Fakultät für Mathematik und Wirtschaftswissenschaften
Institut für Zahlentheorie und Wahrscheinlichkeitstheorie

Spatial Functional Principal Component Analysis and its Application in Diagnostics

Dissertation
zur Erlangung des Doktorgrades Dr. rer. nat.
der Fakultät für Mathematik und Wirtschaftswissenschaften
der Universität Ulm

in Kooperation mit der
Roche Diagnostics GmbH, Penzberg

vorgelegt von
Insa Winzenborg

2011

Amtierender Dekan: Prof. Dr. Paul Wentges

Erstgutachter: Prof. Dr. Ulrich Stadtmüller
Zweitgutachter: Prof. Dr. Volker Schmidt

Tag der Promotion: 21. Juni 2011

Contents

1 Introduction

In many areas (including e.g. medicine, biology, ecology and econometrics), data are measured that have naturally a functional context. One example for measurements in time in the medical area are regular visits of patients where indicators for the medical condition, so-called biomarkers, are measured. Examples in space also occur very often, for example in image analysis and ecology. The functional context of the data shall be directly addressed by using functional data analysis. But what does functional data analysis mean?

To express it in terms of probability theory, let T be an index set and $(\Omega, \mathcal{F}, P)$ a probability space. The function $Y : T \times \Omega \to \mathbb{R}$ is a functional variable, if $Y(t, \cdot) : \Omega \to \mathbb{R}$ is a scalar random variable for each $t \in T$ and if T is *infinite*. If T was finite, we would be in the multivariate case. Hence a functional dataset consists of observations of I functional variables $Y_1, \ldots, Y_I$ identically distributed as Y. This thesis deals with the cases where T is a real interval (e.g. time) or a rectangle in $\mathbb{R}^2$ (e.g. space) and were the paths $Y(\cdot, \omega)$ are continuous for all $\omega \in \Omega$. In practice, one never observes the functions themselves, but measurements that are taken only at discrete measurement points. Furthermore, measurements can be error-prone and taken at irregular measurement points throughout the observations (e.g. in case of patient visits, the visit days usually vary from patient to patient). Hence suitable methods have to be applied to the measured data in order to obtain smooth observations or e.g. smooth moment estimators.

One could pose the question why this kind of data is treated as functional and not as multivariate data. The reason is, except for the treatment of the mentioned irregularities, that we want to include information of the environment, which can be done if some kind of continuity is assumed.

Essentially two main directions exist in the area of functional data analysis. In the first approach a set of basis functions over T is defined and the measurements are represented through this functional basis. Evaluations are based on the coefficients of this representation. Ramsay and Silverman [2006] is a comprehensive application-oriented reference book of this approach with further applications in Ramsay and Silverman [2002]. For a summarized overview see Levitin et al. [2007].

In contrast, there exists an approach that works without this kind of parametrization. Instead, smoothing techniques (mainly nonparametric) are applied in order to derive smooth functions. An introduction to this topic is given by Ferraty and Vieu [2003] and a more comprehensive treatment of techniques in this field in Ferraty and Vieu [2006].

As functional processes have (at least in theory) an infinite number of dimensions, it is crucial to concentrate on the important information in order to get an overview of the structure of the process. A method to do so is principal component analysis (PCA), which allows to extract the major modes of variations and to represent the infinite dimensional process with great accuracy through a small, finite basis. Principal components are an efficient way to represent the data through an orthonormal system, because the principal components system is optimal amongst all possible orthonormal systems in the sense that it retains most of the variability of the original process.

Principal component analysis is a popular method in multivariate analysis as it reduces the number of dimensions of a high dimensional data set to a few relevant ones. The first principle components, which are linear combinations of the original variables, are optimal in the way that they can explain the most variation in the data set of all possible orthogonal linear combinations. For an extensive overview of PCA in the multivariate analysis see Jolliffe [2004]. PCA is conceptually easy to extend to the functional case and, compared to multivariate PCA, it is even of greater use, because multivariate PCA often suffers from a lack of interpretation. The variables in a multivariate data set can explain features with totally different ranges and meanings. Hence a linear combination of them has often no clear meaning. Through the continuous index set in functional analysis, the principal components are also curves and can be seen directly as the major modes of variation.

Early work on functional PCA (FPCA) was done for example by Obhukov [1960], Dauxois et al. [1982], Castro et al. [1986], Besse and Ramsay [1986] and Bouhaddou [1987], but the method became more popular with the progress in computing speed.

FPCA in the context of the basis representation of functional data is treated in detail in Ramsay and Silverman [2006]. The FPC calculation is in this case accomplished through transforming the original problem to an analysis on the coefficients of the basis representation. Ocana et al. [2007] explain the equivalence of FPCA on curves and the multivariate PCA in the basis approach. Johnstone and Yu Lu [2009] discusses the method with emphasis on sparseness.

Nonparametric smooth estimators for mean, covariance and principal components are derived by Rice and Silverman [1991]. Silverman [1996] includes smoothing by choosing the norm appropriately. Boente and Fraiman [2000] treat kernel-based estimation methods for FPCA. Yao et al. [2003] and Yao et al. [2005] examine FPCA intensively, including analysis on regular grids as well as highly irregular and sparse data. Briefly summarized, their method is based on a nonparametric smoothing of the mean and covariance function of the process and principal components are estimated based on a discretized version of the covariance function.

Lots of other approaches and variants of FPCA exist which are adopted to various data situations. To name a few, in Yao and Lee [2006], the authors use penalized spline methods to estimate the mean function and Yao [2007] includes joint modeling of longitudinal and survival data in their FPCA framework. A Bayesian approach to FPCA is given by van der Linde [2008]. Benko et al. [2009] adapts FPCA to a two-sample problem and Cardot [2006] includes a covariate in their analysis. Müller et al. [2006] include a variance process instead of a non-random variance function in their modeling of a functional process and estimate additionally a variance function for each observation. In Müller and Stadtmüller [2005], the authors perform generalized linear modeling with random functions as predictors.

The major topic of this thesis is the extension of a temporal FPCA method to spatial functional data. The possibility of extension is mentioned by some authors, including Jolliffe [2004, Section 12.3], Ramsay and Silverman [2006, Section 8.5.3] and Yao et al. [2005]. An early reference (Preisendorfer and Mobley [1988, Section 2d]) carries out two approaches to FPCA in the spatial case: In the first approach the actual problem is replaced by a discrete dual problem, which is easier to solve. The second approach uses spatial basis functions and is essentially a spatial variant of the approach of Ramsay and Silverman [2006]. Braud et al. [1993] apply the FPCA method of Bouhaddou [1987], which also uses the dual approach. Actual executions of the extension in the nonparametric case including smoothing, convergence results and implementations are lacking up to our knowledge.

Therefore we extend the estimation method of Yao et al. [2003] and Yao et al. [2005] to spatial data and show the consistency of the estimators under certain conditions. In order to have a sound basis for this demonstration, we present the one-dimensional case at first and afterwards extend the theoretic framework and the nonparametric estimation of mean and covariance function as well as principal components to the spatial case. Estimation is performed through local regression smoothing techniques. For the two-dimensional approach, we further define an approach where only observations and principal components, but not the estimated mean and covariance functions, are smoothed, because this is favorable in computing time and also applicable in situations with non-sparse data. In both cases, one- and two-dimensional, we demonstrate and evaluate the estimation on simulations of a Wiener process. During evaluation of the two-dimensional Wiener process, we can show how to handle multi-dimensional eigenspaces. Both, the one- and two-dimensional process are applied in multiple data situations and finally we can successfully derive consistency rates, which are already available by Yao et al. [2005] for a slightly different data situation in the one-dimensional case, for the two-dimensional data situation. Furthermore, the one- and two-dimensional methods are implemented in an R package. This is done by using S4-classes, the current R standard of object-oriented programming, as described in Chapter 9 of Chambers [2009].
Before describing the structure of the thesis, we want to introduce the main application area:

In laboratory diagnostics, assays on patient samples are conducted in order to support the physician in diagnosing diseases. These assays determine the concentration of one or more analytes (substances in a sample) which are indicators for the disease under examination.

Presently a new generation of analysis systems is under development. These systems allow measuring multiple analyte concentrations with only one sample. Particularly in examining complex diseases with lots of analytes to be determined, this is a major facilitation.

The multi-parameter system (see Figure 1.1) is a system which realizes antigen-antibody reactions for different analytes on one unit and measures the level of specific bindings through fluorescent markers for each analyte. The core of this system is a chip with a matrix of spots on its even surface. In the practice variant one spot line (i.e. one column of spots) only contains spots for one type of analyte, but in the development phase, chips with a homogeneous coating on the whole surface are used as well, so that structures on the chip surface can be analyzed. Figure 1.2 shows examples for both types.

Measurements of a chip are taken on a grid, but even if the measurements are taken at discrete points, they can be seen as chosen point measurements of a whole outcome function. The measurement procedure and the data are described in more detail in Chapter 4.

Various questions arise in this context, for example how to evaluate the spatial measured signal structure or how to compare measurements of two platforms or of different units on the same platform. Further the performance shall be monitored over time and outlier measurements detected. These questions we want to address with the method of spatial FPCA.

But not only spatial examinations are of interest. When evaluating real samples, the values of the vertical lines (columns) in Figure 1.2(a) are averaged in a robust way. That is why we also analyze averaged column data using one-dimensional FPCA as well as other methods.

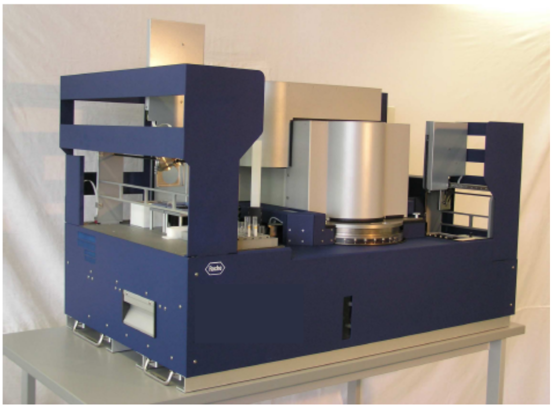
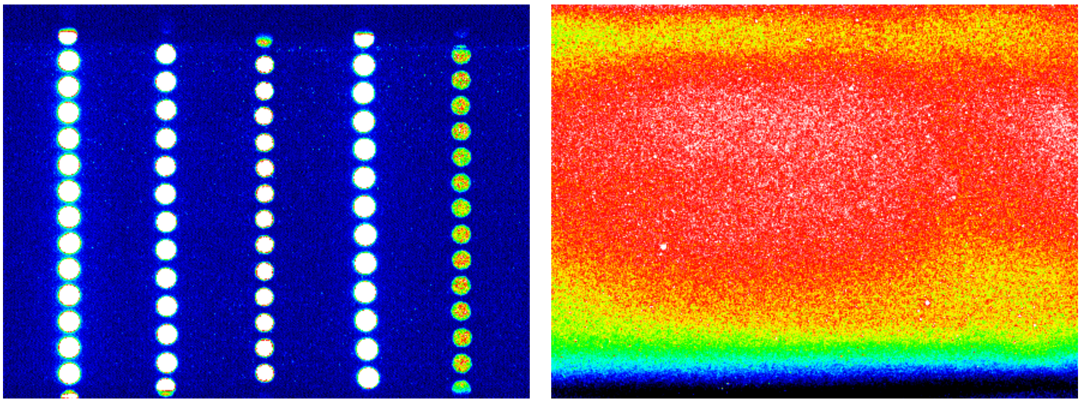


Figure 1.1: Multi parameter analyzer system



(a) Picture of a spotted chip surface

(b) Picture of a full field chip surface

Figure 1.2: Camera pictures

The thesis is organized as follows:
Chapter 2 first provides a brief introduction into multivariate PCA, before presenting the theory of the one-dimensional functional principal component analysis. Afterwards the nonparametric estimation of the principal components is presented and basics concerning kernels, smoothing and the choice of bandwidth are included. The method is applied to a Wiener process, because the FPCs of this process can be determined theoretically, such that the method can be validated using this process. Finally a section treats clustering of the FPCA outcomes, which is one typical way the FPCA results are used.
In Chapter 3 we extend the FPCA method from one- to two-dimensional domains. Prior to the estimation of the spatial FPCA, the theoretic framework for spatial FPCA is given. Like in the one-dimensional case, also in the two-dimensional case a Wiener process with theoretically computable principal components exists and we estimate the FPCs of the Wiener process also

in the spatial case.

Afterwards, the real data applications follow: Chapter 4 includes one-dimensional and Chapter 5 spatial applications. In the first application in Chapter 4 one-dimensional summaries of the measurement chips are clustered by different methods including a clustering method based on FPCA results. The second application is the single application in this thesis which is not settled in the area of the analysis system. In this application longitudinal measurements of clinical parameters are analyzed using FPCA. We use this example to compare the above-mentioned parametric FPCA approach by Ramsay and Silverman [2006] with our nonparametric method.

In the two-dimensional application Chapter 5 two different examples of spatial FPCA in the context of the analysis system are presented. In the first application chips of two different instruments are compared according to their variance structure. The second example includes repeated series of measurements of the analysis system. The aim of that section is to apply spatial FPCA in order to analyze the variance structure respectively the change of the variance structure over time.

In Chapter 6, consistency results for the one- as well as the two-dimensional estimation are calculated. The one-dimensional case is close to the work of Yao et al. [2005], though we consider a slightly different framework and modified the proof at appropriate points.

The results of this chapter are already summarized in Chapters 2 and 3. Therefore only readers interested in the technical details need to read Chapter 6.

Chapter 7 explains the implementation of the one-and the two-dimensional FPCA method in the statistical programming language R and finally Chapter 8 provides a summary of the thesis and the discussion of some main points.

2 Functional Principal Component Analysis

Functional principal component analysis (in the following abbreviated by FPCA) is an extremely useful tool in functional data analysis, because it allows seeing variance structures which are not obvious when simply displaying the data. Furthermore, it provides an evaluation of the complexity of a data set through regarding how many components are needed to represent the data set in satisfying accuracy in the FPC basis. The FPC scores (i.e. the coefficients of representing the data through a principal component basis) contain information about the location of each observed curve in the FPC space and can be used for example to identify outliers or build clusters of curves with similar structure.

2.1 Theory

Before introducing PCA in the functional case, we want to discuss PCA in the well-known multivariate framework first. The multivariate PCA can be interpreted as an axis rotation. If one represents a data set in the new system of coordinates, the coefficients belonging to the first K rotated axes (for each $K = 1, 2, \ldots$) explain as much variation as possible with this number of components. Therefore one can reduce the new system by leaving out the last axes without losing much variation.

