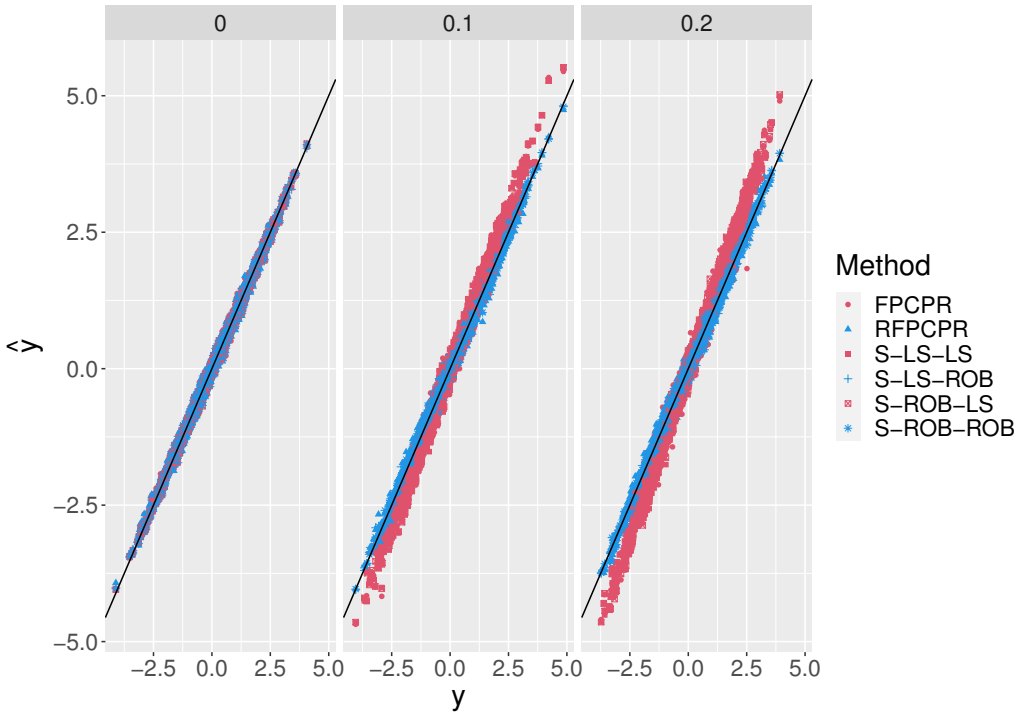

Figure 5.1.4.: Model 1 - Regular: Fitted Values vs. True Values

When increasing the number of observations to 10 to 20 we see a similar effect as in the regular case. Still, the contamination does not affect the estimators as bad as in the regular case. We do observe best estimates by `S-ROB-ROB` and `RFPCPR`. Figure 5.1.6 shows the 20%-trimmed mean of the estimates for the $3-5$ setting. First of all, not a single method is able to approximate the true coefficient function well. Especially, the first peak is not approximated accurately. It also appears that this peak is shifted to the right instead of just having too little amplitude. This effect is seen for every contamination setting. As also suggested by the boxplots in Figure 5.1.5 no clear effect of contamination is present. The estimates remain inadequate. As stated before, a reason might be the really low number of observations per curve. The estimates for $10-20$ observations per curve look similar to the regular case and are not displayed here. For completeness, the corresponding figure is available in Figure B.1.

In terms of predictions it looks somewhat different. Figure 5.1.7 shows the boxplots of the mean squared prediction errors in the non-aggressive setting. While it seems like methods `(R)FPCPR` beat the S-methods in the $3-5$ setting, the S-methods heavily outperform these

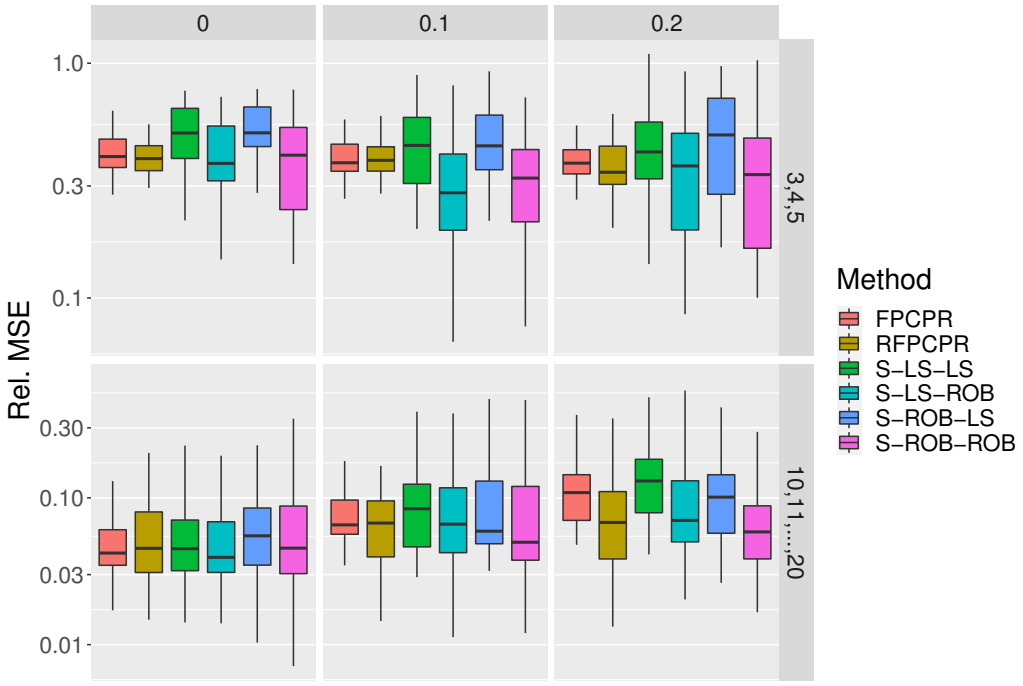

Figure 5.1.5.: Model 1 - Irregular, Non-Aggressive: Boxplots of relative MSEs

methods once enough observations are available. This is mainly due to the imputation needed which introduces unavoidable errors in the estimation process. One has to note that all errors are rather high questioning the viability of these estimations. However, only in the $10 - 20$ setting a clear effect of contamination is present. In the $3 - 5$ setting the contamination effect is seen for all types of estimators. Best results are yielded by `S-LS-ROB` and `S-ROB-ROB` which perform quite equally.

The fitted values vs. true values plot in Figure 5.1.8 shows the badness of the predictions. While in the regular case, we see clear lines, this is not the case here anymore. All methods appear to have problems in predicting the correct values and no clear distinction between robust and non-robust methods is present. However, we do see a denser cloud in the highest contamination level for robust methods (in blue).

The case of $10 - 20$ observations per curve yields a better result as suggested by the boxplots. While there are not as clear lines visible as in the regular case, clear trends are present. Also the tilting can be seen just as in the regular case. The corresponding plot is

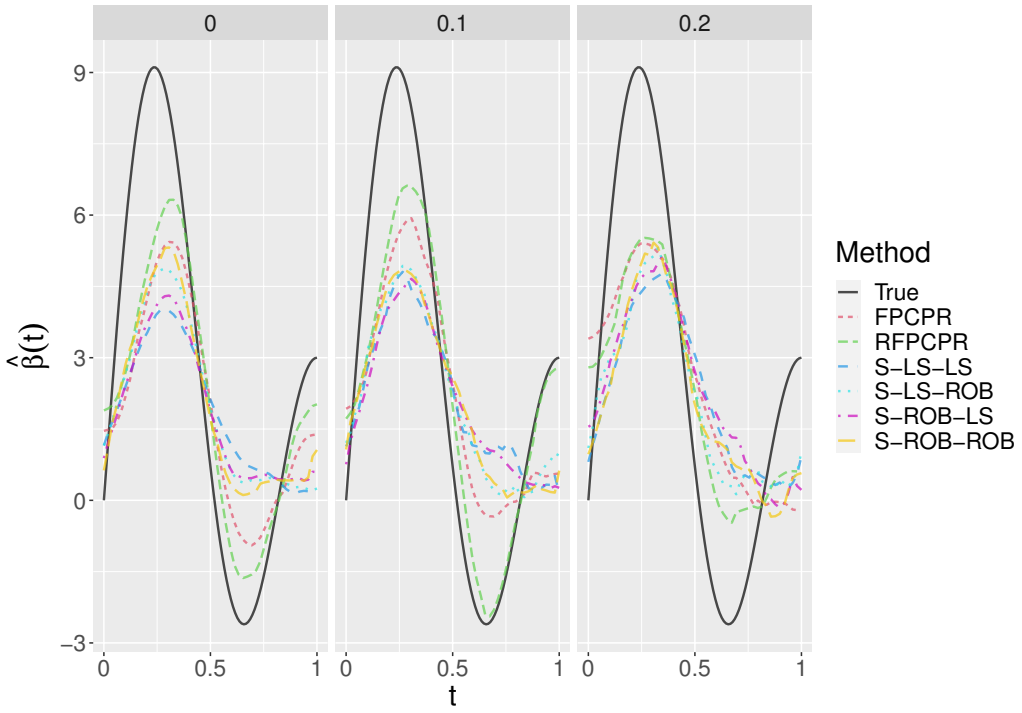

Figure 5.1.6.: Model 1 - Irregular, Non-Aggressive, $3-5$ observations: Estimated Coefficient Functions

available in Figure B.2 in the appendix.

The **aggressive** setting does not show many differences. In Figure B.3 we can see the boxplots of the estimation errors which only show minor distinctions for the $3-5$ setting. The methods `(R)FPCPR` seem to yield better results since the imputation is more sensible if the observed data points are more far away for each other. However, the performances of the S-methods stay more or less the same, some even decrease. Especially, the `S-ROB-ROB` method has problems compared to the non-aggressive setting. In terms of estimated coefficient functions, the aggressive setting for $3-5$ observations yields very similar results as seen in Figure B.4. For $10-20$ observations the results are even more similar to the non-aggressive case and are hence not reported at all.

Figure B.5 in the appendix shows the prediction errors in the aggressive setting. Again, we only observe differences for $3-5$ observations. The major differences are that the methods seem to be closer to each other in terms of predictions. However, the `(R)FPCPR` methods still outperform the S-methods. The fitted values vs. true values plots for settings $3-5$

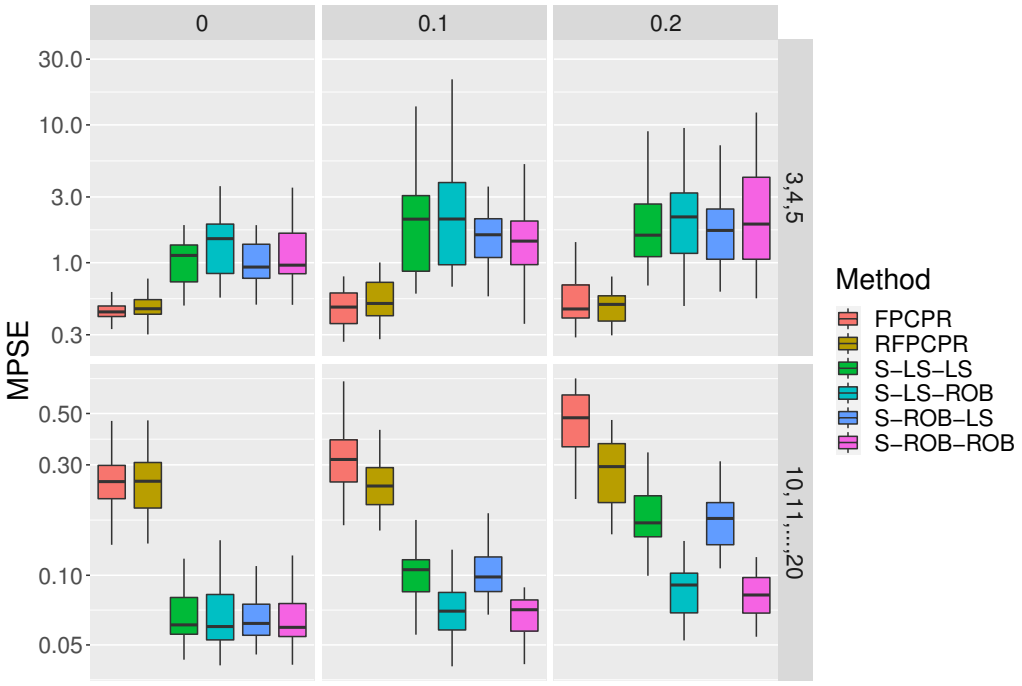

Figure 5.1.7.: Model 1 - Irregular, Non-Aggressive: Boxplots of MSPEs

and $10-20$ observations look very much alike the corresponding plots in the non-aggressive case and are therefore not reported in this analysis.