Assume an N-dimensional random vector $Y = (Y_1, \ldots, Y_N)^T$ with expected value vector

$$\mu = (\mu_1, \ldots, \mu_N)^T \quad \text{and covariance matrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \ldots & \Sigma_{1N} \\ \vdots & \ddots & \vdots \\ \Sigma_{N1} & \ldots & \Sigma_{NN} \end{pmatrix}.$$

Then the principal components can be defined as follows (Jolliffe [2004]):

Definition 1. A vector $\rho_1 = (\rho_{11}, \ldots, \rho_{1N})^T$ with $||\rho_1|| = \rho_1^T \rho_1 = 1$ is called *first principal component* of Y, if the variance $\mathrm{Var}(Y^T\rho_1)$ is maximal within the linear combinations with standardized coefficient vectors:

$$\mathrm{Var}(Y^T\rho_1) = \rho_1^T \Sigma \rho_1 = \max_{\{\rho \in \mathbb{R}^N \,|\, ||\rho||=1\}} \rho^T \Sigma \rho \tag{2.1}$$

For $k = 2, \ldots, N$ a vector $\rho_k = (\rho_{k1}, \ldots, \rho_{kN})^T$ with $||\rho_k|| = \rho_k^T \rho_k$ is called the *kth principal component* of Y if $\rho_k \perp \rho_l$, i.e. $\rho_k^T \rho_l = 0$ for $l < k$, $||\rho_k|| = 1$ and

$$\mathrm{Var}(Y^T\rho_k) = \rho_k^T \Sigma \rho_k = \max_{\{\rho \in \mathbb{R}^N \,|\, ||\rho||=1, \rho \perp \rho_l \text{ for } l<k\}} \rho^T \Sigma \rho. \tag{2.2}$$

This definition is directly based on the variance maximization property. Alternatively principal components can be defined as the eigenfunctions of the covariance matrix. We want to state this equivalent proposition as a theorem:

Theorem 1. *The kth principal component is an eigenvector of the kth largest eigenvalue of the covariance matrix Σ. Hence the principal components can be calculated by finding all pairs (λ_k, ρ_k) with $\lambda_k \in \mathbb{R}$ and $\rho_k \in \mathbb{R}^N, \rho_k \neq 0$ which solve the eigenequation*

$$\Sigma\rho = \lambda\rho \qquad \textit{under the condition } \rho_k \perp \rho_l \textit{ for } k \neq l.$$

Proof. This is a consequence of Lagrange's rule (Heuser [2002, Section 174]) because the rule states that for a solution ρ_1 in (2.1) there exists λ_1 such that (with grad being the gradient of a function)

$$\begin{aligned} \operatorname{grad}_{\rho_1} \left(\operatorname{Var}(Y^T\rho_1) - \lambda_1(1 - \rho_1^T\rho_1)\right) &= 0 \\ \Leftrightarrow 2\Sigma\rho_1 - 2\lambda_1\rho_1 &= 0 \end{aligned}$$

and for a solution ρ_k $(k \geq 2)$ there exist λ_k and $\nu_1, \ldots, \nu_{k-1}$ such that

$$\begin{aligned} \operatorname{grad}_{\rho_k} \left(\operatorname{Var}(Y^T\rho_k) - \lambda_k(1 - \rho_k^T\rho_k) - \sum_{l=1}^{k-1} \nu_l \rho_l^T \rho_k\right) &= 0 \\ \Leftrightarrow 2\Sigma\rho_k - 2\lambda_k\rho_k - 2\sum_{l=1}^{k-1} \nu_l\rho_l &= 0 \end{aligned}$$

Multiplying the formula with ρ_l the first two terms vanish and we can deduce $\nu_l = 0$ for $l = 1, \ldots, k-1$ and therewith

$$\Sigma\rho_k - \lambda_k\rho_k = 0$$

Furthermore, each solution of the eigenequation leads to a principal component. Observe that Definition 1 leads to N principal components due to the N dimensions of the space $\mathbb{R}^N$ and the condition that the principal components must be orthogonal. Therefore Lagrange's rule leads to N orthogonal eigenvectors of the covariance matrix. Moreover, we know from linear algebra that a positive semi-definite $N \times N$ matrix has exactly N orthogonal eigenvectors. Hence we can deduce that each eigenvector is also a principal component. □

In order to see the analogy to the functional case later on, we want to remark that for each element $x \in \mathbb{R}^N$, scores in the principal component space can be defined such that

$$x = \mu + \sum_{k=1}^{N} \xi_k(x)\rho_k \quad \text{with} \quad \xi_k(x) = (x - \mu)^T\rho_k.$$

Naturally we could define the scores also without subtracting the mean but as x is later regarded as an observation of the process Y it is consistent to use this representation.

Further the process Y itself can be represented through its principal components by

$$Y = \mu + \sum_{k=1}^{N} \xi_k(Y)\rho_k \quad \text{with} \quad \xi_k(Y) = (Y-\mu)^T\rho_k.$$

The scores $\xi_k(Y)$ are uncorrelated random variables with mean 0 and variance λ_k:

$$\mathbb{E}(\xi_k(Y)) = \mathbb{E}((Y-\mu)^T\rho_k) = \underbrace{\mathbb{E}(Y-\mu)^T}_{=0_N}\rho_k = 0$$

$$\mathrm{Cov}(\xi_k(Y),\xi_l(Y)) = \mathrm{Cov}((Y-\mu)^T\rho_k,(Y-\mu)^T\rho_l) = \rho_k^T\Sigma\rho_l = \begin{cases} 0 & k \neq l \\ \lambda_k & k = l \end{cases}$$

In the multivariate case, the observations are vectors $x = (x_1,\ldots,x_N) \in \mathbb{R}^N$ with the dot product $\langle x_1, x_2\rangle = x_1^T x_2$ as inner product for $x_1, x_2 \in \mathbb{R}^N$. In the functional case, the vectors are replaced by functions $f : T \longrightarrow \mathbb{R}$ (T is a bounded interval in $\mathbb{R}$). The appropriate inner product in this case is the integral $\langle f_1, f_2\rangle = \int_T f_1(t)f_2(t)\,dt$. In order to have a well-defined setting, we specify the following framework (compare e.g. Chiou and Li [2007]):

We consider functions in the space of square (Lebesque-)integrable functions as observations, i.e. functions in

$$L^2(T) = \left\{ f : T \longrightarrow \mathbb{R} \quad \middle| \quad \left| \int_T f(t)^2 dt \right| < \infty \right\}$$

for a bounded interval T in $\mathbb{R}$ with inner product $< f, g >= \int_T f(t)g(t)dt$ and the corresponding norm $||f|| = \sqrt{< f, f >}$. $L^2(T)$ is a Hilbert space with the defined metric if we identify the functions which are almost everywhere identical (in other words which differ only on a null set) (see Heuser [2006, Chapter 4]).

Furthermore, let $(\Omega, \mathcal{A}, P)$ be a probability space. Then we consider a stochastic process $Y : T \times \Omega \to \mathbb{R}$ such that $|\int_T \mathbb{E}(Y^2(t,\omega))\,dt| < \infty$ and hence almost all paths $t \to Y(t,\omega)$ are in $L^2(T)$. The mean function of the process is given by $\mu(t) = \mathbb{E}(Y(t))$. We furthermore assume that the second moment $\mathbb{E}(Y^2(t))$ exists everywhere, such that the covariance function $G(s,t) = \mathrm{Cov}(Y(s), Y(t))$ exists.

Then we can define the covariance operator $A : L^2(T) \longrightarrow L^2(T)$ with

$$(Af)(t) = \int_T f(s)G(s,t)\,ds \tag{2.3}$$

for $f \in L^2(T)$ and $t \in T$. In order to show that A is well-defined observe that for a function $f \in L^2(T)$ we deduce

$$\int_T \left(\int_T f(s)G(s,t)\,ds \right)^2 dt \leq \underbrace{\left(\int f(s)^2\,ds \right)}_{=||f||^2<\infty} \int\int \mathrm{Cov}(Y(s),Y(t))^2\,ds\,dt$$

$$\leq ||f||^2 \int_T\int_T \left[\mathbb{E}(Y(s)Y(t)) - \mathbb{E}(Y(s))\mathbb{E}(Y(t))\right]^2 ds\,dt$$

$$\leq ||f||^2 2 \int_T \mathbb{E}(Y(s))^2 \, ds \int_T \mathbb{E}(Y(t))^2 \, dt < \infty$$

because $f \in L^2(T)$ and Y is square integrable.
The covariance operator is a linear compact symmetric operator such that the general spectral theory for this kind of operators in Hilbert spaces can be applied (see e.g. Heuser [2006, Chapter 5]).
The principal components of the process Y are the eigenfunctions of the operator A. This time we provide the eigenfunction definition first and deliver the variance maximization criterion later on.

Definition 2 (Eigenfunctions and eigenvalues)**.** If a function $\rho \in L^2(T), \rho \neq 0$ and a constant $\lambda \in \mathbb{C}$ fulfill

$$A\rho = \lambda\rho \quad \Leftrightarrow \quad < G(\cdot, t), \rho > = \lambda\rho(t) \quad \text{for all } t \in T,$$

λ is called eigenvalue with eigenfunction ρ of A respectively G.

The main difference to the multivariate case is that the number of principal components is in general infinite. Despite the infiniteness, many of the properties of multivariate PCA are valid as well (see Heuser [2006, Chapter 5]). We have to keep in mind that the properties apply up to a null set.

Properties 2.

1. *The set of eigenvalues is countable and as A is symmetric and positive semi-definite, the eigenvalues are real, non-negative values, such that we can assume in the following that the eigenvalues are ordered as $\lambda_1 > \lambda_2 > \ldots \geq 0$ with corresponding eigenfunctions $\rho_1, \rho_2, \ldots$.*

2. *$\sum_{k=1}^{\infty} \lambda_k < \infty$, in particular $\lambda_k \to 0$*

3. *If for one $k \in \mathbb{N}$, $\lambda_k = 0$, the process is finite dimensional.*

Often, in the functional case, only eigenvalues > 0 are considered because they are sufficient to describe the underlying process. If the sequence of positive eigenvalues interrupts, there is an infinite (countable) set of eigenvalues equal zero in Definition 2. We will see later on, that this is not of practical relevance, but in this section it is important to notice the difference.

4. *The eigenspace of each eigenvalue > 0, i.e. for an eigenvalue λ the space*

$$\left\{ f \in L^2(T) \,|\, Af = \lambda f \right\}$$

is finite dimensional and the eigenspaces for different eigenvalues are orthogonal.

5. *If the eigenvalues are all > 0, the eigenfunctions define an maximal orthonormal system in the Hilbert space $L^2(T)$ and thereby also a orthonormal basis, provided the norm of the eigenfunctions is standardized. Due to the remark following 3., this is also true if the eigenvalues are zero starting at some index. The corresponding eigenfunctions than span the nullspace of A.*

6. *The covariance operator A and the covariance function G can be expressed through eigenvalues and eigenfunctions as*

$$Af = \sum_{k=1}^{\infty} \lambda_k \langle f, \rho_k \rangle \rho_k$$

$$G(s,t) = \sum_{k=1}^{\infty} \lambda_k \rho_k(s) \rho_k(t)$$

in $L_2(T)$ for $f \in L_2(T)$ and $s, t \in T$.

An important finding is that the stochastic process can be represented as a linear combination of eigenfunctions and random coefficients (see e. g. Bosq [2000, Section 1.2]):

Theorem 3 (Karhunen-Loève expansion)**.** *If Y is a square integrable process with continuous covariance function G, the process has a Karhunen-Loève expansion*

$$Y(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k(Y) \rho_k(t).$$

The coefficients $\xi_k(Y)$ are given via $\xi_k(Y) =< Y - \mu, \rho_k >$ and are called scores *in this context. The scores ξ_k for $k \in \mathbb{N}$ are uncorrelated random values with zero mean and variance λ_k. The Karhunen-Loève series converges in $L^2(T)$ and uniformly.*

Like in the multivariate case we have an equivalence of Definition 2 with the variance maximization criterion (Heuser [2006]). The maximization criterion reads as follows in the functional case:

Theorem 4 (Variance maximization criterion)**.** *ρ_1 is the function ρ that maximizes $|\langle A\rho, \rho \rangle|$ with the constraint $||\rho|| = 1$. Accordingly for $k > 1$, ρ_k is the function ρ that maximizes $|\langle A\rho, \rho \rangle|$ with the constraint $||\rho|| = 1$ and $\langle \rho, \rho_l \rangle = 0$ for $l \leq k - 1$.*

In the following, the properties of the FPCA as stated in Properties 2, Theorem 3 and Theorem 4 shall be proven.

Proof of Properties 2, Theorem 3 and Theorem 4. In order to see the equivalence, first note that $\sup_{\{f|\,||f||=1\}} |\langle Af, f \rangle| = ||A||$ ([Heuser, 2006, Chapter V]). Therefore a sequence f_l with $||f_l|| = 1$ exists such that $\langle Af_l, f_l \rangle \longrightarrow ||A|| =: \lambda$. Furthermore

$$0 \leq ||Af_l - \lambda f_l||^2 = ||Af_l||^2 - 2\lambda \langle Af_l, f_l \rangle + \lambda^2 ||f_l||^2$$

$$\leq ||A||^2 - 2\lambda \langle Af_l, f_l \rangle + ||A||^2 \longrightarrow_{l \to \infty} 0,$$

such that also

$$Af_l - \lambda f_l \longrightarrow_{l \to \infty} 0 \tag{2.4}$$

follows. As A is a compact operator, (Af_l) has a convergent subsequence (Af_{l_m}) and due to (2.4) also f_{l_m} converges to a limit function. Let the limit function be ρ. ρ has norm 1 and fulfills $A\rho - \lambda\rho = 0$. Hence ρ is eigenfunction of eigenvalue λ. Further observe that $|\langle A\rho, \rho\rangle| = \sup_{\{f|\,||f||=1\}} |\langle Af, f\rangle|$, because

$$|\langle A\rho, \rho\rangle| = |\langle \lambda\rho, \rho\rangle| = |\lambda|\,||\rho|| = ||A|| = \sup_{\{f|\,||f||=1\}} |\langle Af, f\rangle|.$$

Furthermore each function ρ that fulfills $|\langle A\rho, \rho\rangle| = \sup_{\{f|\,||f||=1\}} |\langle Af, f\rangle|$ is an eigenfunction of eigenvalue $||A||$ (choose $f_l = \rho$). Now define $\lambda_1 := \lambda$, $\rho_1 := \rho$ and $E_1 := [\rho_1]^\perp$ (the orthogonal space of the span of ρ_1) and A_1 as the restriction of A on E_1. A_1 is likewise a symmetric compact operator on E_1 (see Heuser [2006, Chapter 5]) and if $A_1 \neq 0$ we can conclude like above, that an eigenvalue λ_2 exists with $0 \leq |\lambda_2| = ||A_1|| \leq ||A|| = |\lambda_1|$ and an eigenfunction ρ_2 with norm 1. Proceeding likewise with E_2 etc. one obtains a sequence $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ and associated eigenfunctions $\rho_1, \rho_2, \ldots$. If we only progress as long as the eigenvalues are positive, the series is truncated if A vanishes on $E_k := [\rho_1, \ldots, \rho_k]^\perp$. We can rewrite $L_2(T) = [\rho_1, \ldots, \rho_k] \otimes E_k$ and therefore write each $f \in L_2(T)$ as

$$f = \sum_{l=1}^{k} \langle f, \rho_l\rangle \rho_l + g \quad \text{with} \quad g \in E_k$$

and

$$Af = \sum_{l=1}^{k} \lambda_l \langle f, \rho_l\rangle \rho_l.$$

If the series $(\lambda_k)_k$ does not break off, $\lambda_k \longrightarrow_{k\to\infty} 0$. Otherwise $\left(\frac{\rho_k}{\lambda_k}\right)_k$ would be bounded and therefore the sequence $(A\frac{\rho_k}{\lambda_k})_k = (\rho_k)_k$ would have a convergent subsequence. This is not possible because $||\rho_k - \rho_l|| = \sqrt{2}$ for $l \neq k$ due to the orthogonality which yields a contradiction. Furthermore, note that $g_k := f - \sum_{l=1}^{k} \langle f, \rho_l\rangle \rho_l \in E_k$. Hence

$$||Ag_k|| = ||A_k g_k|| \leq ||A_k||\,||g_k|| = |\lambda_{k+1}|\,||g_k||$$

and

$$\begin{aligned} ||g_k||^2 &= ||f||^2 - 2\sum_{l=1}^{k} \langle f, \rho_l\rangle \langle f, \rho_l\rangle + ||\sum_{l=1}^{k} \langle f, \rho_l\rangle \rho_k||^2 \\ &= ||f||^2 - \sum_{l=1}^{k} \langle f, \rho_l\rangle^2 \leq ||f||^2. \end{aligned}$$

Therewith also

$$||Ag_k|| \leq |\lambda_{k+1}|||g_k|| \leq \underbrace{|\lambda_{k+1}|}_{\to 0}\ \underbrace{||f||}_{\text{const.}}$$

follows. We can deduce that $Ag_k \longrightarrow_{l\to\infty} 0$ and $Af = \sum_{l=1}^{\infty} \lambda_k \langle f, \rho_l\rangle \rho_l$ in $L_2(T)$. Therewith most of the properties are proved. Now we can proceed to prove the Karhunen-Loève representation. Without loss of generality we assume $\mu(t) = 0$ in this proof and abbreviate $\xi_k := \xi_k(Y)$.

It is

$$\begin{aligned}\mathbb{E}\left(\int_T |Y(t)\rho_k(t)|\,dt\right)^2 &\leq \mathbb{E}\left(\int_T Y^2(t)\,dt \int_T \rho_k^2(t)\,dt\right)\\ &\leq \int_T G(t,t)\,dt \int \rho_k^2(t)\,dt < \infty,\end{aligned}$$

hence $\int_T |Y(t)\rho_k(t)|\,dt < \infty$ and

$$\xi_k = \langle Y, \rho_k\rangle = \int_T Y(t)\rho_k(t)\,dt$$

is well-defined. Further $\mathbb{E}(\xi_k^2) < \infty$, $\mathbb{E}(\xi_k) = 0$ and

$$\mathbb{E}(\xi_k, \xi_l) = \int_T \int_T \rho_k(s)G(s,t)\rho_l(t)\,ds\,dt = \lambda_k \delta_{k,l}. \tag{2.5}$$

Hence it follows that

$$\mathbb{E}(Y(t)\xi_k) = \mathbb{E}\left(Y(t)\int_T Y(s)\rho_k(s)\,ds\right) = \int_T G(s,t)\rho_k(s)\,ds = \lambda_k \rho_k(t). \tag{2.6}$$

Using (2.5) and (2.6) we obtain

$$\begin{aligned}\mathbb{E}\left(Y(t) - \sum_{l=1}^{k} \xi_l \rho_l(t)\right)^2 &= \mathbb{E}(Y^2(t)) - 2\sum_{l=1}^{k} \underbrace{\mathbb{E}(Y(t)\xi_l)}_{\lambda_l \rho_l(t)} \rho_l(t) + \underbrace{\mathbb{E}\underbrace{\left(\sum_{l=1}^{k} \xi_l \rho_l(t)\right)^2}_{\sum_{l=1}^{k} \xi_l^2 \rho_l^2(t)}}_{\sum_{l=1}^{k} \lambda_l \rho_l^2(t)}\\ &= G(t,t) - \sum_{l=1}^{k} \lambda_l \rho_l^2(t)\end{aligned}$$

for each $t \in T$ and $k \in \mathbb{N}$. According to Properties 2.6 it follows that

$$\sup_{t \in T} \mathbb{E}\left(Y(t) - \sum_{l=0}^{k} \xi_l \rho_l(t)\right)^2 \longrightarrow_{k\to\infty} 0.$$

$\square$

In practice, one uses a finite approximation of the Karhunen-Loève equation. The expansion is truncated at a value $M < \infty$ and μ and G are estimated from the data. Based on the estimated G, the eigenvalues and -functions can be estimated and ultimately the scores as well. Due to the Karhunen-Loève representation, observations can be reconstructed through the scores. Hence the scores can be used in order to compare observations. Often the first two scores are already sufficient in order to describe the observations in an adequate way. In the following, we explain how the parameters can be estimated.

2.2 Estimation

We observe realizations of a process Y as discretized functions. In order to estimate mean and covariance function and afterwards the principal components, eigenvalues and scores, it is necessary to use smoothing methods to deal with the discreteness. Further possible measurement errors of the observations and irregular data can be handled. This is why we first introduce kernel functions, which are the basis for local smoothing techniques. Afterwards we turn to the here applied smoothing technique. The next two subsections treat the one- as well as the multi-dimensional cases such that they can be applied also in the chapter about spatial FPCA.

2.2.1 Kernel functions

Kernel functions are used to apply weights to each two measurement points and they assign (in the most common cases) greater weight to small and lesser weight to greater distances of measurement points. Therefore they can be used to weight smoothing techniques appropriately. We want to define kernel functions and mention some characteristics that are important later on.

In the one-dimensional case the kernels considered here are compactly supported functions $\mathcal{K} : \mathbb{R} \to \mathbb{R}$, mostly with a support $[-1, 1]$. $\mathcal{K}$ is called a kernel of order (ν, l) with $\nu, l \in \mathbb{N}$ if

$$\int \mathcal{K}(t)t^k \, dt = \begin{cases} (-1)^\nu \nu! & k = \nu \\ 0 & 0 \le k < l, k \neq \nu \\ \neq 0 & k = l \end{cases}$$

(see e.g. Müller [1988, Chapter 4]). The first value ν refers to the derivative to be estimated. The most frequent case is $\nu = 0$ where the original function that created the observations is estimated. In Chapter 6, we will also use kernels for estimating first and second derivatives, i.e. with $\nu = 1$ or $\nu = 2$. l refers to the degree of smoothness the kernel implies. Müller [1988] carries out that higher values of l lead to better convergence rates, but also have stronger boundary effects.

A common kernel function which we apply in the following is the *Epanechnikov quadratic kernel* (see Hastie et al. [2001]):

$$\mathcal{K}(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{if} \quad |t| \le 1 \\ 0 & \text{otherwise} \end{cases},$$

which is of order $(0, 2)$. The black line in Figure 2.1 shows the kernel function.

Derivative kernels can directly be constructed out of kernels of order $(0, 2)$. If $\mathcal{K}$ is a one-dimensional kernel function of order $(0, 2)$, the derivative kernel is defined as $\mathcal{K}_t(t) = \frac{-t\mathcal{K}(t)}{\sigma_{\mathcal{K}}^2}$ with a scaling factor $\sigma_{\mathcal{K}}^2 := \int \mathcal{K}(t)t^2 \, dt$. One can easily recalculate that it has the order $(1, 3)$.

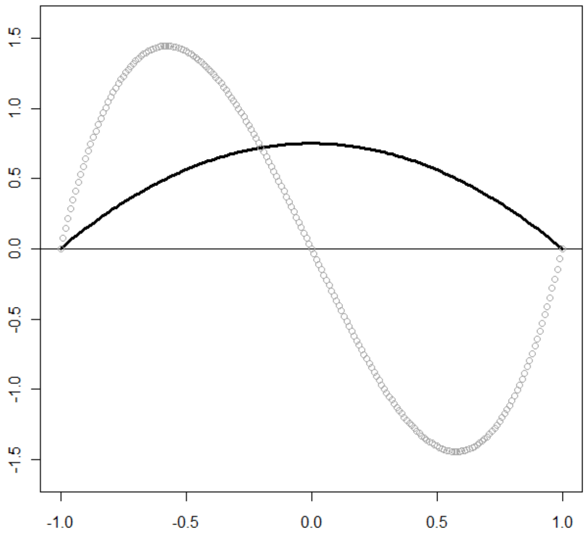


Figure 2.1: The Epanechnikov kernel function (black line) and its derivative kernel (gray dotted line).

The first derivative kernel for the Epanechnikov kernel is for example

$$\mathcal{K}(t) = \begin{cases} \frac{(-t)\frac{3}{4}(1-t^2)}{\frac{1}{5}} \\ 0 \end{cases} = \begin{cases} 3.75(t^3 - t) & \text{if} \quad |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and is plotted in Figure 2.1 as dotted gray line.

The kernel definition can also be extended to the multivariate case. A compactly supported function $\mathcal{K} : \mathbb{R}^q \to \mathbb{R}$ is a kernel of order (ν, l) with $\nu \in \mathbb{N}^q$ and $l \in \mathbb{N}$ if

$$\int \mathcal{K}(t)t^k \, dt = \begin{cases} (-1)^\nu \nu! & k = \nu \\ 0 & 0 \leq |k| < l, k \neq \nu \\ \neq 0 & \text{for at least one } k \text{ with } |k| = l \end{cases},$$

using multi index notation $|k| = k_1 + \ldots + k_q$ and $k! = k_1! \cdots k_q!$.

Multivariate kernels can be composed as product kernels of one-dimensional kernels. In this case one can calculate the order of the multivariate kernel knowing the orders of the one-dimensional kernels. For example in the case where $\mathcal{K}_j$ is a one-dimensional kernel of order

$(0,2)$ for $j = 1, \dots, q$ and $\mathcal{K}(t) = \mathcal{K}_1(t_1) \cdots \mathcal{K}_N(t_q)$, one can derive that

$$\int \mathcal{K}(t) t^k \, dt = \int \mathcal{K}_1(t_1) t_1^k \, dt_1 \dots \int \mathcal{K}_q(t_q) t_q^k \, dt_q = \begin{cases} 1 & k = 0_q \\ 0 & 0 \leq |k| < 2, k \neq 0_q \\ \neq 0 & \text{for at least one } k \text{ with } |k| = |2| \\ & \text{(i.e. } k_1 = 2, k_j = 0 \text{ for } j \neq 1) \end{cases} \tag{2.7}$$

and therefore $\mathcal{K}$ is of order $(0_q, 2)$.

Multivariate derivative kernels are constructed following the same principle as in the univariate case by forming the kernels component-by-component. For example regarding the two-dimensional product kernel $\mathcal{K}(t_1, t_2) = \mathcal{K}_1(t_1)\mathcal{K}_2(t_2)$ with $\mathcal{K}_1$ and $\mathcal{K}_2$ both of order $(0,2)$, $\mathcal{K}$ is of order $((0,0),2)$. The derivative kernel for the derivative in t_1-direction is

$$\mathcal{K}_{t_1}(t_1, t_2) = (-t_1) \frac{\mathcal{K}_1(t_1)}{\sigma_{t_1}^2} \mathcal{K}_2(t_2)$$

and is of order $((1,0),3)$.

2.2.2 Smoothing methods

Normally, data are not observed directly as functional data, but as measurements at discrete points in time or space and often are error-prone. In order to obtain functional observations which can be evaluated at each point in time, one needs to fit a function $g : T \longrightarrow \mathbb{R}$ ($T \subset \mathbb{R}$ or $T \subset \mathbb{R}^2$ in the spatial case) to the observations. If g belongs to a class of functions which varies in a finite number of parameters, we have a parametric regression model. If g is just assumed to be k times continuously differentiable, the regression model is a non-parametric one.

Later on, our aim is to apply smoothing techniques in order to derive estimates for mean and covariance function based on the raw values (see Section 2.2.3). In the following, our smoothing technique of choice is local polynomial regression. If a function value $f(t)$ shall be estimated based on data (t_l, y_l) for $l = 1 \dots, L$, the points in the neighborhood of t are used to fit a polynomial regression function in order to obtain an estimate $\hat{f}(t)$. Assume the relationship $f(t_l) = y_l + \epsilon_l$ with independent errors ϵ_l having mean 0 and variance σ^2. The neighborhood is defined by using a kernel function $\mathcal{K}$ (as defined in section 2.2.1). This kernel function is applied to the distance of t and t_l, adjusted with a bandwidth $h \in \mathbb{R}_{>0}$ which defines the extent of the neighborhood that is taken into account. Points near t are assigned a greater weight in the regression than distant points.

In order to explain the general concept, let $g(t; \beta) : \mathbb{R} \times \mathbb{R}^{d+1} \longrightarrow \mathbb{R}$ (with $\beta = (\beta_0, \dots, \beta_d)$ and $d \in \mathbb{N}$) be a polynomial of small degree, i.e. $g(s) = \sum_{j=0}^d \beta_j s^j$. Then the local least squares estimate of f based on the data (t_l, y_l) for $l = 1 \dots, L$ is given by $\hat{f}(t) = g(0; \hat{\beta}^t) = \hat{\beta}_0^t$ with

$$\hat{\beta}^t \in \arg \min_{\beta \in \mathbb{R}^{d+1}} \sum_{l=1}^L \mathcal{K}\left(\frac{t - t_l}{h}\right) [y_l - g(t - t_l; \beta)]^2 \tag{2.8}$$

for any fixed $t \in \mathbb{R}$ and a smoothing bandwidth h. Further one obtains an estimation for the derivative of f via $\hat{f}'(t) = \beta_1^t$.

The most common case is $d = 0$, where $g(t;\beta) = \beta_0$ is constant. In this case the local estimator is simply a weighted mean of the original data. This estimator is also called Nadaraya-Watson estimator. For $d = 1$, $g(t,\beta) = \beta_0 + \beta_1 t$ is a linear function and for $d = 2$ a parable. Higher orders of d are only scarcely used. The choice of d mainly influences boundary effects, which are always critical in the estimation of functions.

If the data are in $\mathbb{R}^q$ (i.e. the data can be written as $(t_{l_1}, \ldots, t_{l_q}, y_l)$), one can proceed analogously with a kernel function $\mathcal{K} : \mathbb{R}^q \longrightarrow \mathbb{R}_{\geq 0}$ and a multidimensional polynomial of order d: $g(s;\beta) : \mathbb{R}^q \times \mathbb{R}^{D(d,q)} \longrightarrow \mathbb{R}$. The actual number of parameters $D(d)$ is a function of d and q. For example in the linear case $D(d,q) = q+1$ and $g(s;\beta) = \beta_0 + \beta_1 s_1 + \ldots + \beta_q s_q$. The calculation can be expressed through

$$\hat{\beta}^t \in \arg\min_{\beta \in \mathbb{R}^d} \sum_{l=1}^{L} \mathcal{K}\left(\frac{t_1 - t_{l_1}}{h_1}, \cdots, \frac{t_q - t_{l_q}}{h_q}\right) [y_l - g(t_1 - t_{l_1}, ..., t_q - t_{l_q}; \beta)]^2 \qquad (2.9)$$

with bandwidths $h_1, \ldots, h_q \in \mathbb{R}_{>0}$. The estimator is again given by $\hat{f}(t) = g(0; \hat{\beta}^t) = \hat{\beta}_0^t$.