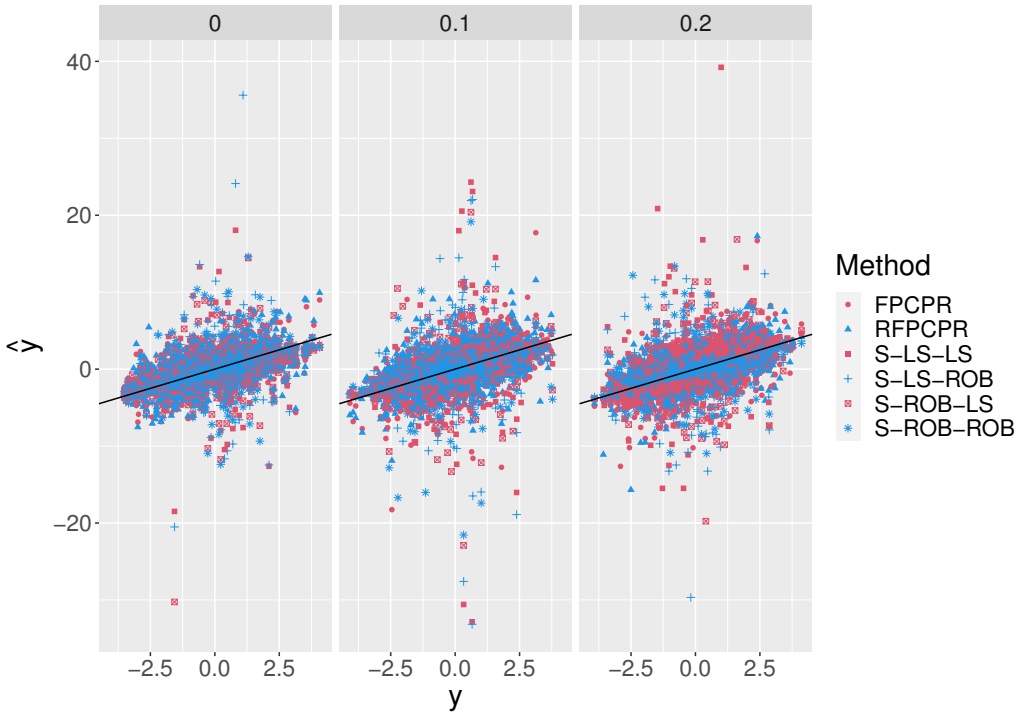

Figure 5.1.8.: Model 1 - Irregular, Non-Aggressive, $3-5$ observations: Fitted Values vs. True Values

5.2. Discussion of Model 2

For model 2 we only report cross-validated results since there is no exact number of components to choose from. We will see that the estimated number of components will vary quite a lot which also implies quite different estimations.

Regular Case

In terms of number of components chosen, the different methods are quite different. Table A.3 shows the mean number of components chosen as well as the corresponding standard deviation. Based on these numbers 3 to 5 components are chosen on average, yet no clear trend is visible. While in clean setting, increasing the number of observations of each curve leads to more components chosen for methods `(R)FPCPR`, it is the opposite for the S-methods where the number of components more or less decreases. Once contamination is added, the methods vary a lot.

Based on the robust methods `RFPCPR`, `S-LS-ROB` and `S-ROB-ROB` the proper number of com-

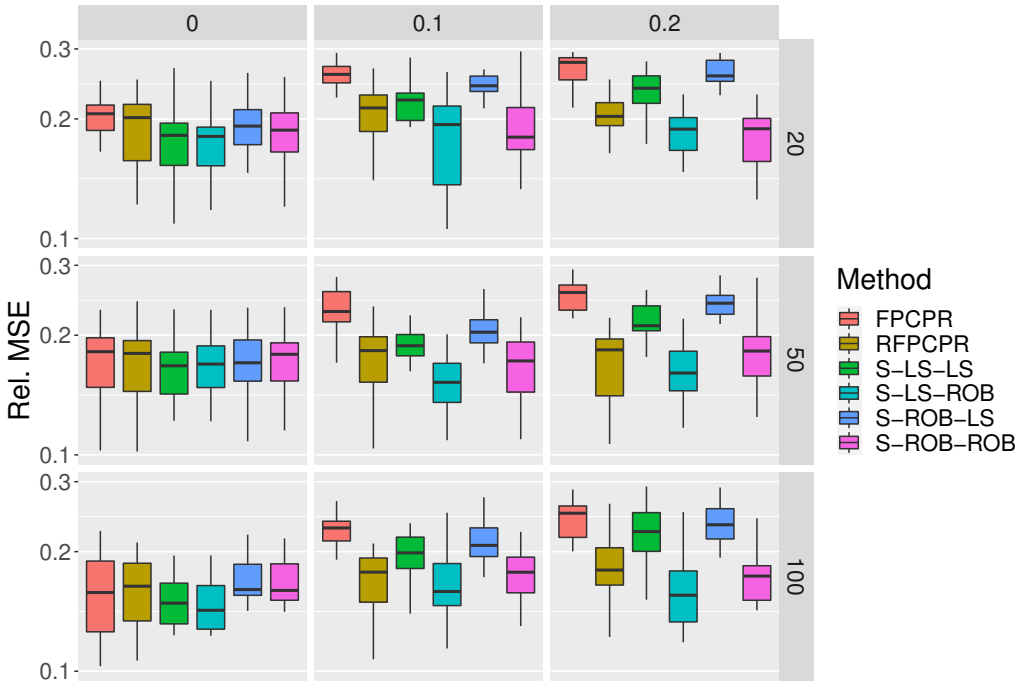

Figure 5.2.1.: Model 2 - Regular: Boxplots of relative MSEs

ponents seems to be 4. Due to the high variability in the chosen number of components, the effects of observation size and contamination level can not be exactly matched anymore as was possible in model 1 where all methods estimated the number of components very equally.

First, we take a look at the boxplots of the estimation error seen in Figure 5.2.1. Looking at the scales of the errors, we observe much higher relative errors than we have in model 1. This is to be expected since the corresponding coefficient function can not be exactly expressed using a finite number of functional principal components and hence an unavoidable error is present. However, we still see a similar behaviour in terms of contamination and observed sample points of the curves. In the clean setting all methods perform quite evenly - there are only minor differences in the cases of $n = 20, 100$. Once contamination is present, we observe higher estimation errors for methods without a robust regression step.

Nevertheless, the `S-LS-LS` method seems to be the most stable non-robust method. In the low-observed setting we see that the S-methods seem to outperform the `(R)FPCPR` methods

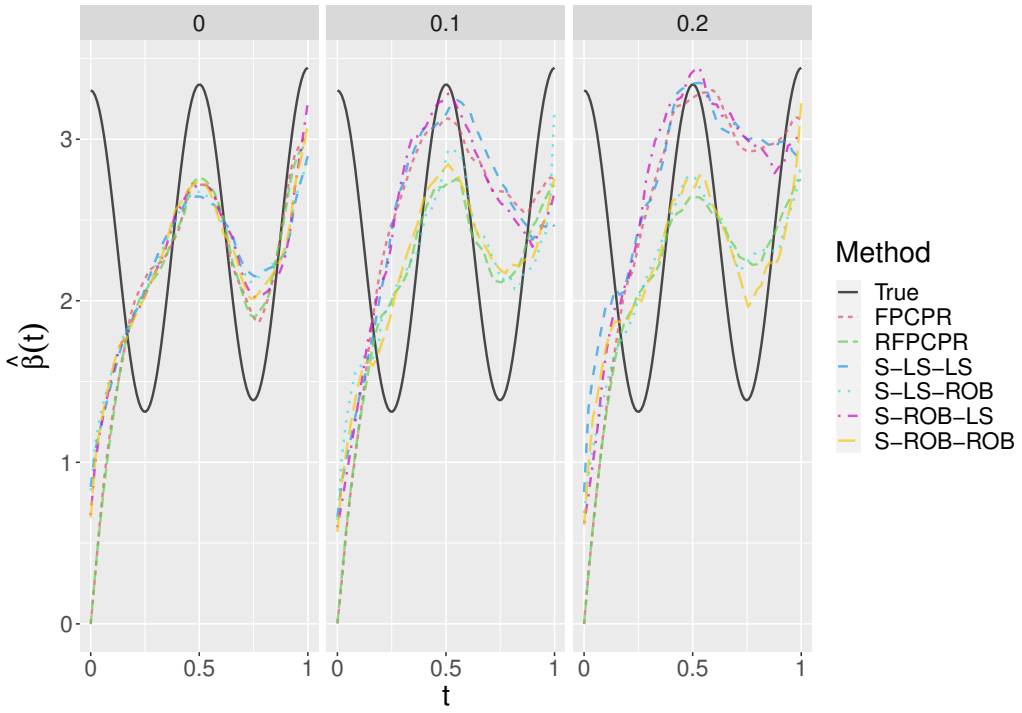

Figure 5.2.2.: Model 2 - Regular: Estimated Coefficient Functions

since the observed grid is not dense enough for these methods based on numerical integration. This effect is not as prominent anymore once the observation size is increased but the S-methods still appear to be at least on the same level as the numerical integration methods in contrast to the first model.

Figure 5.2.2 shows the 20%-trimmed means of each estimator. We observe bad approximation at the beginning of the interval, however, the main oscillation of the true coefficient function after that is approximated quite well. As expected, the robust methods remain stable with contamination and only differ by minor amounts. The non-robust methods are contaminated which is especially seen at the second and third peak of the function. While the second peak is randomly approximated better now, the third peak's approximation is way off.