The regression smoothing technique with $d \geq 1$ compared to the standard Nadaraya-Watson kernel-weighted average reduces bias at the boundaries. Using a polynomial of order two instead of one increases variance a lot and is mainly of use if curvature in the interior shall be estimated with only a small bias (see Hastie et al. [2001]).

In the context of functional data analysis the data points (t_l, y_l) could be the discrete measurements of one subject. In this case the result is a smoothed version of the observation. Otherwise the (t_l, y_l) could be the pooled data of several subjects or pooled raw covariances values of each two subjects. Then the mean respectively covariance function over all observations is estimated. The precise explanation of these cases is given in the next section.

Under some regularity conditions, one can show the uniform convergence of this kind of estimators. In our situation this is carried out in Chapter 6. For the general case please refer to Fan and Gijbels [1996].

2.2.3 Estimation of moments

For estimating the FPCs, we first need an estimation of mean and covariance function. Before the estimators are introduced, the general stochastic framework shall be defined. Our main assumption is that we have independent observations of a process over time (or another continuous one-dimensional variable). Further we allow additional measurement errors to occur as long as they show no correlations.

Let X be a stochastic process in $L_2(T)$ and the trajectory X^i of the ith subject an independent identical copy of X. To incorporate measurement errors (see Yao et al. [2005]), let Y^i be a noisy observation made of X^i, such that the nth observation of Y^i made at time t_n^i (for $n = 1, \ldots, N_i$) can be written as

$$Y_n^i = X^i(t_n^i) + \epsilon_n^i,$$

wherein $\mathbb{E}(\epsilon_n^i) = 0$ and $\mathrm{Var}(\epsilon_n^i) = \sigma^2$. The errors ϵ_n^i are assumed to be iid for all points in time and all subjects.

We assume that the mean and covariance function

$$\mu(t) = \mathbb{E}(X(t)) \quad \text{and} \quad G(t_1, t_2) = \mathrm{Cov}(X(t_1), X(t_2))$$

exist.

According to the theory of section 2.1, the covariance function G has an orthogonal expansion

$$G(t_1, t_2) = \sum_{k=0}^{\infty} \lambda_k \rho_k(t_1) \rho_k(t_2), t_1, t_2 \in T$$

with eigenfunctions ρ_k and non-increasing eigenvalues λ_k ($\sum_k \lambda_k < \infty, \lambda_1 \geq \lambda_2 \geq ... \geq 0$) and the X_i can be represented as

$$X_i(t) = \mu(t) + \sum_{k=0}^{\infty} \xi_{ik} \rho_k(t), t \in T,$$

with uncorrelated random variables ξ_{ik} with mean 0 and variance λ_k.

For the calculation of principal components based on a data set we follow the nonparametric approach of Yao et al. [2005]. They use local linear regression for estimating the mean function and two-dimensional local polynomial fitting for the covariance function. The eigenfunctions and PCA scores are estimated via discrete approximations.

The eigenfunctions shall be evaluated at equally spaced points in time $t_{n'} (n' = 1, \dots, N')$, not necessarily the same as the measurement points t_n^i, which are potentially irregular.

Estimation of the mean function In order to estimate the mean function, all single measurement values are pooled. A polynomial of degree one is used for smoothing via local regression. Therefore (2.8) leads to minimizing

$$\sum_{i=1}^{I} \sum_{n=1}^{N_i} \mathcal{K}_1 \left(\frac{t_n^i - t}{h_1} \right) \left[Y_n^i - \beta_0 - \beta_1 (t - t_n^i) \right]^2 \tag{2.10}$$

with respect to β_0 and β_1 in order to obtain $\hat{\beta}_0^t$ and $\hat{\beta}_1^t$ for $t \in T$. $\mathcal{K}_1$ is a kernel function from $\mathbb{R}$ to $\mathbb{R}$ and $h_1 \in \mathbb{R}_{>0}$ its bandwidth. The estimated mean function is then given by $\hat{\mu}(t) = \hat{\beta}_0^t$. $\hat{\beta}_1^t$ is further an estimation for the first derivative of μ.

Estimation of the covariance function For estimating the covariance function we first calculate the raw covariances

$$G_i(t_{n_1}^i, t_{n_2}^i) = (Y_{n_1}^i - \hat{\mu}(t_{n_1}^i))(Y_{n_2}^i - \hat{\mu}(t_{n_2}^i)) \tag{2.11}$$

and afterwards minimize according to (2.9)

$$\sum_{i=1}^{I} \sum_{1 \leq n_1 \neq n_2 \leq N_i} \mathcal{K}_2 \left(\frac{t_{n_1}^i - t_1}{h_2}, \frac{t_{n_2}^i - t_2}{h_2} \right) \left[G_i(t_{n_1}^i, t_{n_2}^i) - \beta_0 - \beta_{11}(t_1 - t_{n_1}^i) - \beta_{12}(t_2 - t_{n_2}^i) \right]^2$$

with respect to β_0, β_{11} and β_{12} to obtain $\hat{\beta}_0^{t_1,t_2}$, $\hat{\beta}_{11}^{t_1,t_2}$ and $\hat{\beta}_{12}^{t_1,t_2}$ for $(t_1,t_2) \in T^2$. $\mathcal{K}_2$ is a kernel function from $\mathbb{R}^2$ to $\mathbb{R}$ and $h_2 \in \mathbb{R}_{>0}$ its bandwidth. This provides the estimation for the non-diagonal values via $\hat{G}(t_1,t_2) = \hat{\beta}_0^{t_1,t_2}$. Further $\hat{\beta}_{11}^{t_1,t_2}$ is an estimator for $G_{t_1}(t_1,t_2)$, i.e. for the derivative of G according to t_1, and $\hat{\beta}_{12}^{t_1,t_2}$ an estimator for $G_{t_2}(t_1,t_2)$. Diagonal elements are excluded in the estimation, because the raw observations have larger errors than the non-diagonal elements:[1]

$$\begin{aligned} G_i(t^i_{n_1}, t^i_{n_2}) &= (X^i(t^i_{n_1}) + \epsilon^i_{n_1} - \hat{\mu}(t^i_{n_1}))(X^i(t^i_{n_2}) + \epsilon^i_{n_2} - \hat{\mu}(t^i_{n_2})) \\ &\approx \operatorname{Cov}(X(t^i_{n_1}), X(t^i_{n_2})) + \sigma^2 \delta_{n_1 n_2} \end{aligned}$$

In order to obtain estimations for the diagonal elements, we fit a local quadratic component orthogonal to the diagonal because the covariance is maximal on the diagonal (see also Yao et al. [2005]). To achieve this, first rotate the values by 45 degrees via

$$\begin{pmatrix} t^{i*}_{n_1} \\ t^{i*}_{n_2} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ -\sqrt{2} & \sqrt{2} \end{pmatrix} \begin{pmatrix} t^i_{n_1} \\ t^i_{n_2} \end{pmatrix}$$

and afterwards minimize

$$\sum_{i=1}^{I} \sum_{1 \le n_1 \ne n_2 \le N_i} \mathcal{K}_2 \left(\frac{t^{i*}_{n_1} - t_1}{h_2}, \frac{t^{i*}_{n_2} - t_2}{h_2} \right) \left[G_i(t^{i*}_{n_1}, t^{i*}_{n_2}) - \gamma_0 - \gamma_1 (t_1 - t^{i*}_{n_1}) - \gamma_2 (t_2 - t^{i*}_{n_2})^2 \right]^2$$

with respect to γ_0, γ_1 and γ_2. Then set

$$\tilde{G}(t_1, t_2) := \gamma_0(t_1, t_2)$$

and

$$\hat{G}(t_1, t_1) = \tilde{G}(0, \sqrt{2}\, t_1).$$

which finally is the estimation for the diagonal elements.

2.2.4 Estimation of functional principal components

Given the estimate $\hat{G}$ for the covariance function as derived in the section before, the estimates of the eigenfunctions and eigenvalues are given by the solutions $\hat{\rho}_k$ and $\hat{\lambda}_k$ of the eigenequations

$$\int_T \hat{G}(s,t) \rho_k(s)\, ds = \lambda_k \rho_k(t). \tag{2.12}$$

This is still a functional operator-eigenvalue problem. Hence we reduce the problem to a multivariate matrix-eigenvalue problem. Approximated eigenfunctions $\hat{\rho}_k$ and eigenvalues $\hat{\lambda}_k$ are obtained by calculating the multivariate eigenvalues and eigenvectors of the discretized covariance function. As the evaluation points $t_{n'}$ as defined on p. 18 are equally spaced, let

[1] $\delta(a,b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$

$\Delta_t = t_2 - t_1$ be their distance. The discretized covariance is then given by $\tilde{G}(t_{n'_1}, t_{n'_2}) := \hat{G}(t_{n'_1}, t_{n'_2})\Delta_t$ for $n'_1, n'_2 = 1, \dots, N'$ and the (multivariate) eigenvectors of $\tilde{G}$ can be calculated solving the equation

$$\tilde{G}\tilde{\rho}_k = \hat{\lambda}_k \tilde{\rho}_k$$

for $\tilde{\rho}_k \in \mathbb{R}^{N'}$, $\tilde{\rho}_k \neq 0_{N'}$ and $\hat{\lambda}_k \in \mathbb{R}$. The norm of $\tilde{\rho}_k$ in L_2 is approximated via

$$\sqrt{\sum_{n'=1}^{N'} \tilde{\rho}_k^2(t_{n'})\Delta_t}.$$

Therefore the discretized standardized eigenfunctions are given by

$$\hat{\rho}_k(t_{n'}) = \tilde{\rho}_k(t_{n'}) \left(\sqrt{\sum_{n'=1}^{N'} \tilde{\rho}_k^2(t_{n'})\Delta_t} \right)^{-1}$$

to have norm one.

In order to control the (arbitrary) direction of the eigenfunctions we further demand the majority of signs of the discrete eigenvector to be positive.

Apart from the eigenvalues one is often interested in the share of variance explained by each principal component. Therefore we often regard the variance components defined as

$$\mathrm{VC}(\lambda_k) := \frac{\lambda_k}{\sum_{l=1}^{\infty} \lambda_l},$$

often expressed in %. This value is estimated by

$$\widehat{\mathrm{VC}}(\hat{\lambda}_k) = \frac{\hat{\lambda}_k}{\sum_{l=1}^{M} \hat{\lambda}_l}.$$

In practice M is being chosen such that $\hat{\lambda}_M$ is (almost) zero and usually higher than the number K of eigenfunctions used for representations.

Now that the basis system existing of the functional principal components is estimated, we can proceed identifying each observation in the new basis system which means estimating the FPC scores. The FPC scores are classically estimated by numerical integration of

$$\hat{\xi}_{ik} = \int_T (Y^i(t) - \hat{\mu}(t))\hat{\rho}_k(t)\,dt \approx \sum_{n'=1}^{N'} (Y^i_{n'} - \hat{\mu}(t_{n'}))\hat{\rho}_k(t_{n'})\Delta_t. \tag{2.13}$$

If the measurements have additional errors (remember that we assumed $Y^i_n = X^i(t^i_n) + \epsilon^i_n$), the FPC scores are generally estimated too high. Therefore Yao et al. [2003] propose to use so-called shrinkage estimates instead:

$$\tilde{\xi}_{ik} = \frac{\hat{\lambda}_k}{\hat{\lambda}_k + \frac{|T|\hat{\sigma}^2}{N'}} \hat{\xi}_{ik}, \tag{2.14}$$

in which $|T|$ is the length of the interval T. In order to build the shrinkage estimates it is necessary to estimate the error variance $\hat{\sigma}^2$. This is done by comparing the *error-free* estimation $\hat{G}(t,t)$ for $t \in T$ with an estimation $\hat{V}(t)$ for $V(t) = G(t,t) + \sigma^2$ which one obtains by using local polynomial smoothing like in (2.10) based on diagonal estimation with $G_i(t_n^i, t_n^i)$ (see (2.11)) as input. In order to get a stable estimate, only the interval $T_1 = [\min(T) + |T|/4, \max(T) - |T|/4]$ is taken as input. The estimation is as follows:

$$\hat{\sigma}^2 = \frac{2}{|T|} \int_{T_1} \left(\hat{V}(t) - \hat{G}(t,t) \right) \, dt$$

if $\hat{\sigma}^2 > 0$ and $\hat{\sigma}^2 = 0$ otherwise (see Yao et al. [2005]). The last case could occur if the measurements are not error-prone.

Furthermore Yao et al. [2005] extend the theory to sparse estimations and propose conditional estimates for the principal components as the best way to estimate them in this context under normal assumptions. This method is necessary if the number of measurements per subject is too small in order to obtain a good approximation of the integral in (2.13).

As the data mainly in our focus are not sparse, we implemented the estimation of FPCs by numeric integration and give the shrinkage estimation method as an additional option. Like in the sparse case, our algorithms are also suitable to incorporate missing data or data not measured on a regular grid.

With the estimated eigenfunctions $\hat{\rho}_k$ for $k = 1, \dots, K$ and the scores $\hat{\xi}_{ik}$, observation i can now be represented through

$$\hat{Y}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\rho}_k(t) \tag{2.15}$$

which is the finite version of the Karhunen-Loève representation (see Theorem 3).

One can judge the accuracy of this representation for example by evaluating the following error measure via numerical integration:

$$||\hat{Y}_i - Y_i||_2 = \sqrt{\int |\hat{Y}_i(t) - Y_i(t)|^2 \, dt}$$

A first application of this estimation is given in Section 2.3, but before we want to say a word regarding choice of parameters.

2.2.5 Choice of bandwidth and number of eigenfunctions

Studies revealed that the crucial parameter in kernel regression is the choice of the bandwidth h (see e.g. Hastie et al. [2001]). A small bandwidth leads to high variance, a large bandwidth to high bias, so that it is not trivial to find the right trade-off. In order to find the suitable bandwidth, many data-driven methods were developed. Above all, cross validation is very popular.

The idea behind cross validation is to use a training data set in order to fit a method and to use an independent test data set to verify the method respectively to calculate a prediction error. If the method is dependent on more than one parameter, as in our case, the bandwidths h_1 for the mean and h_2 for the covariance estimation, the prediction error can be calculated for multiple values of h_1 and h_2. The combination with the smallest prediction error is the combination of choice.

For the general cross validation method (Hastie et al. [2001]) define an indexing function

$$\kappa : \{1, \dots, I\} \Rightarrow \{1, \dots, G\}$$

which partitions the observations in G approximately equally sized parts. For each observation $i \in 1, \dots, I$ the FPCA is calculated leaving out all data belonging to part $\kappa(i)$. Let the resulting mean function be $\hat{\mu}^{-\kappa(i)}$ and the eigenfunctions $\hat{\rho}_k^{-\kappa(i)}$. The scores of observation i are calculated based on (2.13):

$$\hat{\xi}_{ik}^{-\kappa(i)} = \int_T (Y^i(t) - \hat{\mu}^{-\kappa(i)}(t))\hat{\rho}_k^{-\kappa(i)}(t)\,dt$$

With $\hat{Y}_i^{-\kappa(i)}$ calculated like in (2.15) based on the reduced data set, we can now calculate the prediction error:

$$CV(h_1, h_2) = \frac{1}{I}\sum_{i=1}^{I} ||Y_i - \hat{Y}_i^{-\kappa(i)}||_2 \qquad (2.16)$$

This is typically done for several combinations of h_1 and h_2 and both are chosen to minimize (2.16).

The method with $G = I$ is called *leave-one-out cross validation*. Other typical values for G, dependent on the available amount of data, are $G = 2$ (one training and test data set), $G = 5$ or $G = 10$.

The optimal number of eigenfunctions can also be determined through cross validation. In both cases cross validation helps to avoid overfitting, e.g. fitting noise characteristics of the data which cannot be found again in another data set, if the data are independent.

According to Ramsay and Silverman [2006] for most practical problems it is sufficient to evaluate the bandwidth graphically. As this method of FPCA estimation is relatively time-consuming depending on the number of measurements, it is sometimes no possibility to use cross validation. In Chapter 6, the consistency rate for general kernel regression estimators are calculated depending on the bandwidth. The choice of the kernel function itself has only a small influence on the results.

In the case of the number of eigenfunctions, we say that we require the eigenfunctions to explain a certain amount of variation in the data and skip all eigenfunctions which only have a minor influence.

2.2.6 Overview of consistency results

As we are interested in the quality of our estimation, we show in Chapter 6 under certain assumptions the uniform consistency of the estimators and calculate convergence rates. At this point we give a short overview and refer the reader to Chapter 6 for technical details. For those not familiar to the Landau symbolism we use to notate the rates, please consider the short introduction in Section 6.1. The calculations are based on the assumption that the measurement points are equally spaced and that the bandwidths h_1 and h_2 are appropriately chosen dependent on the number of observations I and the number of points in time N. Exactly we require for the bandwidth for mean smoothing h_1 that

$$h_1 \to 0, Ih_1^2 \to \infty, Ih_1^6 = O(1), \frac{1}{N} = O(h_1^5) \quad \text{for} \quad I \to \infty$$

and accordingly for the bandwidth for covariance smoothing h_2:

$$h_2 \to 0, Ih_2^4 \to \infty, Ih_2^8 = O(1), \frac{1}{N} = O(h_2^7) \quad \text{for} \quad I \to \infty.$$

For the mean and covariance function we obtain the following uniform consistency rates in Theorems 9 and 10:

$$\sup_{t \in T} |\hat{\mu}(t) - \mu(t)| = O_P\left(\frac{1}{\sqrt{I}h_1}\right)$$

$$\sup_{t_1, t_2 \in T} |\hat{G}(t_1, t_2) - G(t_1, t_2)| = O_P\left(\frac{1}{\sqrt{I}h_2^2}\right)$$

For the proof of these rates, the estimators for mean and covariance are written explicitly and each of them is split into a Nadaraya-Watson estimator corrected with derivative estimators. A lemma tells how to evaluate the single components such that a final rate can be deduced from the single rates for each component.

Directly from the rate for the covariance estimation the rates for eigenvalue and eigenfunction estimation can be derived for a fixed $k \in \mathbb{N}$. The rates for the eigenfunctions are only achievable, if λ_k has the multiplicity 1 (Theorem 12):

$$|\hat{\lambda}_k - \lambda_k| = O_P\left(\frac{1}{\sqrt{I}h_2^2}\right)$$

$$||\hat{\rho}_k - \rho_k|| = O_P\left(\frac{1}{\sqrt{I}h_2^2}\right)$$

$$\sup_{t \in T} |\hat{\rho}_k(t) - \rho_k(t)| = O_P\left(\frac{1}{\sqrt{I}h_2^2}\right).$$

That means that we have a convergence in the L_2 sense as well as a uniform convergence rate for the eigenfunctions. For the derivation of the rate for the eigenvalues Weyl's criterion is used and in the case of eigenfunctions a general eigenanalysis theory for Hilbert-Schmidt operators. The detailed proofs are carried out in Section 6.2.

2.3 Example: Wiener process

In order to give a first example of the FPCA calculation and at the same time evaluate our approach we construct a stochastic process with a simple structure such that its eigenvalues and eigenfunctions can be calculated theoretically.

This property is fulfilled by a Wiener process which is a stochastic process with independent normally distributed increments (see e.g. Csörgö and Révész [1981]). A disadvantage of this process is that its sum representation converges slowly due to the independent increment condition. Thereby many principal components are needed to represent the process accurately through a finite sum of principal components. For this process we compare the approximated eigenfunctions calculated accordingly to Section 2.2 with the theoretic ones.

Definition 3 (Wiener process)**.** A stochastic process $\{W(t)|\, 0 \leq t < \infty\}$ in $\mathbb{R}$ is called *Wiener process* if

1. $W(t) - W(s) \sim \mathcal{N}(0, t-s)$ for $0 \leq s < t < \infty$ and $W(0) = 0$

2. $W(t)$ is a process with independent increments, i.e.

$$W(t_2) - W(t_1), W(t_4) - W(t_3), \dots, W(t_{2i}) - W(t_{2i-1})$$

   are independent for all $0 \leq t_1 < t_2 \leq t_3 < t_4 \leq \ldots \leq t_{2i-1} < t_{2i} < \infty$ and all $i \in \mathbb{N}$.

3. The path $t \rightarrow W(t, \omega)$ is continuous in t with probability one.

Directly from the definition one can derive the following properties:

$$\mu(t) := \mathbb{E}(W(t)) = 0 \quad \text{for all} \quad 0 \leq t < \infty$$

$$G(s,t) := \mathrm{Cov}(W(s), W(t)) = \min(s,t)$$

With some calculational effort one can obtain a representation of $W(t)$ as an infinite sum of random variables (see Csörgö and Révész [1981] for details):

$$W(t) = X_0 t + \sqrt{2} \sum_{l=1}^{\infty} X_l \frac{\sin l\pi t}{l\pi} \tag{2.17}$$

wherein X_l are independently $\mathcal{N}(0,1)$ distributed random variables for $l \in \mathbb{N}$.

Like the eigenfunction expansion this is an orthogonal expansion, but it does not have the optimality property of the eigenfunction expansion in the amount of variance explained by a certain number of components (compare Definition 1).

Next, we will calculate the eigenvalues and eigenfunctions of the Wiener process on the interval $[0,1]$. According to Theorem 1, it is necessary to find all pairs of scalars $\lambda_k \in \mathbb{R}$ and nonzero functions $\rho_k : [0,1] \longrightarrow \mathbb{R}$ for $k \in \mathbb{N}$ which solve the equation

$$\langle G(\cdot, s), \rho \rangle = \lambda \rho$$

under the conditions $\rho_k(0) = 0$, $\int_0^1 \rho^2(t)\, dt = 1$ and $\int_0^1 \rho_k^2(t) \rho_l^2(t)\, dt = 0$ for $k \neq l$. A short calculation shows that this leads to a differential equation:

$$\begin{aligned}
&\langle G(\cdot,s),\rho\rangle = \lambda\rho(s)\\
\Leftrightarrow\quad &\int_0^1 \min(s,t)\rho(t)\,dt = \lambda\rho(s)\\
\Leftrightarrow\quad &\int_0^s t\rho(t)\,dt + \int_s^1 s\rho(t)\,dt = \lambda\rho(s) \qquad (2.18)\\
\Rightarrow\quad &\frac{\partial}{\partial s}\int_0^s t\rho(t)\,dt + \frac{\partial}{\partial s}s\int_s^1 \rho(t)\,dt = \lambda\frac{\partial}{\partial s}\rho(s)\\
\Rightarrow\quad &s\rho(s) + \int_s^1 \rho(t)\,dt - s\rho(s) = \lambda\frac{\partial}{\partial s}\rho(s)\\
\Rightarrow\quad &\frac{\partial}{\partial s}\int_s^1 \rho(t)\,dt = \lambda\frac{\partial^2}{\partial^2 s}\rho(s)\\
\Rightarrow\quad &\lambda\frac{\partial^2}{\partial^2 s}\rho(s) = -\rho(s)
\end{aligned}$$

This differential equation has the solution

$$\rho(s) = c\sin\left(\frac{s}{\sqrt{\lambda}}\right)$$

with a constant $c \in \mathbb{R}$. The constant can be determined through the scaling condition:

$$\begin{aligned}
&\int_0^1 \rho^2(t)\,dt = 1\\
\Leftrightarrow\quad &\int_0^1 \sin^2\left(\frac{t}{\sqrt{\lambda}}\right)\,dt = \frac{1}{c^2}\\
\Leftrightarrow\quad &c = \sqrt{2} \quad \text{or} \quad x = -\sqrt{2}
\end{aligned}$$

In order to determine λ we substitute the solution into the equation (2.18) and obtain

$$\int_s^1 sc\sin\left(\frac{t}{\sqrt{\lambda}}\right)\,dt + \int_0^s tc\sin\left(\frac{t}{\sqrt{\lambda}}\right)\,dt = \lambda c\sin\left(\frac{t}{\sqrt{\lambda}}\right).$$

Solving this equation for λ leads to the set of possible solutions $\left\{ \frac{4}{\pi^2(2k-1)^2} \,\middle|\, k \in \mathbb{N} \right\}$ for λ. Hence the pairs of eigenvalues and eigenfunctions of the Wiener process are

$$\lambda_k = \frac{4}{\pi^2(2k-1)^2} \quad \text{and} \quad \rho_k(s) = \sqrt{2}\sin\left(\left(k - \frac{1}{2}\right)\pi s\right) \quad \text{for} \quad k \in \mathbb{N}. \qquad (2.19)$$

Next we want to test the algorithm described in Section 2.2 by applying it to realizations of a Wiener process. One can either use the sum representation (2.17) to simulate outcomes of the Wiener process or else successively according to Definition 3, which we do in the following:

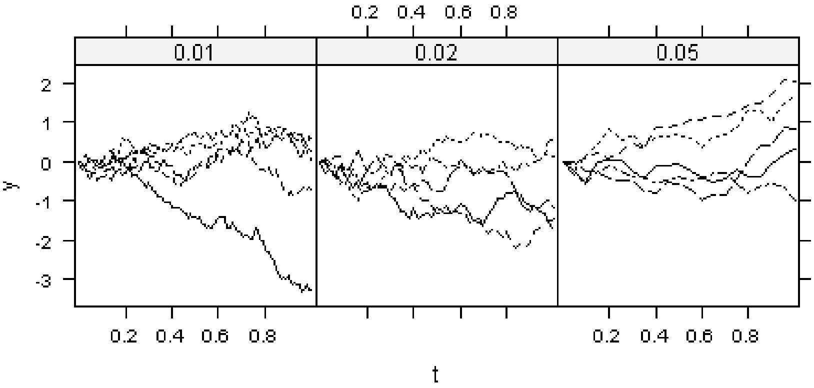


Figure 2.2: Wiener processes with the step widths $\frac{1}{N}$ indicated above the plots. Each graph shows five realizations.

Method 1 (Step by step simulation of a Wiener process)**.** The step by step simulation needs the step width $\frac{1}{N}$ as the only parameter.
Here $x_1, \ldots, x_N$ are N normally distributed random numbers with mean 0 and variance $\frac{1}{N}$. The Wiener process is simulated via

$$\begin{aligned} w(0) &= 0 \\ w\left(\frac{1}{N}\right) &= x_1 \\ &\vdots \\ w\left(\frac{n}{N}\right) &= \sum_{i=1}^{n} x_i = w\left(\frac{n-1}{N}\right) + x_n \quad \text{for} \quad n = 1, \ldots, N. \end{aligned}$$

$\big(w(0), w\left(\frac{1}{N}\right), \ldots, w(1)\big)$ is an approximated realization of a Wiener process at 0, $\frac{1}{N}$, $\frac{2}{N}$, ..., 1.

Empirical findings of the estimation of eigenfunctions depend on the bandwidth h used for smoothing mean and covariance function (here the same bandwidth is used for both), on the number of steps N and on the number of simulated observations I.
As the sign of the eigenfunctions is arbitrary, we standardize it by requesting the greater part of the function being positive. In practice it is done by demanding more positive than negative signs in the discretized eigenfunctions.
Figure 2.2 shows simulated realizations of the Wiener process for different step widths $\frac{1}{N}$, Figure 2.3 an example for an estimated covariance function and Figure 2.4 an example for estimated versus true eigenfunctions. We observe that the factor combination presented here seems to estimate the eigenfunctions quite accurate. In order to analyze the influence of the three parameters, we performed a small simulation study. Each parameter was varied on three levels and the results were judged by calculating the mean absolute deviation between true and estimated eigenfunction.

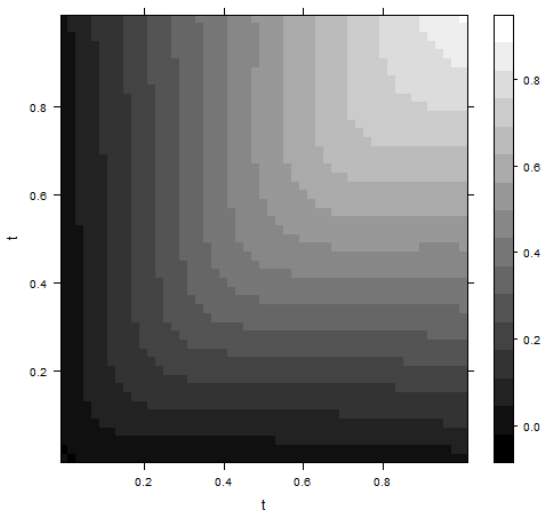


Figure 2.3: Estimated covariance function of a simulated Wiener processes with step width 0.02, bandwidth 0.2 and 50 realizations.