In Figure 5.2.3 the errors in terms of prediction are displayed in boxplots. Again this plot looks similar to the one of Model 1. Contamination affects the non-robust methods immensely whereas all methods perform evenly on clean data. All methods with a robust

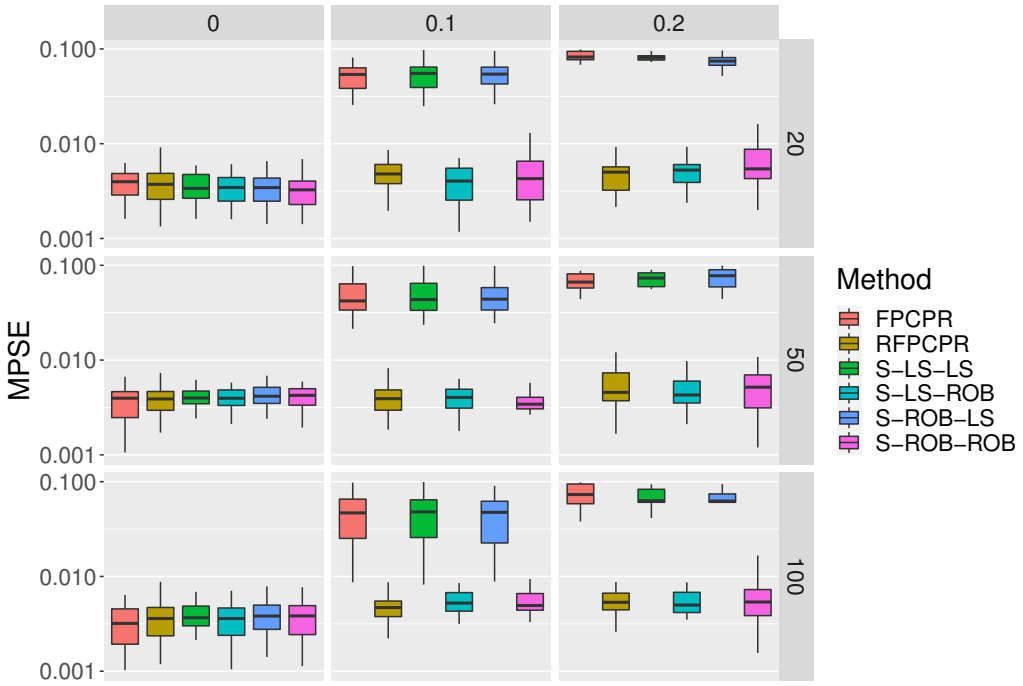

Figure 5.2.3.: Model 2 - Regular: Boxplots of MSPEs

regression step seem to yield similar results indicating the importance of the regression being robust compared to non-robust regression combined with robust FPCA, see the high errors of `S-ROB-LS`. This might be due to numerical instability of the robust FPCA which might output quite unusual scores which, in turn, heavily influence the followed regression. In terms of actual fitted values (Figure 5.2.4) we observe a similar picture as for Model 1. Once contamination is present, the line is tilted due to bad leverage points.

Irregular Case

In the irregularly observed, **non-aggressive** setting, we see quite different results. First, the number of estimated components does differ. Table A.4 shows the corresponding means and standard deviations. While in the regular case the numbers varied between 3 and 5 we see some estimated numbers smaller than 3, especially for the S-methods. Otherwise the effect of increased sample size per curve is quite similar to the regular case. Overall, we can say that methods with robust regression estimate the number of components to be larger than their non-robust alternatives. Additionally, the methods based on numerical integration `(R)FPCPR` also estimate a larger number of components compared to the S-methods.

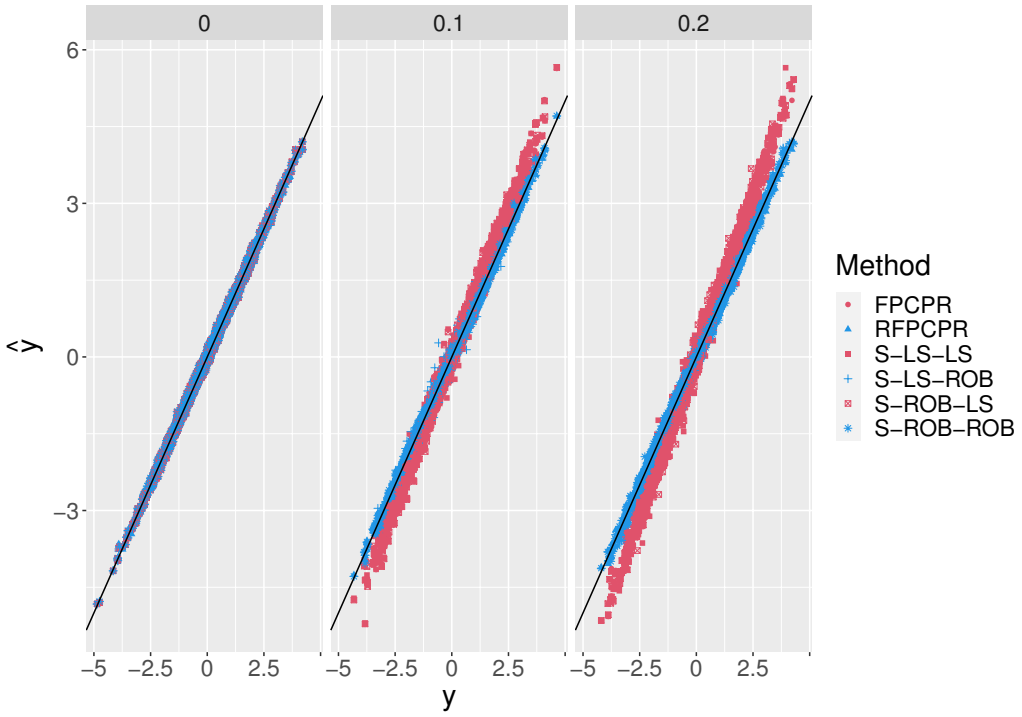

Figure 5.2.4.: Model 2 - Regular: Fitted Values vs. True Values

Figure 5.2.5 shows boxplots of the relative MSEs of each setting and estimator. As for Model 1, having only $3-5$ observations, non-aggressively sampled does not seem to be sufficient to properly estimate the true coefficient function. Once the observation size is increased to $10-20$ observations per curve we observe a similar behaviour as in the regular case. What is quite interesting, is the fact that the non-robust `FPCPR` estimator yields non-worse results in terms of median relative MSE compared to the robust S-methods. This might be due to the S-methods being very numerically unstable resulting in higher estimation errors.

In the low-observed case we can see in Figure 5.2.6 the reasons for that behaviour. It seems like the second peak is somewhat approximated. Especially the S-methods show weird behaviour due to numerical issues (see `S-ROB-ROB` in the highest contamination setting). As for $10-20$ observations, the coefficient functions look pretty similar to the regular case with just minor distinctions at the end of the interval where in the irregular case the methods can not approximate the last inclination at all (see Figure B.6).

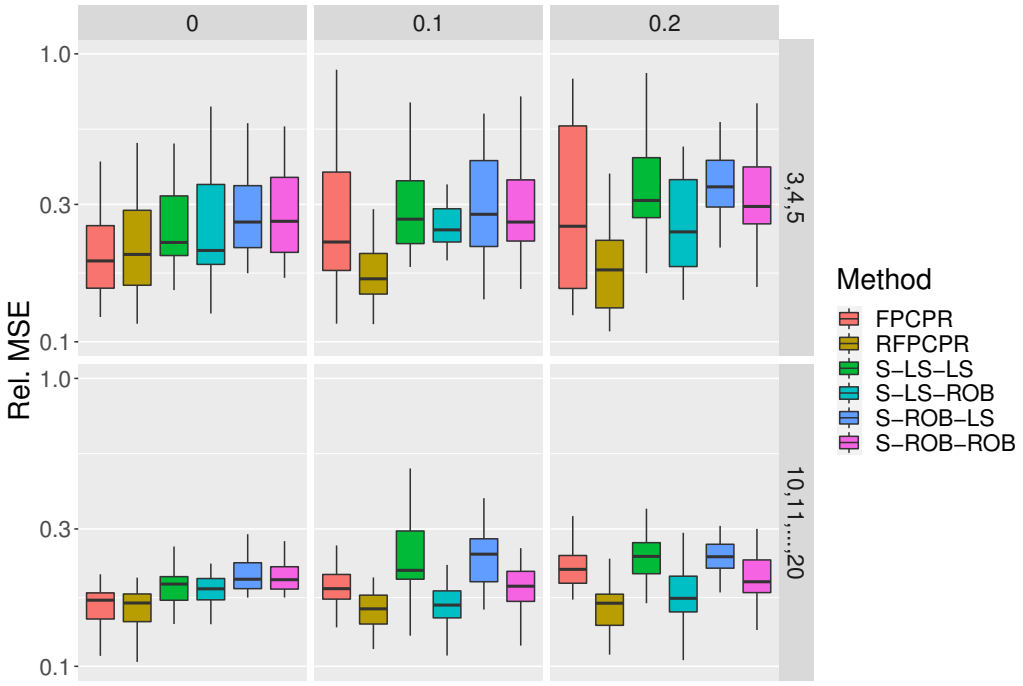

Figure 5.2.5.: Model 2 - Irregular, Non-Aggressive: Boxplots of relative MSEs

Prediction errors are seen in Figure 5.2.7. Just as in Model 1 we observe high errors for $3-5$ observations where the S-methods do not seem to work sufficiently. Here, the detour using inter- and extrapolation appears to be better even though this approximation in the non-aggressive setting where observations might be very close to each other, does not always make sense. However, once the number of observations is increased, we see a clear trend in the S-methods performing way better than the methods `(R)FPCPR`. This is true for both clean and contaminated data. Again, we observe the `S-LS-ROB` estimator to be somewhat better than the fully robust `S-ROB-ROB` estimator which shows higher variability and a higher median error. This might be due to numerical issues in the robust FPCA estimation step.

In the appendix we can see the plot of fitted values vs. true values for this case (see Figure B.7). As suggested by the boxplots, the predictions are quite bad, and all methods seem to have problems in correctly estimating the true response values. Additionally, we only observe a slight improvement when using robust methods in the contaminated settings. As for $10-20$ observations, the plot looks much better and is already quite similar to the

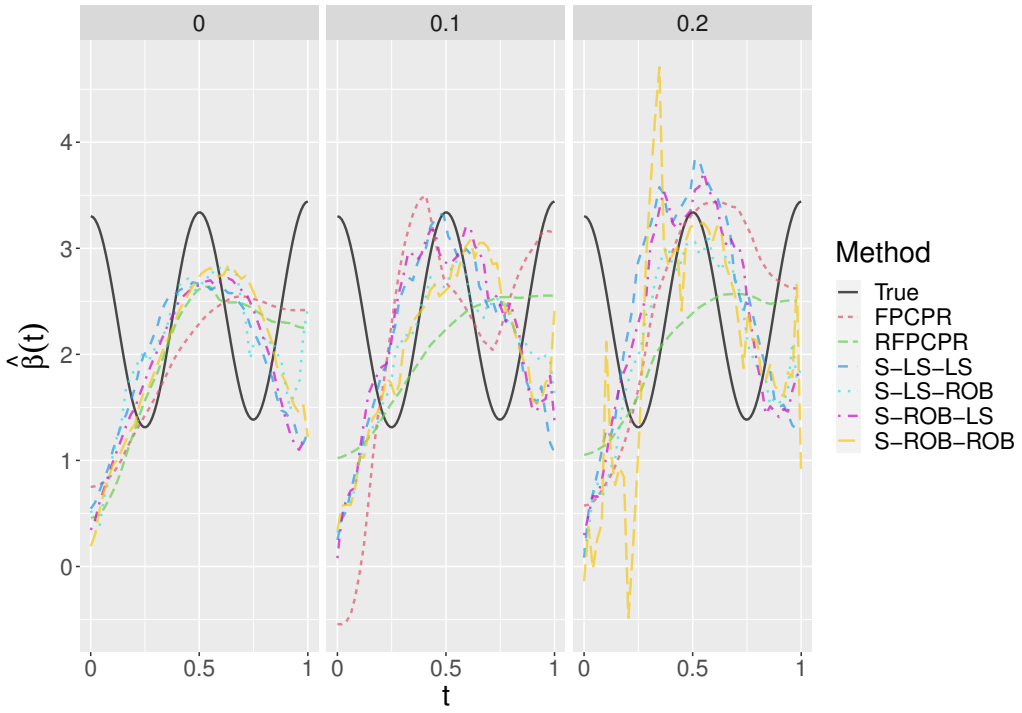

Figure 5.2.6.: Model 2 - Irregular, Non-Aggressive, $3-5$ Observations: Estimated Coefficient Functions

regular case, hence it is also not reported in this analysis.

In the **aggressive** setting, the boxplots of the estimation errors does not differ a lot from the non-aggressive setting. Minor differences are marginally lower errors for the S-methods as well as the `FPCPR` estimator. Due to similarity we do not report this plot. Naturally, the estimated coefficient functions exhibit very similar behaviour, thus the corresponding figure is not reported. As already seen a lot before, once the observations are increased to $10-20$ observations per curve, the resulting coefficient function estimates look very much alike the estimates in the regular case.

As for predictions, we do observe a substantial difference in terms of mean squared prediction error. Corresponding boxplots can be seen in Figure 5.2.8. Especially in the $3-5$ observations per curve case, the errors are much smaller for every methods. However, the `(R)FPCPR` estimators still seem to outperform the S-methods to some extent. One reason might be that in the aggressive setting, imputation might make more sense and that the overall increase of available data points is more influential than the possible introduction

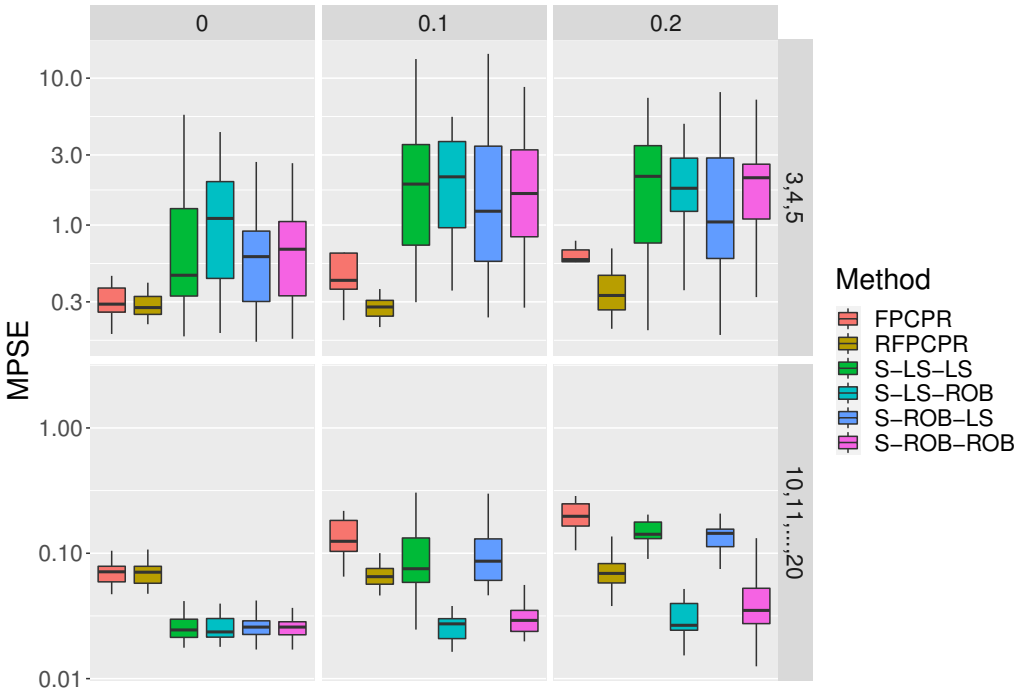

Figure 5.2.7.: Model 2 - Irregular, Non-Aggressive: Boxplots of MSPEs

of very incorrectly estimated curves by imputation. The S-methods also show quite some variability in their estimations due to complex numerical operations.

Figure B.8 in the appendix shows the corresponding plot of fitted values vs. true values which is more much improved in contrast to the non-aggressive setting with $3-5$ observations per curve. The point clouds are much denser and the effect of contamination is also visible to some extent, seen by the tilting of the red points. Nevertheless, some predictions fail miserably which is not always visible in the boxplots since outliers are not displayed there. In the $10-20$ observations case this plot looks very much alike the one in the regular case and is not reported here.

5.3. Conclusions

In this simulation study we compared 6 non-robust and robust estimators on 2 different models. The models differ by their coefficient function. In the first model, the coefficient function can be expressed using just 3 eigenfunctions of the covariance operator while in the

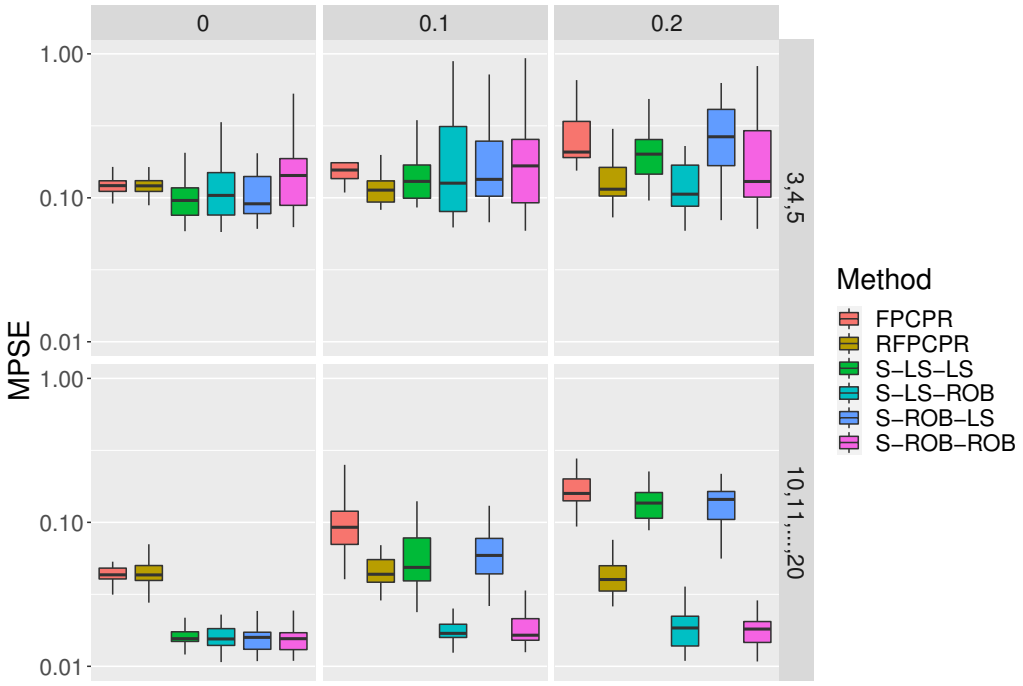

Figure 5.2.8.: Model 2 - Irregular, Aggressive: Boxplots of MSPEs

second model an infinite number of eigenfunctions would be needed. The estimators can be grouped by how the FPCA part of the algorithm is executed. The `(R)FPCPR` estimators are based on projection pursuit and heavily depend on the goodness of numerical integration. Hence they need data densely observed to be able to approximate well the necessary integrals. This is also what we observed in the regular cases for both models where the corresponding approximations lacked accuracy in the case of $n = 20$. The second group of estimators do not depend on numerical integration - they rather have a distributional assumption and estimate the FPCA part by conditional expectations. Their strengths can be seen in low-observed cases.

We also considered irregularly observed data where we differed between $3-5$ or $10-20$ observations per each curve. We observed the $10-20$ case to be quite similar to the regular case already whereas major problems arose in the $3-5$ setting. No method could properly estimate the coefficient function in any model except for Model 2 where useful approximations could be made in the aggressive setting.

To conclude, the robust methods introduced in this thesis have shown their usefulness in certain cases. In terms of estimation of the coefficient function, the `(R)FPCPR` estimators outperform the S-methods in the first model once a sufficiently large number of observations is available while in the second model it seems to be the other way around. As for predictions, the S-methods almost always yield better results than the competing estimator based on numerical integration. So depending on the goal, the proper estimator has to be chosen.

6. Real World Example

To test the estimators on real data, we take a look at the Canadian Weather of Ramsay and Silverman (2005) which is also used in Kalogridis and Van Aelst (2019). This data contains weather data such as temperature and precipitation of 35 weather stations all over Canada. The data was collected from 1960 to 1994 and averaged on a daily basis. The goal is to explain the response, the logarithmic annual precipitation, by the temperature curves of each weather station using a scalar-response model as described before. Optimally we will see which time of the year influences the annual precipitation the most.

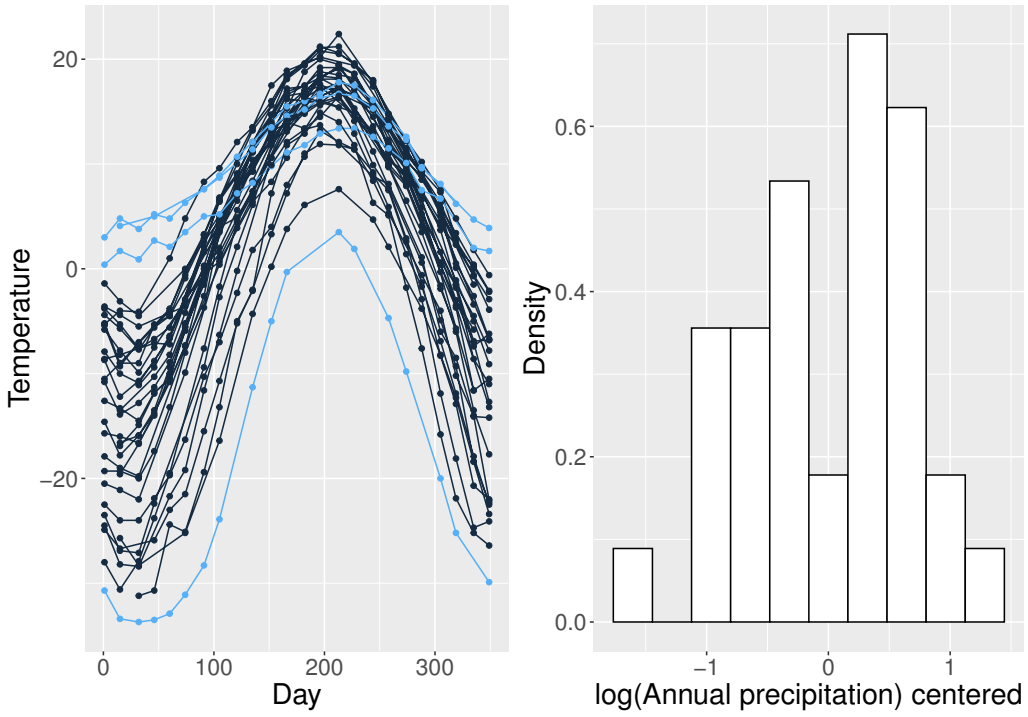

Figure 6.1.: Temperature Curves and Histogram of Annual Precipation of Canadian Weather Data

However, we only consider 2 data points per month, namely, the temperature measured at the first and 15-th day of each month, respectively. We have also encountered some missing

measurements, resulting in 35 observations with 16 to 24 observations per each temperature curve. Respective plots can be seen in Figure 6.1 where outliers in both temperature (light blue) and precipitation are present.

We set a maximum $K = 8$ and obtain following estimations and predictions. The `FPCPR` method chooses 3 functional principal components while its robust alternative selects 4 FPCs. Three components are also chosen by every S-method, see Figure 6.2. Note that the `(R)FPCPR` methods preprocess the raw data using B-splines. This way of smoothing the raw data is important to obtain actual functional objects which help in the estimation process. In contrast, no preprocessing is done for the S-methods since they are based on sparse data for which no presmoothing is usually suited. However, internal smoothing is still done on the covariance level to obtain a smooth covariance estimator to be able to properly estimate the eigenfunctions.