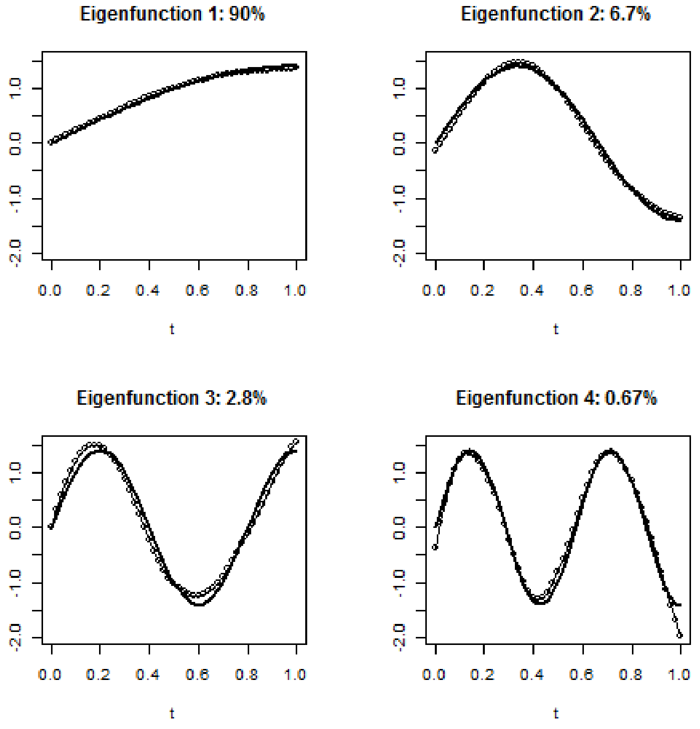


Figure 2.4: Estimated (lines with dots) and true (lines) eigenfunctions of a simulated Wiener processes with the same parameters as Figure 2.3.

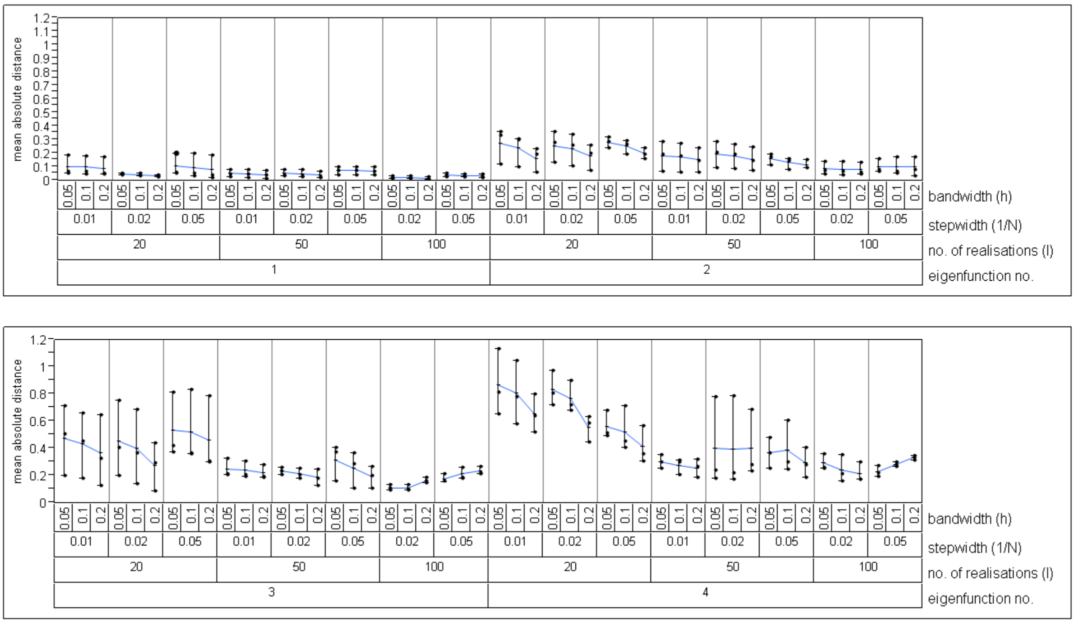


Figure 2.5: Influence of bandwidth, step width and number of realizations on the mean absolute deviation between true and estimated eigenfunctions.

Figure 2.5 shows the results of three repeated simulations for each factor combination. One can see that the estimation generally gets worse for eigenfunctions of larger order. Furthermore the estimation accuracy is higher when increasing the bandwidth or the number of realizations in the tested ranges. The step width has no clear influence. This is consistent with the results of Section 6.2, where the order of convergence of the estimation is evaluated theoretically.

2.4 Clustering based on functional principal components

In multivariate analysis it is common to group observations into clusters by using methods like K-means clustering or hierarchical clustering (the methods are explained in the application in Section 4.2). If many variables are observed, it is sometimes necessary to conduct a dimension reduction before the clustering procedure. If principal component analysis is used, the clustering is done on the resulting scores. In functional data analysis it is not trivial to use cluster analysis directly on the observations as they are not necessarily made at equally spaced points in time and also not at the same points in time for each observation. Therefore it is very convenient to carry out cluster analysis based on the FPC scores. We will demonstrate this clustering in the applications later on.

Chiou and Li [2007] have pushed this concept further and developed a method, the K-centers functional clustering method, which promises to be an enhancement of FPCA and shall lead to a better description of the single clusters. This method shall be shortly described in the next subsection.

2.4.1 K-centers functional clustering

KCFC (K-centers functional clustering) is a method which calculates principal components for each cluster and re-sorts each element to that cluster where it can be best approximated via the first principal components.

Assumptions for this method are that

- the process Y is a mixed process of subprocesses $Y^{(c)}$ in $L^2(T)$.
- Each subprocess is associated with one cluster.
- A random variable $\mathcal{C}$ on $\{1, \dots, C\}$ describes the cluster membership.

Above the common mean one considers here the marginal mean and covariance functions of the subprocesses which are denoted as follows:

$$\mu^{(c)}(t) = \mathbb{E}(Y(t) \,|\, C = c) \quad \text{and} \quad G^{(c)}(s,t) = \mathrm{Cov}(Y(s), Y(t) \,|\, C = c).$$

It is further assumed that each subprocess has a Karhunen-Loève expansion

$$Y^{(c)}(t) = \mu^{(c)}(t) + \sum_{k=1}^{\infty} \xi_k^{(c)}(Y)\rho_k^{(c)}(t) \tag{2.20}$$

with $< G^{(c)}(\cdot, t), \rho_k^{(c)} >= \lambda_k^{(c)} \rho_k^{(c)}(t)$ and $\xi_k^{(c)}(Y) =< Y - \mu^{(c)}, \rho_k^{(c)} >$.

For applications it is necessary to truncate the infinite sum in (2.20) after K_c summands. This leads to the finite presentation

$$\tilde{Y}^{(c)}(t) = \mu^{(c)}(t) + \sum_{k=1}^{K_c} \xi_k^{(c)}(Y)\rho_k^{(c)}(t).$$

In order to prevent additional distribution assumptions, the cluster membership for a observation y is determined through

$$c^*(y) = \arg \min_{c \in \{1,\dots,C\}} ||y - \tilde{y}^{(c)}||,$$

in which $\tilde{y}^{(c)}$ is the finite Karhunen-Loève expansion of y in cluster c. Hence, one determines in which cluster the observation can be represented with the smallest error.

Based on the aforementioned assumptions the clustering procedure works as follows:
As before, let I be the number of observations and N_i the number of measurements for observation i. Further t_n^i is the time point of measurement n in observation i and $y_i(t)$ the ith observed curve at time t.
For clustering one needs at first estimations of the moments, the eigenfunctions, eigenvalues and the FPC scores. They are achieved by the methods explained in Section 2.2.
The KCFC algorithm is based on an initial cluster assignment on the FPC scores $(\hat{\xi}_{i1}, \dots, \hat{\xi}_{iK})$, which can for example be reached through a standard clustering procedure like K-means clustering with K being the number of principal components taken into account.
After the initial clustering is finished, the algorithm works as follows:
Let $g_i^{(l)} \in \{1, \dots, C\}$ be the cluster membership of the ith observation in the lth iteration and $G^{(l)} = \{g_i^{(l)} | i = 1, \dots, I\}$ the set of cluster memberships.

1. Choose $i \in \{1, \dots, I\}$ and calculate $\hat{\mu}_{(-i)}^{(c)}$ and $\hat{\rho}_{(-i)}^{(c)}$ based on all observations with $g_j^{(l)} = c$ and $j \neq i$.

2. Calculate the ith predicted observation for each cluster c via

$$\hat{y}_{(i)}^{(c)}(t) = \hat{\mu}_{(-i)}^{(c)}(t) + \sum_{k=1}^{K_c} \hat{\xi}_{ik}^{(c)} \hat{\rho}_{k(-i)}^{(c)}(t)$$

   with $\xi_{ik}^{(c)} = \int_T [y_i(t) - \hat{\mu}_{(-i)}^{(c)}(t)] \hat{\rho}_{k(-i)}^{(c)} \, dt$.

3. Assign observation i to cluster

$$g_i^{(l+1)} = \arg \min_{c=1,\dots,C} ||y_i - \hat{y}_{(i)}^{(c)}||.$$

   Set $G^{(l+1)} = \{g_i^{(l+1)} | i = 1, \dots, I\}$.

4. Repeat steps 1-3 until no reclassification occurs anymore.

The value K_c in step 2 is determined for each cluster individually and is recalculated in each iteration. For details please refer to Chiou and Li [2007].

An advantage of this method in comparison to simple K-means clustering based on the FPC scores is that it takes different distributions of the single clusters into account. But the price for the greater precision is that much data are needed for executing multiple FPC analyses in single clusters, which makes it unfeasible in many practical applications.

3 Spatial Functional Principal Component Analysis

Up to now we treated the one-dimensional functional principal component analysis method (FPCA). The theoretic part was explained, estimators were derived and their application to a Wiener process was conducted. In this chapter we want to generalize the previous results to the case of data measured on two-dimensional domains.

Many sources treating one-dimensional FPCA mention the possibility to extend the method to spatial data (for example Yao et al. [2005] or Ramsay and Silverman [2006]), but do not carry out the exact way it is done. Furthermore, software implementations are only written for the one-dimensional case up to our knowledge. In this section, we will explain the framework of spatial FPCA and carry out all details concerning the nonparametric estimation of spatial functional principal components. Moreover, convergence rates for the estimators are deduced and an implementation for this method is accomplished, described in detail in Section 7.2.

The reader is advised to consider Chapter 2 before and focus on the changes that occur while extending the theory. Topics that are very similar to the one-dimensional case will not be treated again as thorough as in Chapter 2.

3.1 Theory

In the one-dimensional case we considered a Hilbert space of square integrable functions over a bounded interval in $\mathbb{R}$ and looked at eigenvalues and eigenfunctions in that Hilbert space. In this chapter we want to extend our theory to a spatial setting in order to analyze data measured on a two-dimensional interval, i.e. a rectangle, in $\mathbb{R}^2$. Therefore we consider the Hilbert space of square integrable functions over a bounded rectangle $T \times \mathcal{T} \subset \mathbb{R}^2$:

$$L^2(T \times \mathcal{T}) = \left\{ f : T \times \mathcal{T} \longrightarrow \mathbb{R} \,\middle|\, \left| \int_T \int_{\mathcal{T}} f(t,\tau)^2 dt\, d\tau \right| < \infty \right\}$$

This space is a Hilbert space with the inner product $< f, g >= \int_T \int_{\mathcal{T}} f(t,\tau) g(t,\tau) dt\, d\tau$ and the norm $|| \cdot || = \sqrt{< \cdot, \cdot >}$ analogous to the one-dimensional case. Again, to be precise, only the space of equivalence classes of functions in $L^2(T \times \mathcal{T})$ which are identical up to a null set forms a Hilbert space with the defined norm. This issue will be neglected in the following presentation.

In the spatial case we consider a stochastic process $\{Y(t,\tau); t \in T, \tau \in \mathcal{T}\}$. The mean and covariance function are here given by

$$\mu(t,\tau) = \mathbb{E}(Y(t,\tau)) \quad \text{respectively} \quad G(t_1,t_2,\tau_1,\tau_2) = \mathrm{Cov}(Y(t_1,\tau_1), Y(t_2,\tau_2))$$

for $t, t_1, t_2 \in T$ and $\tau, \tau_1, \tau_2 \in \mathcal{T}$. It is assumed that μ and G exist, i.e. that

$$\left| \int_T \int_{\mathcal{T}} \mathbb{E}(Y^2(t, \tau, \omega)) \, dt \, d\tau \right| < \infty$$

The covariance operator A (compare to (2.3)) is in the two-dimensional case an operator from $L^2(T \times \mathcal{T}) \longrightarrow \mathbb{R}$ with

$$(Af)(t_1, \tau_1) = \int_T \int_{\mathcal{T}} f(t_2, \tau_2) G(t_1, t_2, \tau_1, \tau_2) \, dt_2 \, d\tau_2$$

for $f \in L^2(T \times \mathcal{T})$ and $(t_1, \tau_1) \in T \times \mathcal{T}$.
As the situation is principally the same as in the multivariate or one-dimensional functional case, the eigenfunctions and eigenvalues are defined in an equivalent way:

Definition 4 (Eigenfunctions and eigenvalues)**.** If a function $\rho \in L^2(T \times \mathcal{T}), \rho \neq 0$ and a constant $\lambda \in \mathbb{C}$ fulfill

$$A\rho = \lambda\rho \quad \Leftrightarrow \quad < G(\cdot, t_2, \cdot, \tau_2), \rho >= \lambda\rho(t_2, \tau_2) \quad \text{for all } t_2 \in T, \tau_2 \in \mathcal{T},$$

λ is called eigenvalue with eigenfunction ρ of G.

The properties of eigenvalues and eigenfunctions mentioned and proved in Properties 2 correspond absolutely analogously in the spatial case, such that we do not repeat them here one-by-one. We again obtain a countable set of real, non-negative eigenvalues that can be ordered as $\lambda_1 > \lambda_2 > ... \geq 0$ with corresponding eigenfunctions $\rho_1, \rho_2, \ldots$. This is due to the symmetry of the positive-semidefinite covariance function G. In the spatial case the eigenspaces are likewise finite-dimensional. Furthermore, the process can be represented with its eigenfunctions as

$$Af = \sum_{k=1}^{\infty} \lambda_k \langle f, \rho_k \rangle \rho_k$$

in $L_2(T \times \mathcal{T})$ for $f \in L_2(T \times \mathcal{T})$.
The Karhunen-Loève theorem (Theorem 3) can also be transferred:

Theorem 5 (Spatial Karhunen-Loève expansion)**.** *If Y is a square integrable process in space with continuous covariance function G, the process has a Karhunen-Loève expansion*

$$Y(t, \tau) = \mu(t, \tau) + \sum_{k=1}^{\infty} \xi_k(Y) \rho_k(t, \tau).$$

The coefficients $\xi_k(Y)$ are given via $\xi_k(Y) =< Y - \mu, \rho_k >$ and are called scores. *The scores ξ_k for $k \in \mathbb{N}$ are uncorrelated random values with zero mean and variance λ_k.*

Hence the theoretic framework is analogous to the one-dimensional case. The proofs of Section 2.1 mainly use properties of the norm which are directly applicable here as are the other steps. Though the theory is similar, nevertheless the estimation of the spatial functional principal components is more complex due to the high dimensionality, especially of the covariance function. This is treated in the next section.