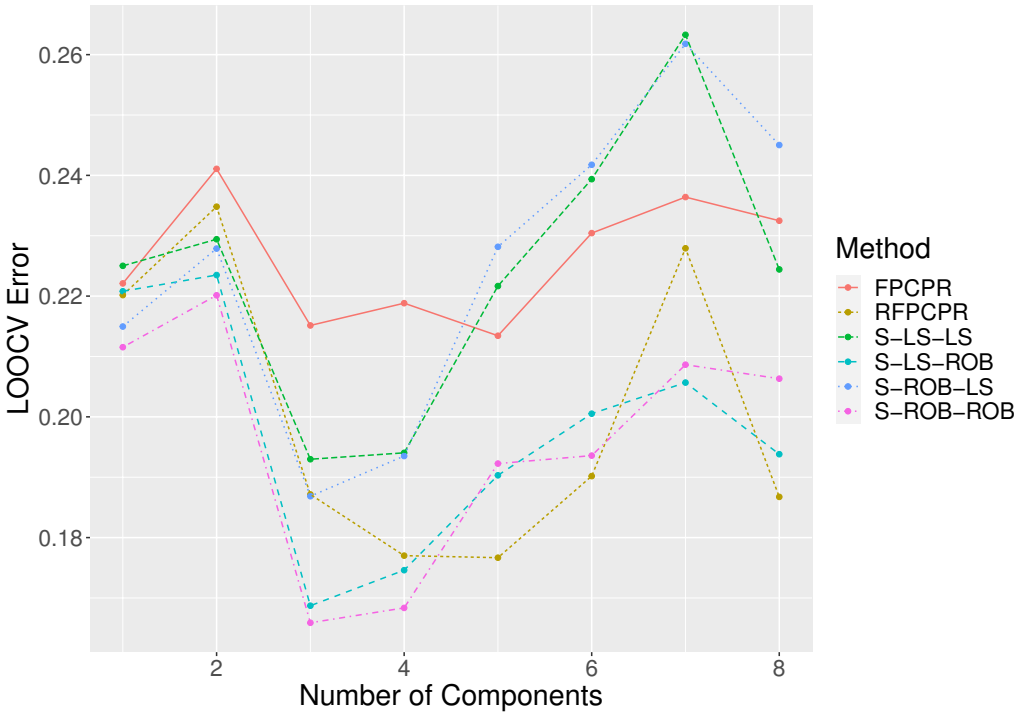

Figure 6.2.: LOOCV Errors of Sparse Canadian Weather Data

One notable difference in the two types of estimators is the quite different scale of values (Figure 6.3). One reason might be that the time points of curves are not sampled

very densely, hence the `(R)FPCPR` methods naturally have problems since they are based on numerical integration. However, similar effects are still present. While the effect of temperature is the biggest at the beginning and end of the year for almost all estimators, the lowest effect is quite different. For the `(R)FPCPR` methods, the minimum is obtained around the middle of the year whereas the S-methods have their minimum around March. It also appears that robust and non-robust estimators look very similar.

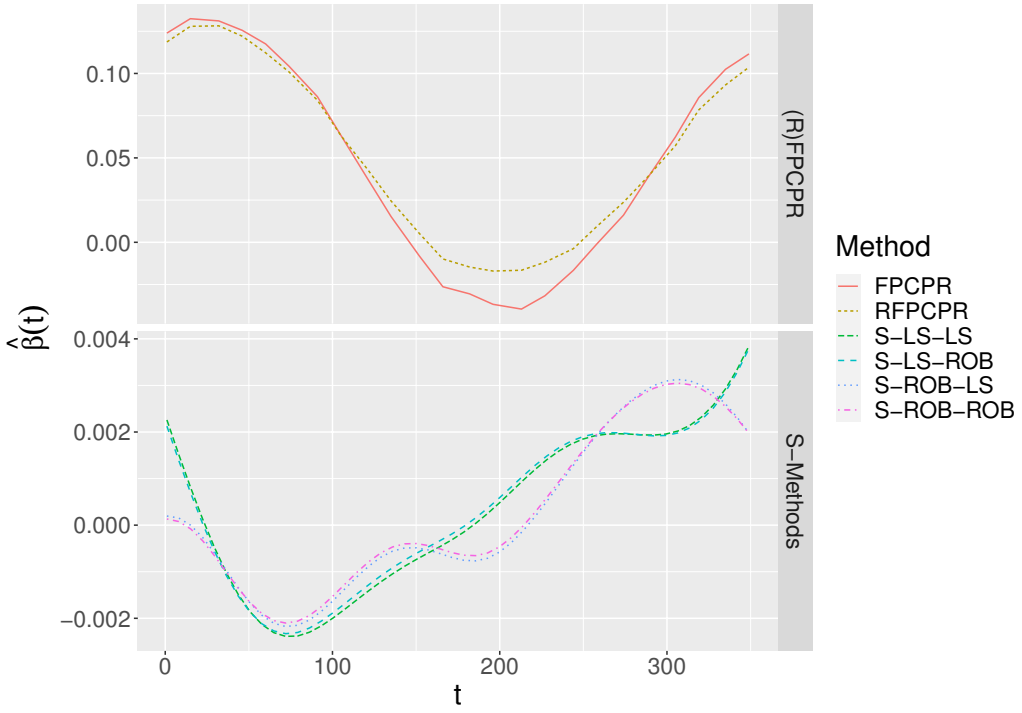

Figure 6.3.: Coefficient Function Estimates of Sparse Canadian Weather Data

Next, we might look at how the estimator actually predicts the annual precipitations. In Figure 6.4 we do see that the `(R)FPCPR` methods lack general accuracy while the S-methods predict better the true values. Nevertheless, no methods can predict the correct values very well. Based on these plots, the S-methods still seem to be better suited for that kind of data which is also why we trust the estimated coefficient functions of these methods more. Figure 6.2 does not only show the number of components chosen, it also shows the prediction errors in terms of LOOCV errors. Based on this plot, `S-LS-ROB` and `S-ROB-ROB` yield the best results which is consistent to the simulation study.

Finally we look at the corresponding QQ-plots of the residuals (Figure 6.5). Most points

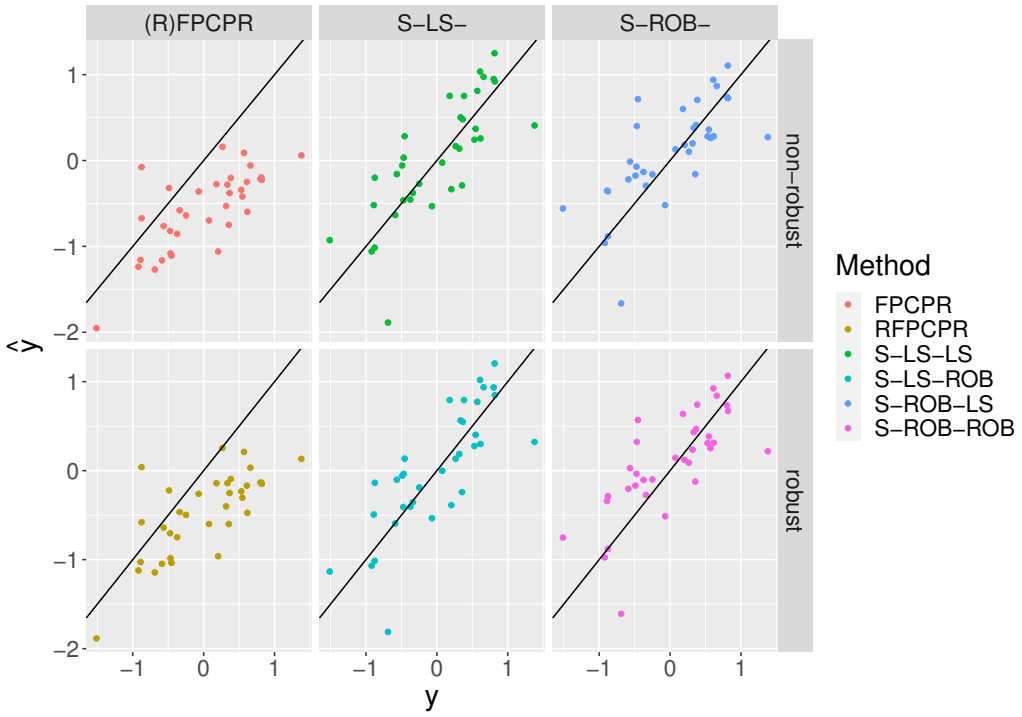

Figure 6.4.: Fitted Values vs. True Values of Sparse Canadian Weather Data

lie on the respective QQ-line which may indicate normality of the errors. But the points of the S-LS-methods do not seem to fully fit the corresponding QQ-line indicating some non-normality. Further, we also marked very extreme outliers for each estimator. The fully non-robust methods (`FPCPR`, `S-LS-LS`) mark Pr. Rupert and Scheffervll as potential outliers. However, for robust methods these are not really considered outliers anymore. Just Pr. Rupert is detected as an outlier in the `S-ROB-ROB` estimator. For the `RFPCPR` and `S-ROB-LS` estimator Dawson is also considered to be an outlier. Indeed, the temperature of Pr. Rupert has a quite flat curve while the temperatures in Dawson and Scheffervll are very low at the beginning and end of the year. In terms of annual precipation, it is quite high in Pr. Rupert compared to the other weather stations indicating a vertical outlier as well.

To compare, we also applied all methods to the daily sampled data. Resulting coefficient function estimates can be seen in Figure 6.6 whereby the S-methods using robust regression are not displayed here. These methods do not seem to work due to chosing too many components (5 for `S-LS-ROB`, 8 for S-ROB-ROB) and hence overfitting the coefficient function

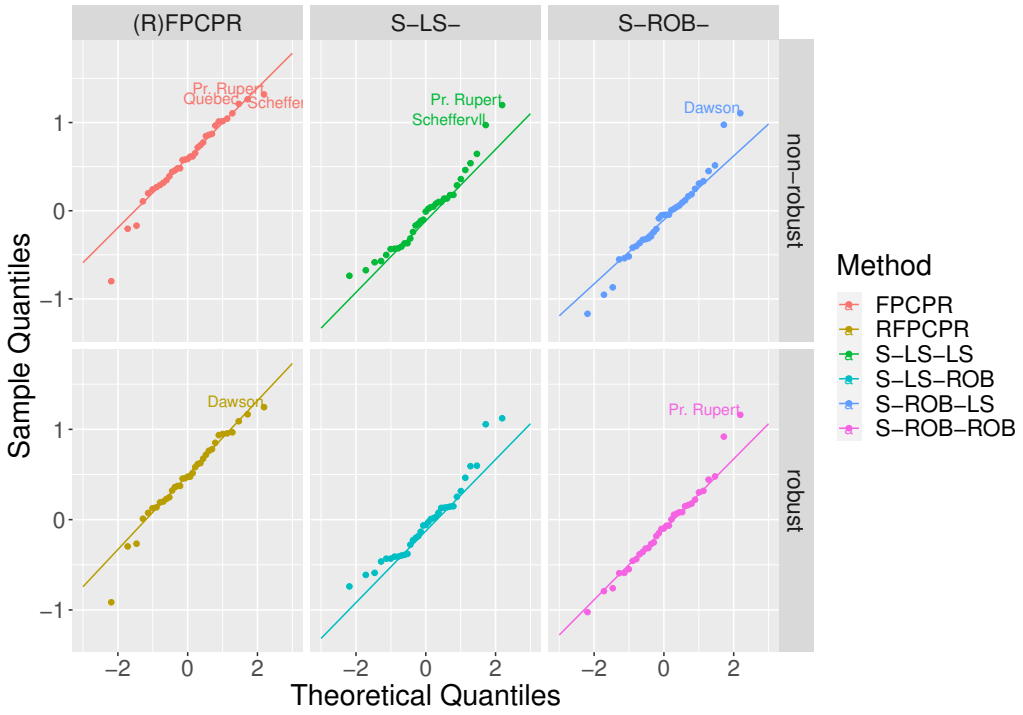

Figure 6.5.: QQ-Plot of Residuals of Sparse Canadian Weather Data

(see Figure C.1). This is also a problem we encountered in the simulation study in the regular settings. Methods `S-LS-LS` and `S-ROB-LS` yield the same number of chosen components and very similar looking estimators for the coefficient function. In terms of the `(R)FPCPR` methods we also encounter overfitting of the `RFPCPR` method since it chooses 6 functional principal components. However, the basic form of the estimator is still comparable to the S-methods. Its non-robust alternative only chooses 4 components. The resulting estimator looks very much alike the one obtained from the sparse case.

In conclusion, the newly proposed S-methods work rather well on sparse, longitudinal data but encounter problems once the data turns out to be too regular due to numerical issues. When comparing these results to the ones of Kalogridis and Van Aelst (2019) where similar methods were applied to the daily data, we do see some differences. First, the basic form of the coefficient function estimates is quite similar to the results yielded by the S-methods. The temperature effect is higher at the start and end of the year, and takes its minimum around March. Regarding outliers, the authors found Inuvik, Pr. Rupert and Kamloops to be significant outliers whereas our analysis yielded Pr. Rupert and Dawson as

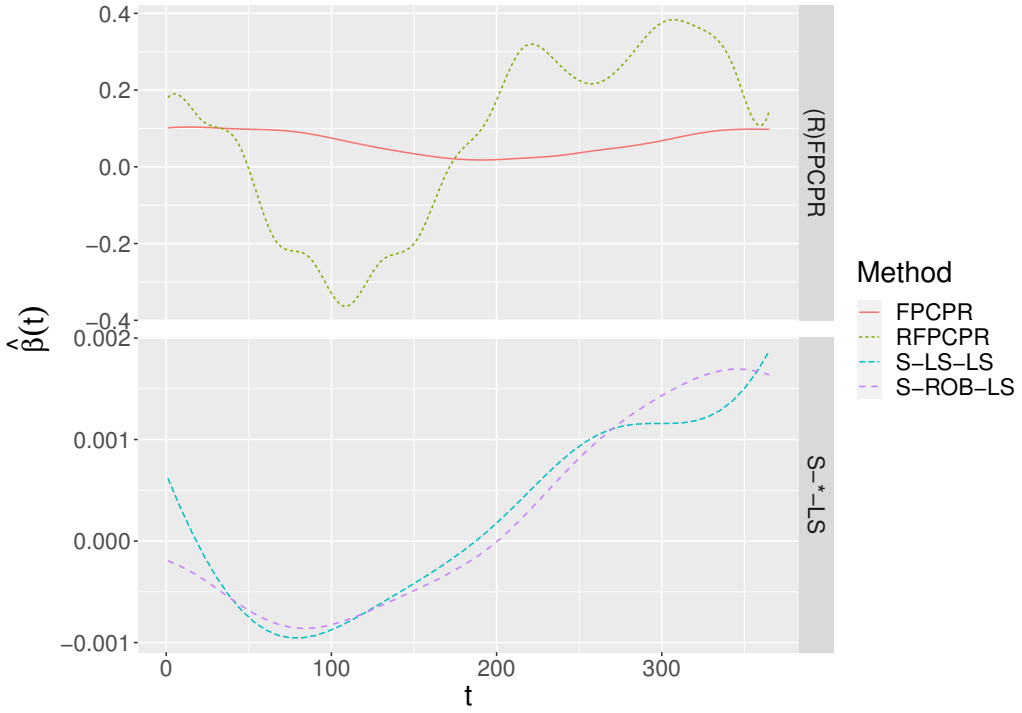

Figure 6.6.: Coefficient Function Estimates of Daily Canadian Weather Data

very influential observations. Reasons for these distinctions might be the irregularity and sparseness of the data since we compressed the data by over 90% in terms of observations per curve and the possible use of an intercept which we have not used in this analysis but the authors may have used. Nevertheless, the estimations shown in this analysis still result in sensible estimators and outliers.

7. Conclusions

In this diploma thesis we looked at robust methods for scalar-on-function regression using functional principal components. We focused on settings where the number of observations of each curve is limited. Overall we differed between two approaches of estimators. On the one hand we applied the work of Kalogridis and Van Aelst (2019) and adapted it to also work on irregular, sparse data, i.e. we generalized the estimators. This is an often seen use-case in real word data. This approach is based on robust estimation of the mean function, eigenbasis using a projection pursuit, and the β-coefficient using MM-estimation. On the other hand, we applied the FPCA methods of Boente and Salibián-Barrera (2021) which are based on distributional assumptions and robust eigen-decomposition, in a similar fashion to build a comparable yet more general functional regression framework for scalar-response models. For comparison reasons we also consider non-robust alternatives for both types of estimators. We also saw that the regression part being robust compared to the FPCA part appears to be more important since the `S-LS-ROB` and `S-ROB-ROB` estimators oftentimes yield comparable results. This is a new insight and has not been considered yet in case of robust sparse functional regression to our best knowledge.

A simulation study showed good performance of the proposed estimators in a regular setting, i.e. where all curves have been sampled at the same time points. We also saw that in the case of just a few observed time points the methods based on conditional expectations outperform the other ones. However, for irregular data with a very low amount of observations per curve, the results are not as satisfiable as systematic errors are present.

The estimators were also tested on a real world example using the Canadian Weather data where the annual precipitation was to be explained using temperature curves of various weather stations. The proposed S-methods did a good job at estimating the effect of temperature, and it was also possible to identify outliers which were not identified by ordinary methods. However, the all methods using robust regression had problems of overfitting on the daily data resulting in bad estimates.

It would be interesting to extend these ideas to the case of a fully-functional model as well. Based on past literature, methods using conditional expectations should work rather well similar to the work of Yao et al. (2005).

A. Simulation Results

Table A.1.: Model 1 - Regular: Number of Components

n	ε	FPCPR	RFPCPR	S-LS-LS	S-LS-ROB	S-ROB-LS	S-ROB-ROB
20	0	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)
	0.1	3.03 (0.18)	3.03 (0.18)	3.00 (0.00)	3.03 (0.18)	3.00 (0.00)	3.00 (0.00)
	0.2	3.00 (0.00)	3.03 (0.18)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)
50	0	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)
	0.1	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)
	0.2	2.97 (0.18)	3.00 (0.00)	2.97 (0.18)	3.00 (0.00)	2.97 (0.18)	3.00 (0.00)
100	0	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)
	0.1	3.03 (0.18)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)
	0.2	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)

Table A.2.: Model 1 - Irregular: Number of Components

	n	ε	FPCPR	RFPCPR	S-LS-LS	S-LS-ROB	S-ROB-LS	S-ROB-ROB
Aggressive	$3,4,5$	0	3.00	3.27	2.76	2.83	2.90	2.86
			(0.79)	(0.74)	(0.58)	(0.47)	(0.77)	(0.74)
		0.1	3.03	3.07	2.63	2.83	2.67	3.00
			(0.89)	(0.69)	(0.61)	(0.65)	(0.55)	(0.79)
		0.2	2.93	3.03	2.72	2.76	2.80	3.20
			(0.91)	(0.89)	(1.16)	(0.58)	(0.71)	(0.61)
	$10,11,\ldots,20$	0	3.00	3.00	3.00	3.00	3.00	3.00
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
		0.1	2.97	3.00	2.97	3.00	2.97	3.00
			(0.18)	(0.00)	(0.18)	(0.00)	(0.18)	(0.00)
		0.2	3.00	2.97	3.03	3.07	3.07	3.13
			(0.26)	(0.18)	(0.32)	(0.25)	(0.37)	(0.35)
Non Aggressive	$3,4,5$	0	3.67	3.77	2.80	2.93	3.10	3.27
			(1.06)	(1.01)	(0.81)	(0.74)	(0.76)	(0.87)
		0.1	3.41	3.87	3.17	3.67	2.90	3.17
			(1.43)	(1.31)	(1.26)	(1.45)	(0.84)	(0.99)
		0.2	3.00	3.43	3.03	3.52	3.00	3.37
			(1.31)	(1.41)	(1.12)	(1.21)	(1.11)	(1.25)
	$10,11,\ldots,20$	0	3.00	3.03	3.03	3.03	3.10	3.07
			(0.00)	(0.18)	(0.18)	(0.18)	(0.31)	(0.25)
		0.1	3.00	3.00	3.03	3.03	3.03	3.00
			(0.00)	(0.00)	(0.32)	(0.18)	(0.18)	(0.00)
		0.2	3.00	3.00	3.10	3.10	3.10	3.07
			(0.00)	(0.00)	(0.48)	(0.40)	(0.40)	(0.37)

Table A.3.: Model 2 - Regular: Number of Components

n	ε	FPCPR	RFPCPR	S-LS-LS	S-LS-ROB	S-ROB-LS	S-ROB-ROB
20	0	3.53 (1.48)	3.93 (1.72)	3.93 (1.53)	4.03 (1.50)	3.80 (1.40)	3.90 (1.37)
	0.1	3.21 (1.78)	4.30 (1.80)	3.28 (2.03)	4.93 (1.36)	3.36 (1.83)	5.36 (1.66)
	0.2	3.34 (1.76)	4.28 (1.60)	2.85 (1.64)	4.08 (1.29)	3.00 (1.35)	4.64 (1.66)
50	0	3.60 (1.25)	3.73 (1.28)	3.40 (1.25)	3.67 (1.49)	3.60 (1.25)	3.87 (1.46)
	0.1	3.43 (1.75)	4.13 (1.68)	2.79 (1.23)	3.75 (1.62)	3.62 (1.88)	4.04 (1.48)
	0.2	3.79 (2.08)	4.10 (1.45)	3.37 (1.47)	3.80 (1.65)	3.70 (1.73)	4.13 (1.74)
100	0	4.27 (1.39)	4.40 (1.33)	3.32 (1.28)	3.50 (1.32)	3.64 (1.37)	3.71 (1.38)
	0.1	3.04 (1.67)	4.20 (1.58)	2.66 (1.45)	3.97 (2.01)	2.88 (1.45)	3.65 (1.70)
	0.2	3.40 (1.79)	4.43 (1.77)	3.10 (1.54)	3.70 (1.42)	3.07 (1.65)	4.24 (1.75)

Table A.4.: Model 2 - Irregular: Number of Components

	n	ε	FPCPR	RFPCPR	S-LS-LS	S-LS-ROB	S-ROB-LS	S-ROB-ROB
Aggressive	$3,4,5$	0	2.73	3.20	2.46	2.68	2.38	2.85
			(1.26)	(1.92)	(1.20)	(1.31)	(0.94)	(1.19)
		0.1	3.05	3.30	2.70	3.22	2.59	3.06
			(1.61)	(1.64)	(1.52)	(1.81)	(1.50)	(1.56)
		0.2	2.79	2.83	2.57	2.76	2.38	3.43
			(1.08)	(1.53)	(1.80)	(1.70)	(0.80)	(2.20)
	$10,11,\ldots,20$	0	3.53	3.70	2.17	2.40	2.37	2.50
			(1.20)	(1.24)	(0.46)	(0.86)	(0.61)	(0.78)
		0.1	3.12	3.83	2.52	3.04	2.76	3.16
			(1.67)	(1.76)	(1.25)	(1.51)	(1.51)	(1.46)
		0.2	2.65	3.33	2.50	2.82	2.48	2.78
			(1.16)	(1.71)	(1.17)	(1.22)	(1.22)	(1.09)
Non Aggressive	$3,4,5$	0	3.70	3.70	2.38	3.71	2.25	2.67
			(1.69)	(1.75)	(1.32)	(1.93)	(0.68)	(1.24)
		0.1	4.56	4.04	2.82	3.18	3.14	3.64
			(2.24)	(1.76)	(1.38)	(1.63)	(1.98)	(1.68)
		0.2	3.50	5.15	2.94	3.65	3.05	3.79
			(1.58)	(2.32)	(1.25)	(1.66)	(1.90)	(1.93)
	$10,11,\ldots,20$	0	2.87	3.20	2.23	2.37	2.50	2.83
			(1.07)	(1.19)	(0.77)	(0.93)	(1.01)	(1.42)
		0.1	2.93	3.47	3.18	2.61	3.15	2.63
			(1.47)	(1.53)	(1.61)	(1.13)	(1.63)	(0.93)
		0.2	2.89	3.90	2.71	2.96	2.50	3.58
			(1.10)	(1.60)	(1.18)	(1.43)	(0.76)	(1.79)