3.2 Estimation

The estimation of the spatial functional principal components is performed equivalently to the one-dimensional case. Mean and covariance functions are estimated again through local polynomial regression, only in higher dimensions. For estimation of the eigenfunctions the covariance function is discretized and re-sorted into a two-dimensional structure such that the eigenfunctions can be calculated like in the one-dimensional case. Afterwards the now one-dimensional eigenfunctions are restructured again to spatial functions. The exact procedure is carried out in the following.

For the explanation of kernel functions and the smoothing method please refer to Sections 2.2.1 and 2.2.2. As for the covariance estimation in the one-dimensional case already two-dimensional kernels were necessary, multi-dimensional kernels were already treated there.

3.2.1 Estimation of moments

Before the estimators for mean and covariance function can be stated, the general framework has to be explained.

Let X be a stochastic process in $L_2(T \times \mathcal{T})$ and the trajectory X^i an independent identical copy of X. We assume having observed error-prone realizations Y_i for $i = 1, \dots, I$ for this process. The measurements can generally be made irregularly at different points in space for the subjects i, so we name the space points with (t_n^i, τ_n^i) with $i = 1, \dots, I$ and $n = 1, \dots, N_i$ being the number of measurements per subject.

Like in the one-dimensional case we incorporate measurement errors ϵ_n^i which are iid for all subjects i and all points in space (t_n^i, τ_n^i). The measurement error has an expected value $\mathbb{E}(\epsilon_n^i) = 0$ and variance $\mathbb{E}(\epsilon_n^i) = \sigma^2$. The observations can be written as

$$Y_n^i = X^i(t_n^i, \tau_n^i) + \epsilon_n^i$$

for $i = 1, \dots, I$, $n = 1, \dots, N_i$.

For the covariance function $G(t_1, t_2, \tau_1, \tau_2) = \mathrm{Cov}(X(t_1), X(t_2), X(\tau_1), X(\tau_2))$ we assume that it has the orthogonal expansion

$$G(t_1, t_2, \tau_1, \tau_2) = \sum_{k=1}^{\infty} \lambda_k \rho_k(t_1, \tau_1) \rho_k(t_2, \tau_2)$$

for $t_1, t_2 \in T, \tau_1, \tau_2 \in \mathcal{T}$ with eigenfunctions ρ_k and non-increasing eigenvalues λ_k.

We further require that the X_i can be represented as

$$X_i(t, \tau) = \mu(t, \tau) + \sum_{k=1}^{\infty} \xi_{ik} \rho_k(t, \tau) \quad \text{for } t \in T, \tau \in \mathcal{T},$$

with uncorrelated random variables ξ_{ik} with mean 0 and variance $\mathbb{E}(\xi_{ik}^2) = \lambda_k$.

The observations can be made at different and irregular points in space for the different subjects and also missing values are allowed, as the following estimation procedures are designed to incorporate irregularity. The estimated functions are calculated for points in space $(t_{n'}, \tau_{m'})$ on a regular equidistant grid with $n' = 1, \dots, N'$ and $m' = 1, \dots, M'$. The distances are denoted by $\Delta t'$ respectively $\Delta\tau'$.

Estimation of the mean function The mean function is again estimated by minimizing

$$\sum_{i=1}^{I}\sum_{n=1}^{N_i}\mathcal{K}_2\left(\frac{t_n^i-t}{h_{2_t}},\frac{\tau_n^i-\tau}{h_{2_\tau}}\right)\left[Y_n^i-\beta_0-\beta_{11}(t-t_n^i)-\beta_{12}(\tau-\tau_n^i)\right]^2 \tag{3.1}$$

with respect to β_0, β_{11} and β_{12} to obtain $\hat{\beta}_0^{t,\tau}$, $\hat{\beta}_{11}^{t,\tau}$ and $\hat{\beta}_{12}^{t,\tau}$ for $(t,\tau)\in T\times\mathcal{T}$. $\mathcal{K}_2$ is a kernel function from $\mathbb{R}^2$ to $\mathbb{R}$ with the bandwidths $h_{2_t}, h_{2_\tau}\in\mathbb{R}_{>0}$. The estimated mean function is then given by $\hat{\mu}(t,\tau)=\hat{\beta}_0^{t,\tau}$. Furthermore $\hat{\beta}_{11}^{t,\tau}$ is an estimation for $\mu_t(t,\tau)$, i.e. the derivative of μ in t-direction and $\hat{\beta}_{12}^{t,\tau}$ for $\mu_\tau(t,\tau)$.

Estimation of the covariance function For the estimation of the four-dimensional covariance function, we consider again the raw covariances of each subject, that are given by

$$G_i(t_{n_1}^i,t_{n_2}^i,\tau_{n_1}^i,\tau_{n_2}^i)=(Y_{n_1}^i-\hat{\mu}(t_{n_1}^i,\tau_{n_1}^i))(Y_{n_2}^i-\hat{\mu}(t_{n_2}^i,\tau_{n_2}^i))$$

for $i=1,\dots,I$ and $n_1,n_2=1,\dots,N_i$.
Afterwards the smoothed function can be obtained by minimizing

$$\sum_{i=1}^{I}\sum_{\substack{n_1,n_2\\ n_1\neq n_2}}\mathcal{K}_4\left(\frac{t_{n_1}^i-t_1}{h_{4_t}},\frac{t_{n_2}^i-t_2}{h_{4_t}},\frac{\tau_{n_1}^i-\tau_1}{h_{4_\tau}},\frac{\tau_{n_2}^i-\tau_2}{h_{4_\tau}}\right) \tag{3.2}$$

$$\times\left[G_i(t_{n_1}^i,t_{n_2}^i,\tau_{n_1}^i,\tau_{n_2}^i)-\beta_0-\beta_{11}(t_1-t_{n_1}^i)-\beta_{12}(t_2-t_{n_2}^i)-\beta_{21}(\tau_1-\tau_{n_1}^i)-\beta_{22}(\tau_2-\tau_{n_2}^i)\right]^2$$

with respect to β_0, β_{11}, β_{12}, β_{21} and β_{22} in order to obtain $\hat{\beta}_0^{t_1,t_2,\tau_1,\tau_2}$ etc. for $(t_1,t_2,\tau_1,\tau_2)\in T^2\times\mathcal{T}^2$. The diagonal line is excluded again due to the measurement error (see explanation for one-dimensional covariance estimation in Section 2.2.3). $\mathcal{K}_4$ is a kernel function from $\mathbb{R}^4$ to $\mathbb{R}$ and $h_{4_t}, h_{4_\tau}\in\mathbb{R}_{>0}$ the estimation bandwidths.
As the method is already computationally very intense, we do not estimate the diagonal line in a different way as was done in the one-dimensional case.

3.2.2 Estimation of functional principal components

Like in equation (2.12) we have to find solutions $\hat{\lambda}_k\in\mathbb{R}$ and $\hat{\rho}_k: T\times\mathcal{T}\to\mathbb{R}$ to the equation

$$\int_T\int_{\mathcal{T}}\hat{G}(t_1,t_2,\tau_1,\tau_2)\rho(t_1,\tau_1)\,dt_1\,d\tau_1=\hat{\lambda}\rho(t_2,\tau_2)$$

Now consider the discretized version (expressed on the regular grid as defined in the foregoing section):

$$\sum_{n_1'=1}^{N'}\sum_{m_1'=1}^{M'}\hat{G}(t_{n_1'},t_{n_2'},\tau_{m_1'},\tau_{m_2'})\rho_k(t_{n_1'},\tau_{m_1'})\Delta_{t'}\Delta_{\tau'}=\lambda_k\rho_k(t_{n_2'},\tau_{m_2'})$$

We solve this equation via re-sorting $\tilde{G} := \hat{G}\Delta_t\Delta_\tau$ according to

$$G^* = \begin{pmatrix} \tilde{G}_{1111} & \tilde{G}_{2111} & \cdots & \tilde{G}_{N111} & \tilde{G}_{1121} & \cdots & \tilde{G}_{N121} & \cdots & \tilde{G}_{N1M1} \\ \tilde{G}_{1211} & \tilde{G}_{2211} & & & & & & & \\ \vdots & & \ddots & & & & & & \\ \tilde{G}_{1N11} & & & \tilde{G}_{NN11} & & & & & \\ \tilde{G}_{1112} & & & & \tilde{G}_{1122} & & & & \vdots \\ \vdots & & & & & \ddots & & & \\ \tilde{G}_{1N12} & & & & & & \tilde{G}_{NN22} & & \\ \vdots & & & & & & & \ddots & \\ \tilde{G}_{1N1M} & & & & \cdots & & & & \tilde{G}_{NNMM} \end{pmatrix}$$

with $\tilde{G}_{n_1n_2m_1m_2} := \tilde{G}(t_{n'_1}, t_{n'_2}, \tau_{m'_1}, \tau_{m'_2})$.

Then the eigenfunctions can be estimated like in the one-dimensional case by solving

$$G^*\rho^* = \lambda\rho^*$$

with

$$\rho^* = \begin{pmatrix} \rho_{11} \\ \rho_{21} \\ \vdots \\ \rho_{N1} \\ \rho_{12} \\ \vdots \\ \rho_{N2} \\ \vdots \\ \rho_{NM} \end{pmatrix}$$

for $\rho^* \in \mathbb{R}^{NM}$ and $\lambda \in \mathbb{R}$. The resulting estimated eigenvalues $\hat{\lambda}_k \in \mathbb{R}$ are also estimations of eigenvalues for the initial problem and the discretized eigenfunction estimations $\rho_k^* \in \mathbb{R}^{NM}$ can be reshaped back to their two-dimensional form (called here $\tilde{\rho}_k$) according to the scheme

$$\tilde{\rho} = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1M} \\ \vdots & & \vdots \\ \rho_{N1} & \cdots & \rho_{NM} \end{pmatrix}.$$

The estimations still need to be standardized according to the right norm. This is done by

$$\hat{\rho}_k(t_{n'}, \tau_{m'}) := \tilde{\rho}_k \left(\sqrt{\sum_{n'=1}^{N'} \sum_{m'=1}^{M'} \tilde{\rho}_k^2(t_{n'}, \tau_{m'})\Delta_{t'}\Delta_{\tau'}} \right)^{-1}$$

Finally the scores are estimated via

$$\begin{aligned}\hat{\xi}_{ik} &= \int_T \int_{\mathcal{T}} (Y^i(t,\tau) - \hat{\mu}(t,\tau))\hat{\rho}_k(t,\tau)\,dt\,d\tau \\ &\approx \sum_{n'=1}^{N'} \sum_{m'=1}^{M'} (Y^i_{n',m'} - \hat{\mu}(t_{n'},\tau_{m'}))\hat{\rho}_k(t_{n'},\tau_{m'})\Delta t' \Delta \tau'.\end{aligned}$$

For this calculation the observations Y^i need to be available at the evaluation points $(t_{n'}, \tau_{m'})$, which is not generally the case. If the measurement points do not comprise the evaluation points, the evaluation is done by linear interpolation.

The calculation of shrinkage estimates as in (2.14) could be conducted likewise in the two-dimensional case by estimating the error variance $\hat{\sigma}$ accordingly and adjust the scaling parameters to the two-dimensional case.

The considerations about bandwidth and number of eigenfunctions of Section 2.2.5 apply also in the two-dimensional case, hence we do not repeat them here.

The computational effort of equation (3.2) is growing fast with the number of data points per observation. This method is in the two-dimensional case only feasible if the number of data points is not too high. For other cases we further implemented a method that omits the mean and covariance smoothing and instead smoothes the single observations. Smoothing of single observations is essentially done like the estimation of the mean function in (3.1), only for each function individually. The advantage is that the whole FPCA estimation can be performed on the discretized observations and therefore like in the multivariate case, only the discretized observations have to be re-sorted from matrices into vectors and the calculated vectorized eigenfunctions have to be re-sorted back into matrices and standardized appropriately. If necessary, they can also be smoothed.

Regarding computational cost, this approach is preferable and if the measurements per observation are sufficiently dense, it is regarding our experience no problem to do so. But if the measured data are too sparse or if many measurements are missing, such that the single observations cannot be estimated satisfyingly for themselves, we have to use the computationally more intense approach presented here. All theoretic results in this thesis are based on the method with the smoothed covariance.

3.2.3 Overview of consistency results

Like in the one-dimensional case we calculated consistency results under certain assumptions which can be found in detail in Section 6.3. h_1 denotes here the bandwidth for mean estimation and h_2 the bandwidth for covariance estimation (we assume for the proofs equal bandwidths in both, the t and τ direction), whereas I as always denotes the number of observations. The number of points in space N and M are here assumed to be identical for all observations I and fulfill $M \geq N$. Additionally the measurements are taken at regular distances in both directions. According to Section 6.3, the parameters have to fulfill the following relationships:

$$h_1 \to 0, Ih_1^4 \to \infty, Ih_1^8 = O(1), \frac{1}{N} = O(h_1^7), M \geq N \quad \text{for} \quad I \to \infty$$

for h_1 and

$$h_2 \to 0, Ih_2^6 \to \infty, Ih_2^{12} = O(1), \frac{1}{N} = O(h_2^9), M \geq N \quad \text{for} \quad I \to \infty$$

for h_2.

For the mean and covariance function we obtain in the spatial case these consistency rates (see Theorems 15 and 16:)[1]:

$$\sup_{\substack{t \in T \\ \tau \in \mathcal{T}}} |\hat{\mu}(t,\tau) - \mu(t,\tau)| = O_P\left(\frac{1}{\sqrt{I}h_1^2}\right)$$

$$\sup_{\substack{\vec{t} \in T^2 \\ \vec{\tau} \in \mathcal{T}^2}} |\hat{G}(\vec{t},\vec{\tau}) - G(\vec{t},\vec{\tau})| = O_P\left(\frac{1}{\sqrt{I}h_2^4}\right)$$

Directly from these rates the rates for eigenvalue and eigenfunction estimation can be derived for a fixed $k \in \mathbb{N}$ (Theorem 12):

$$|\hat{\lambda}_k - \lambda_k| = O_P\left(\frac{1}{\sqrt{I}h_2^4}\right) \tag{3.3}$$

$$||\hat{\rho}_k - \rho_k|| = O_P\left(\frac{1}{\sqrt{I}h_2^4}\right) \tag{3.4}$$

$$\sup_{\substack{t \in T \\ \tau \in \mathcal{T}}} |\hat{\rho}_k(t,\tau) - \rho_k(t,\tau)| = O_P\left(\frac{1}{\sqrt{I}h_2^4}\right) \tag{3.5}$$

We obtain a result for the convergence in the L_2 sense as well as a uniform convergence rate for the eigenfunctions. Like in the one-dimensional case we need the assumption that he eigenspaces are one-dimensional for deriving a consistency rate for the eigenfunctions. Otherwise, for eigenspaces of greater dimensions, only rates for projections on the eigenspace can be given. The reason is that in this case the eigenfunctions themselves are not unambiguously defined. A demonstration of this problem is given in Section 3.3.

In comparison of these results and the one-dimensional results in Section 2.2.6 one can see that the number of dimensions directly influence the h-powers in the denominator.