B. Simulation Figures

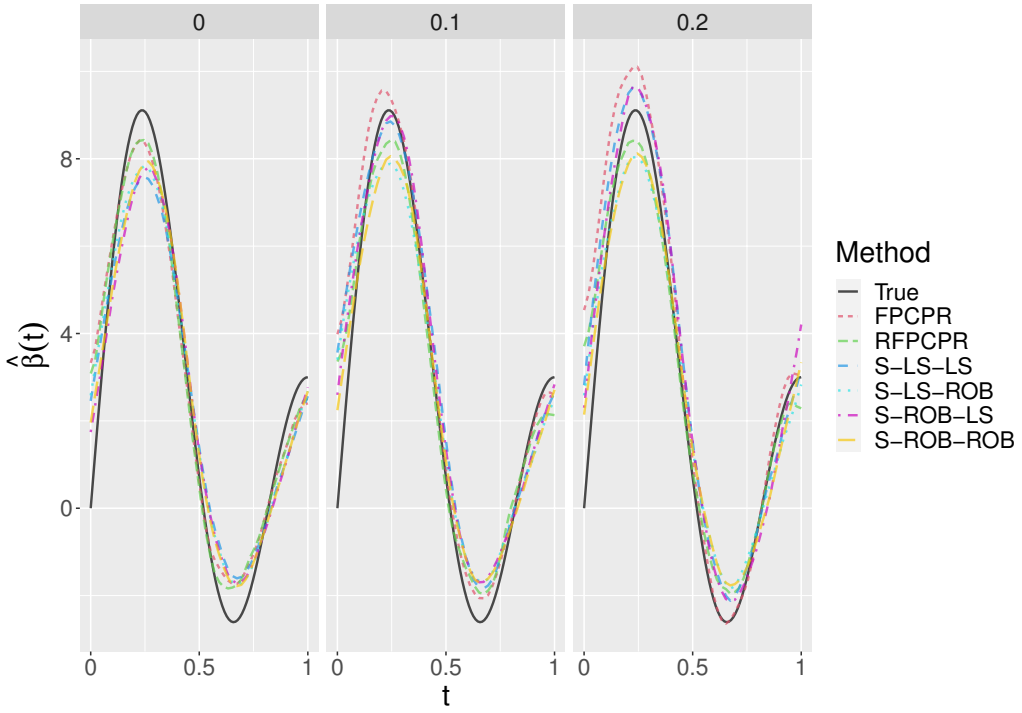

Figure B.1.: Model 1 - Irregular, Non-Aggressive, $10 - 20$ observations: Estimated Coefficient Functions

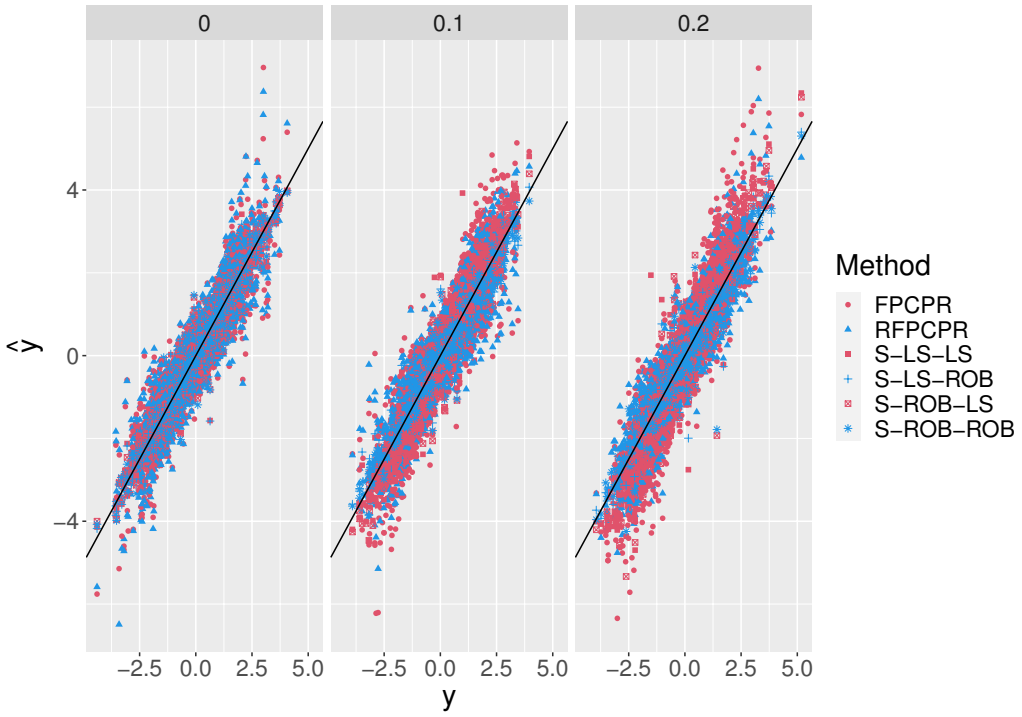

Figure B.2.: Model 1 - Irregular, Non-Aggressive, $10 - 20$ observations: Fitted Value vs. True Values

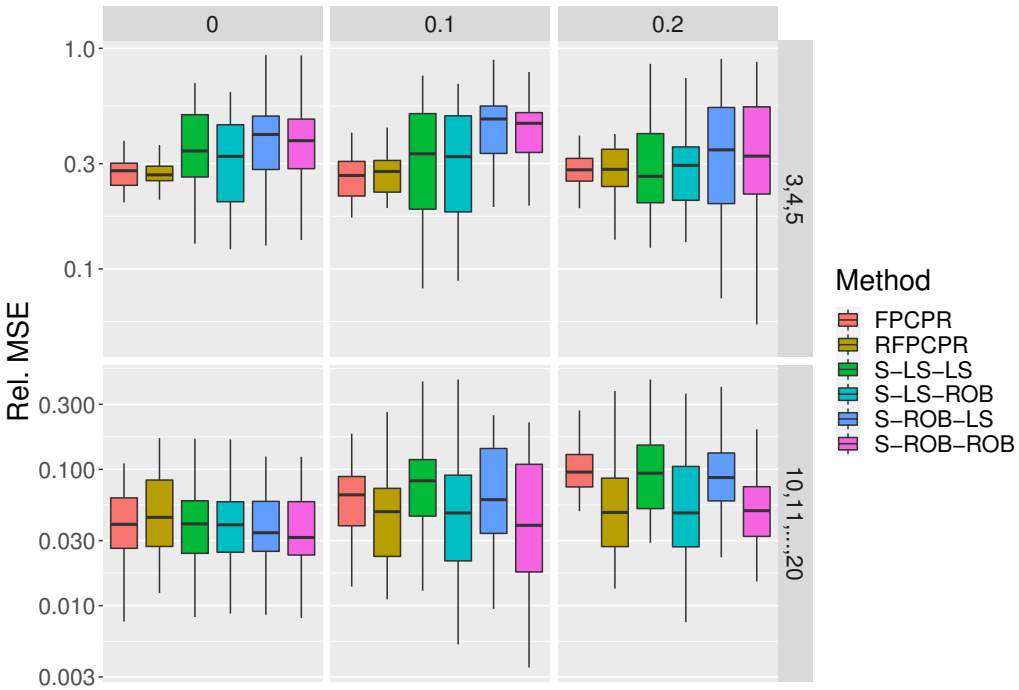

Figure B.3.: Model 1 - Irregular, Aggressive: Boxplots of relative MSEs

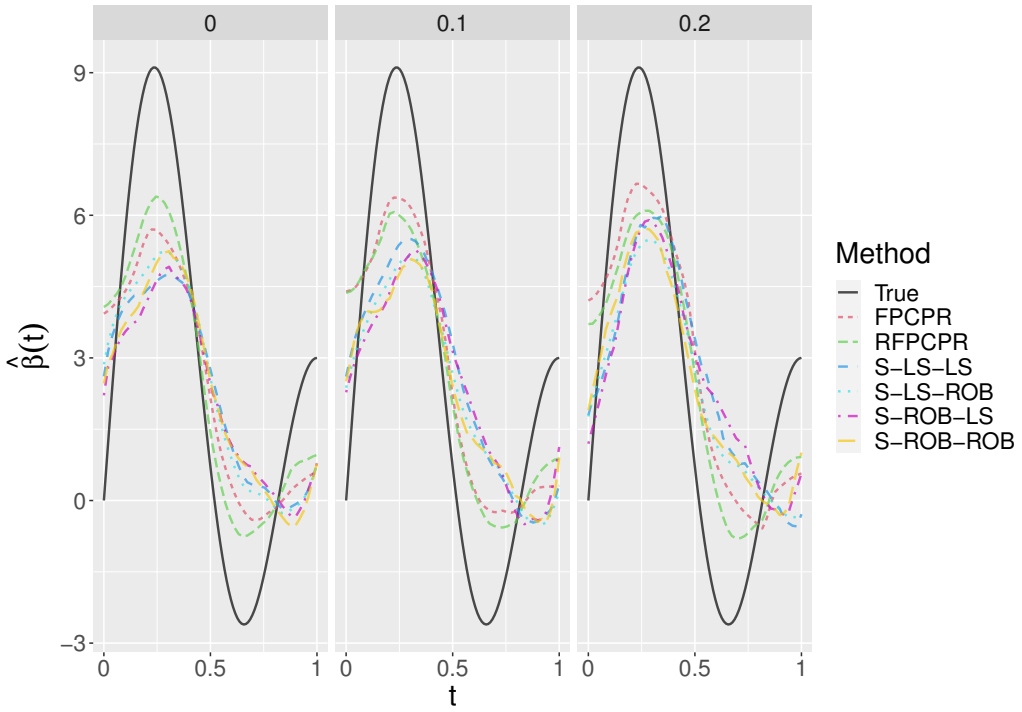

Figure B.4.: Model 1 - Irregular, Aggressive, $3 - 5$ observations: Estimated Coefficient Functions

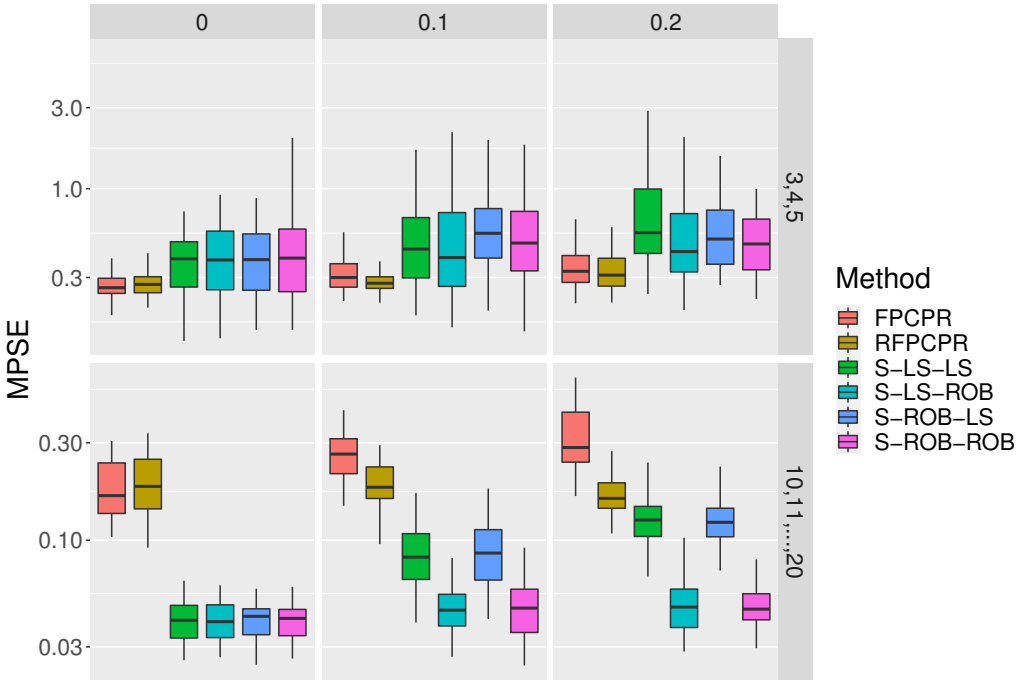

Figure B.5.: Model 1 - Irregular, Aggressive: Boxplots of MSPEs

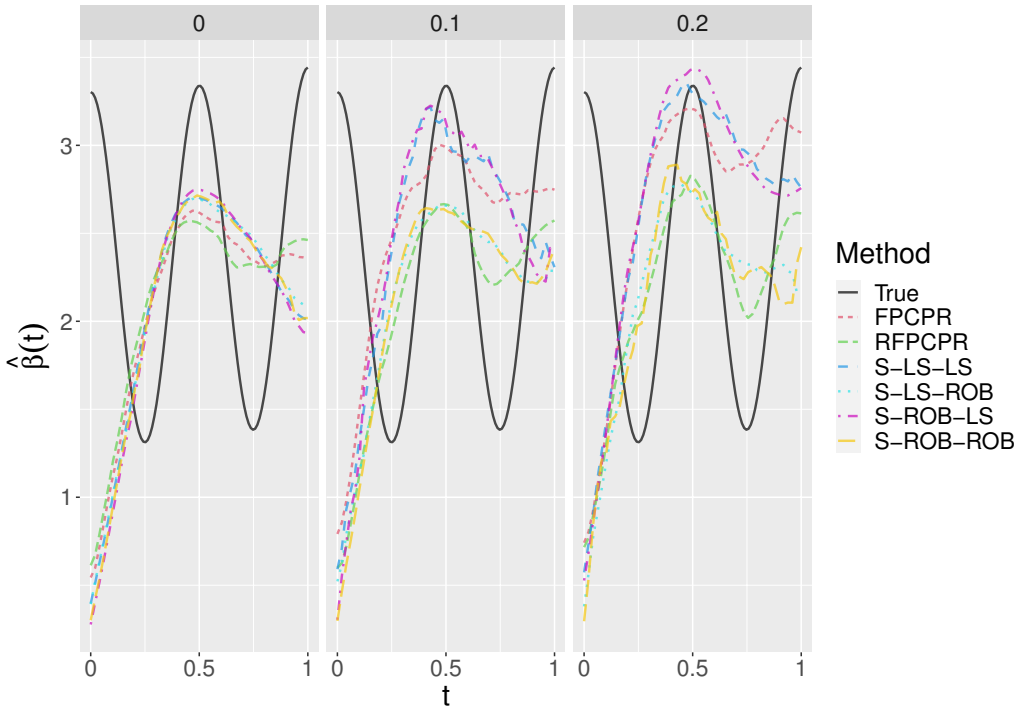

Figure B.6.: Model 2 - Irregular, Non-Aggressive, $10 - 20$ Observations: Estimated Coefficient Functions

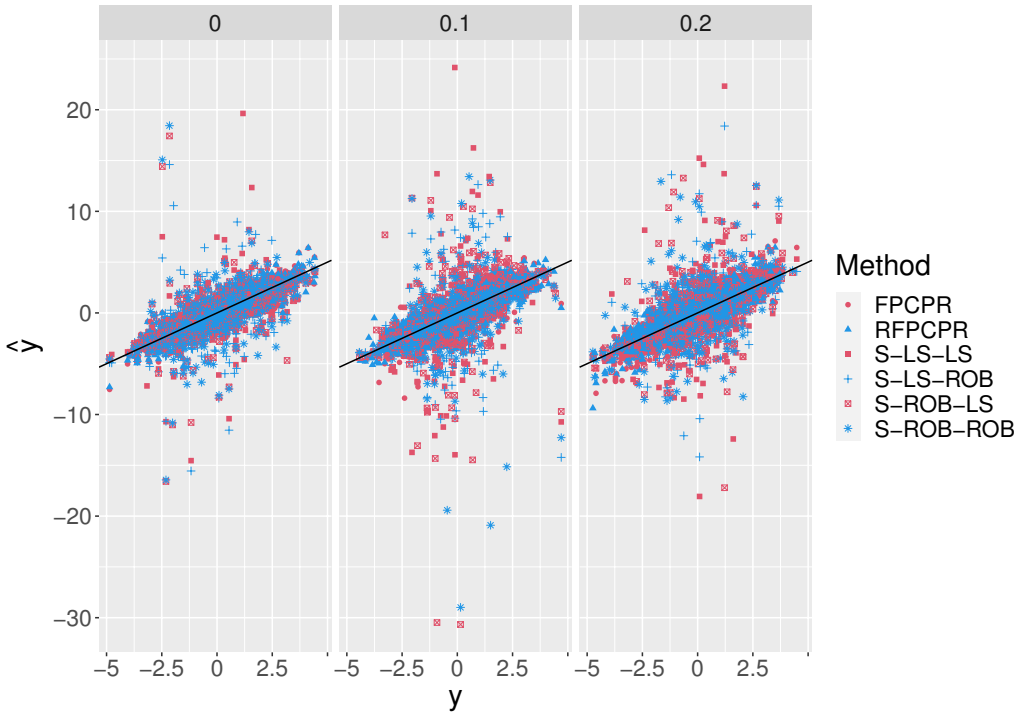

Figure B.7.: Model 2 - Irregular, Non-Aggressive, $3 - 5$ Observations: Fitted Values vs. True Values

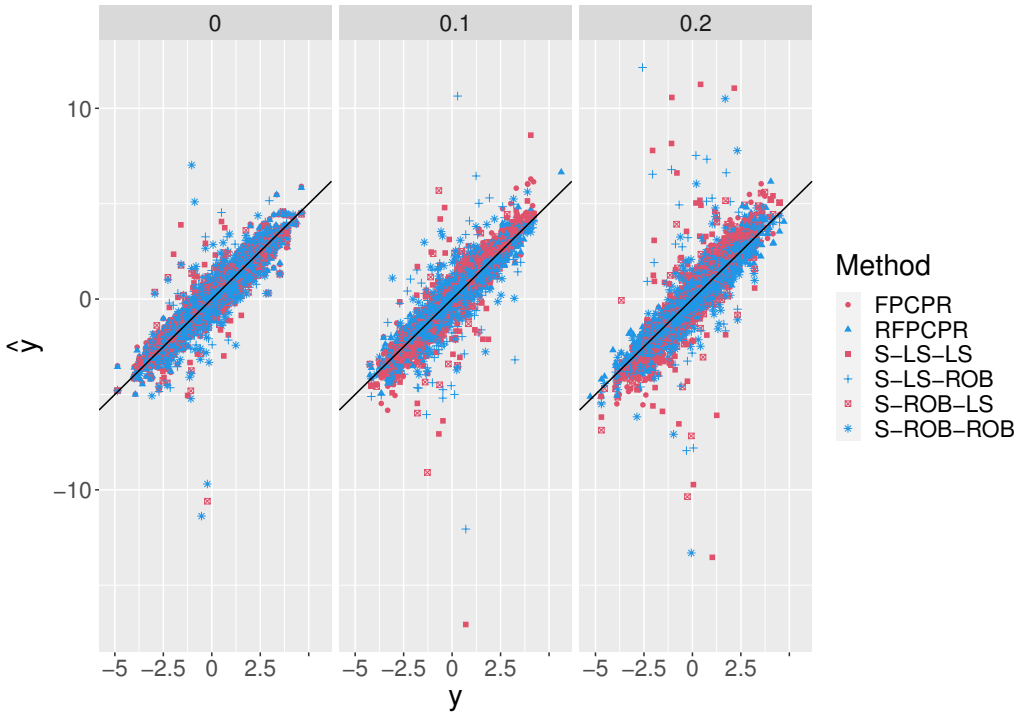

Figure B.8.: Model 2 - Irregular, Aggressive, $3-5$ Observations: Fitted Values vs. True Values

C. Real World Example Figures

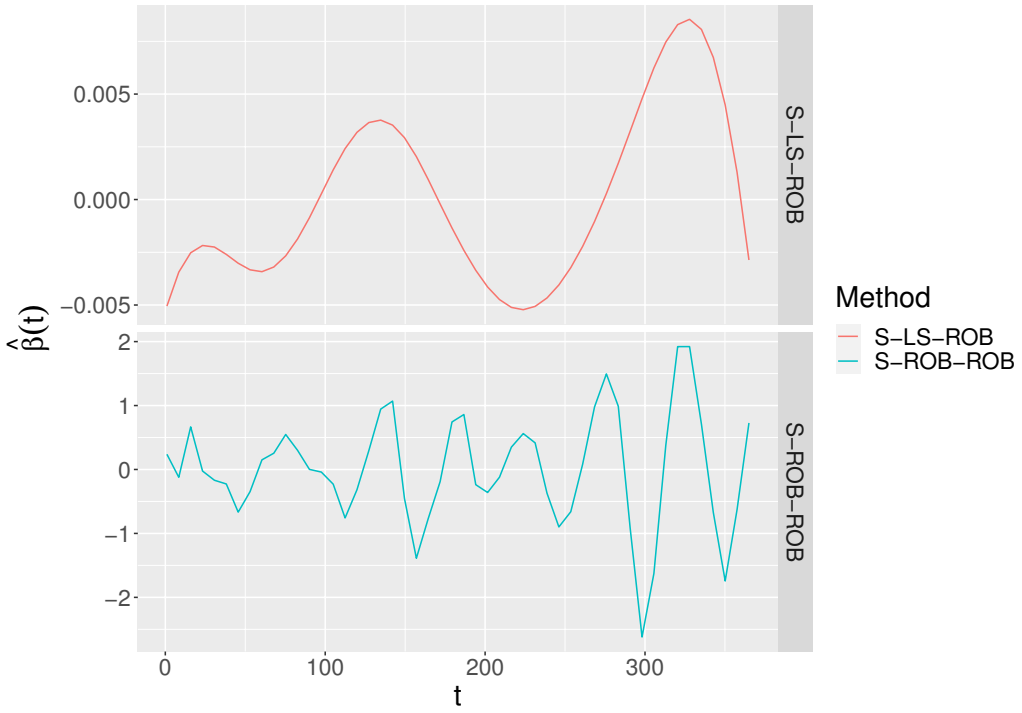

Figure C.1.: Coefficient Function Estimates for S-*-ROB Methods of Daily Canadian Weather Data

Bibliography

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

Bali, J. L. and Boente, G. (2009). Principal points and elliptical distributions from the multivariate setting to the functional case. *Statistics & Probability Letters*, 79(17):1858–1865.

Boente, G. (2021). *sparseFPCA: Robust Functional PCA for Longitudinal Data*. R package version 0.0.0.1.

Boente, G. and Salibián-Barrera, M. (2021). Robust functional principal components for sparse longitudinal data. *METRON*.

Boente, G., Salibián Barrera, M., and Tyler, D. E. (2014). A characterization of elliptical distributions and some optimality properties of principal components for functional data. *Journal of Multivariate Analysis*, 131:254–264.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *Int. Stat. Rev.*, 85(1):61–83.

Filzmoser, P. (2020). Multivariate statistics course notes. TU Wien, Austria.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second edition. Data mining, inference, and prediction.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data With Applications*. Springer Series in Statistics. Springer, New York.

Kalogridis, I. and Van Aelst, S. (2019). Robust functional regression based on principal components. *J. Multivariate Anal.*, 173:393–415.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. Theory and methods.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O., Graves, S., and Hooker, G. (2020). *fda: Functional Data Analysis*. R package version 5.1.9.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, second edition.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.*, 33(6):2873–2903.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.*, 83(402):406–413.

List of Figures

List of Tables