3.3 Example: Spatial Wiener process

Like in the one-dimensional case we want to use a Wiener process in order to verify the FPC calculation. The two-dimensional Wiener process is like its one-dimensional equivalent a stochastic process with independent increments only that in this case the increments occur in two directions, which leads apart from the additional dimension to a crucial difference as we will see later on. The process is defined as follows (see Csörgö and Révész [1981]).

Definition 5 (Wiener process)**.** Let $\{W(t,\tau)|\,(t,\tau) \in \mathbb{R}^2_{>0}\}$ be a two-dimensional stochastic

[1] In the following, we will from time to time use the abbreviations $\vec{t} = (t_1, t_2)$ and $\vec{\tau} = (\tau_1, \tau_2)$.

process. For a rectangle $R = [t_1, t_2) \times [\tau_1, \tau_2) \subset \mathbb{R}^2_{>0}$ with $0 \leq t_1 < t_2 < \infty$ and $0 \leq \tau_1 < \tau_2 < \infty$ we define

$$W(R) := W(t_2, \tau_2) - W(t_1, \tau_2) - W(t_2, \tau_1) + W(t_1, \tau_1)$$

Then $\{W(t, \tau) | \, (t, \tau) \in \mathbb{R}^2_{>0}\}$ is called a two-dimensional Wiener process if

1. $W(R) \sim \mathcal{N}(0, A(R))$ for all $R = [t_1, t_2) \times [\tau_1, \tau_2)$ with $A(R) := (t_2 - t_1)(\tau_2 - \tau_1)$.

2. $W((0, \tau)) = W((t, 0)) = 0$ for all $t, \tau \in \mathbb{R}_{\geq 0}$

3. $W(t, \tau)$ has independent increments, e.g. $W(R_1), W(R_2), \ldots, W(R_n)$ are independent for disjoint rectangles $R_1, \ldots, R_n$.

4. The path $(t, \tau) \to W((t, \tau), \omega)$ is continuous in (t, τ) with probability one.

Analogous to the one-dimensional case the covariance function has again a simple structure. We can derive that

$$G(\vec{t}, \vec{\tau}) := \mathrm{Cov}(W(t_1, \tau_1), W(t_2, \tau_2)) = \min(t_1, t_2) \min(\tau_1, \tau_2)$$

for $\vec{t} = (t_1, t_2)$ and $\vec{\tau} = (\tau_1, \tau_2)$.
In order to calculate the eigenvalues and eigenfunctions of the spatial Wiener process we have to find all pairs of scalars λ and nonzero functions ρ which solve the eigenequation

$$\langle G(\cdot, t_2; \cdot, \tau_2), \rho \rangle = \lambda \rho(t_2, \tau_2) \tag{3.6}$$

$$\Leftrightarrow \int \int G(\vec{t}; \vec{\tau}) \rho(t_1, \tau_1) \, dt_1 d\tau_1 = \lambda \rho(t_2, \tau_2)$$

$$\Leftrightarrow \int \int \min(t_1, t_2) \min(\tau_1, \tau_2) \rho(t_1, \tau_1) \, dt_1 d\tau_1 = \lambda \rho(t_2, \tau_2)$$

As the covariance functions can be split into separate multiplicands for each dimension and those are the same as in the one-dimensional case (see calculation (2.18)), one can assume that ρ can also be separated into multiplicands for each dimension:

$$\rho(t, \tau) = \rho_1(t) \rho_2(\tau)$$

Therefore we obtain solutions for (3.6) by solving the following equations:

$$\int \min(t_1, t_2) \rho_1(t_1) \, dt_1 = \lambda_1 \rho_1(t_2)$$

$$\int \min(\tau_1, \tau_2) \rho_2(\tau_1) \, d\tau_1 = \lambda_2 \rho_2(\tau_2)$$

with $\lambda_1 \lambda_2 = \lambda$.

As the reduced equations correspond to the one-dimensional case, the solutions are (see (2.19)):

$$\lambda_{1,j} = \frac{4}{\pi^2(2j-1)^2} \qquad \text{and} \quad \rho_{1,j}(t) = \sqrt{2}\sin\left(\left(j - \frac{1}{2}\right)\pi t\right) \qquad \text{for} \quad j \in \mathbb{N},$$

$$\lambda_{2,l} = \frac{4}{\pi^2(2l-1)^2} \qquad \text{and} \quad \rho_{2,l}(\tau) = \sqrt{2}\sin\left(\left(l - \frac{1}{2}\right)\pi \tau\right) \qquad \text{for} \quad l \in \mathbb{N},$$

so that we obtain as solutions for the initial equation:

$$\lambda_{j,l} = \frac{16}{\pi^4(2j-1)^2(2l-1)^2}$$

$$\text{and} \quad \rho_{j,l}(t,\tau) = 2\sin\left(\left(j - \frac{1}{2}\right)\pi t\right)\sin\left(\left(l - \frac{1}{2}\right)\pi \tau\right) \quad \text{for} \quad j,l \in \mathbb{N}.$$

Hence the eigenvalues and eigenfunctions are simply all possible compositions of the one-dimensional eigenvalues and eigenfunctions. As there is a kind of symmetry, for example $\lambda_{j,l} = \lambda_{l,j}$, the eigenspaces are not all one-dimensional, but eigenspaces of dimension two and more occur which is the crucial difference we spoke of in the beginning of this section.

For example the first five sorted eigenvalues, which we index here with k^*, of the spatial Wiener process are shown in table 3.1.

unique eigenvalue no. (k^*)	dimension of eigenspace	M	j	l
1	1	1	1	1
2	2	9	1	2
			2	1
3	2	25	1	3
			3	1
4	2	49	1	4
			4	1
5	3	81	1	5
			5	1
			2	2

Table 3.1: The first five eigenvalues of the spatial Wiener process with the dimensions of the eigenspaces. It is $\lambda_{k^*} = \frac{16}{\pi^4 M}$ using the abbreviation $M = (2j-1)^2(2l-1)^2$.

This table shows that the first eigenvalue has a one-dimensional eigenspace, but eigenvalues number 2 to 4 have two-dimensional spaces, eigenvalue 5 even a three-dimensional space.

Like in the one-dimensional analysis we want to simulate outcomes of a Wiener process in order to compare theoretic and estimated principal components. The step by step simulation method described in method 1 is easy to transfer to two dimensions. The Wiener process shall be simulated at the points in space $\left(\frac{n}{N}, \frac{m}{N}\right)$ for $n, m \in \{0, \ldots, N\}$.

Method 2 (Step by step simulation of a spatial Wiener process)**.** Let x_{nm} with $n, m \in \{1, \ldots, N\}$ be N^2 normally distributed random numbers with mean 0 and variance $\frac{1}{N^2}$ and

$$w\left(\frac{n}{N}, 0\right) = 0 \quad \text{for} \quad n \in \{0, \ldots, N\}$$

$$w\left(0,\frac{m}{N}\right)=0 \quad \text{for} \quad m\in\{1,\dots,N\}$$

$$w\left(\frac{n}{N},\frac{m}{N}\right)=\sum_{n'=1}^{n}\sum_{m'=1}^{m}x_{n'm'}=w\left(\frac{n-1}{N},\frac{m}{N}\right)+w\left(\frac{n}{N},\frac{m-1}{N}\right)-w\left(\frac{n-1}{N},\frac{m-1}{N}\right)+x_{nm}$$

for $\quad n,m\in\{1,\dots,N\}$.

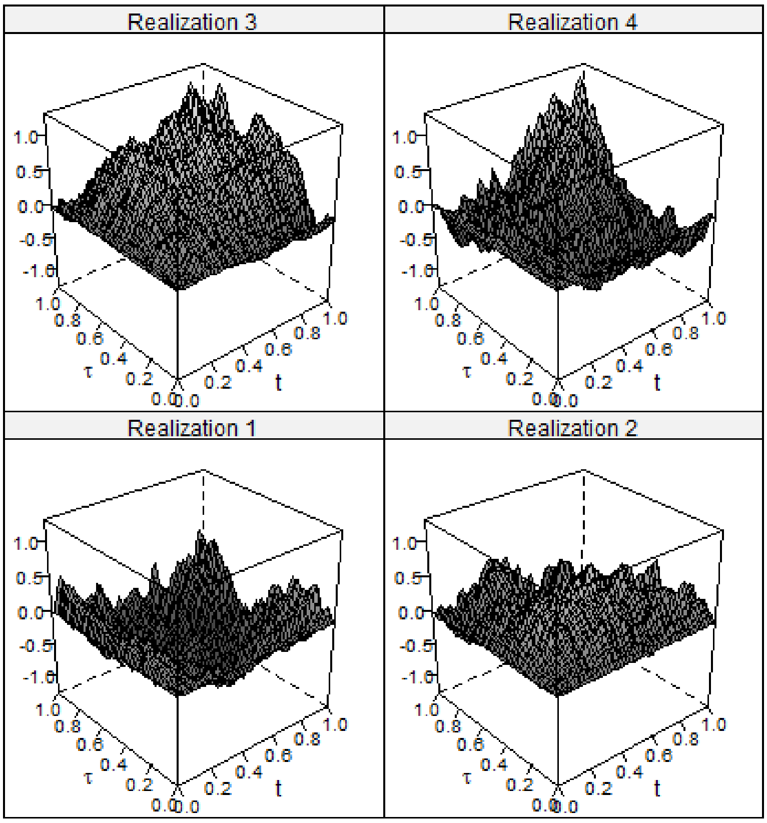


Figure 3.1: Realizations of a spatial Wiener process with a step width of 0.02.

In order to test the two-dimensional FPCA estimation, we simulated 200 realizations of the two-dimensional Wiener process with a step width of 0.02. Figure 3.1 shows examples for realizations. As the simulated data set is large and the estimation method with smoothed mean and covariance smoothing is not feasible in this case, we used here the alternative described at the end of subsection 3.2.2 with smoothing of observations and eigenfunctions. Therefore, the Epanechnikov kernel with a bandwidth of 0.2 was used.

The comparison of theoretic and estimated eigenfunctions gets more complicated than in the one-dimensional case. For the first eigenfunction, we can still compare the estimated and theoretic eigenfunction as before. The result is plotted in Figure 3.2.

According to Table 3.1 the second (theoretic) eigenvalue has a two-dimensional eigenspace. This causes the problem that the eigenfunctions cannot be unambiguously defined. In general in a multi-dimensional eigenspace, if ρ_1 and ρ_2 are orthogonal eigenfunctions of the same eigenvalue

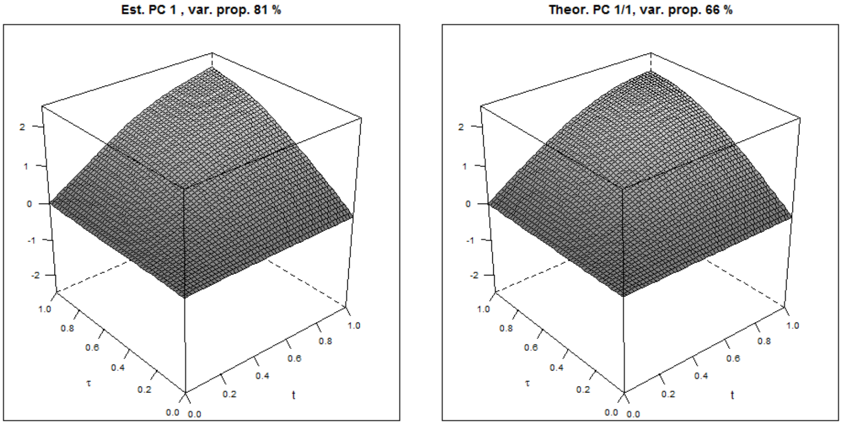


Figure 3.2: Estimated and true eigenfunctions for the first eigenvalue.

λ, each standardized linear combination $r\rho_1 + (1-r)\rho_2$ $(r \in [0,1])$ is also an eigenfunction. Therefore we most likely do not obtain two estimated eigenfunctions corresponding to the second theoretic eigenvalue λ_{2^*}, but in the best case a linear combination of them. Hence we cannot compare eigenfunction and eigenfunction, but have to make a detour via projections. Furthermore, the probability is zero that we obtain two identical eigenvalues in the estimated case and therefore we do not have multi-dimensional eigenspaces in that case, exactly spoken. Nonetheless we can identify the estimated eigenfunctions $\hat{\rho}_2$ and $\hat{\rho}_3$ with the theoretic eigenfunctions $\rho_{1,2}$ and $\rho_{2,1}$ respectively $\hat{\rho}_4$ and $\hat{\rho}_5$ with $\rho_{1,3}$ and $\rho_{3,1}$ and speak in this context about the estimated and true second and third eigenspaces.

What we did in regard of the comparison is to represent the estimated eigenfunctions in the basis of the theoretic eigenfunctions and vice versa the theoretic eigenfunctions in the basis of the estimated ones. Tables 3.2 and 3.3 show the resulting scores.

For Table 3.2 we calculated the scores according to

$$\xi_{j,l}(\hat{\rho}_k) = \langle \hat{\rho}_k, \rho_{j,l} \rangle. \tag{3.7}$$

The mean function of the Wiener process is zero, hence it is not subtracted here.

Regarding Table 3.2 we can see that $\hat{\rho}_1$ is almost completely represented by $\rho_{1,1}$ and that $\hat{\rho}_2$ and $\hat{\rho}_3$ are linear combinations (respectively rotated versions) of $\rho_{1,2}$ and $\rho_{2,1}$. In the ideal case the sum of the scores $\xi_{1,2}(\hat{\rho}_2)$ and $\xi_{2,1}(\hat{\rho}_2)$ would be one. In the same way $\hat{\rho}_4$ and $\hat{\rho}_5$ can be compared with $\rho_{1,3}$ and $\rho_{3,1}$.

To push this comparison further, we plotted in Figure 3.3 the estimated eigenfunctions $\hat{\rho}_2$ and $\hat{\rho}_3$ together with their representations $\tilde{\hat{\rho}}_2$ and $\tilde{\hat{\rho}}_3$ in the second theoretic eigenspace, e.g.

$$\tilde{\hat{\rho}}_2 = \xi_{1,2}(\hat{\rho}_2)\rho_{1,2} + \xi_{2,1}(\hat{\rho}_2)\rho_{2,1} \quad \text{and}$$

$$\tilde{\hat{\rho}}_3 = \xi_{1,2}(\hat{\rho}_3)\rho_{1,2} + \xi_{2,1}(\hat{\rho}_3)\rho_{2,1}$$

as well as the same for the third and fourth estimated eigenfunctions in Figure 3.5. The figures support the results of Table 3.2 that the representation works very well.

Scores (est. by theor.)	j	l	Est. PC 1	Est. PC 2	Est. PC 3	Est. PC 4	Est. PC 5
Theor. PC 1	1	1	1.01	-0.02	-0.01	0.06	0.05
Theor. PC 2	1	2	-0.03	0.90	-0.32	-0.01	0.00
	2	1	-0.03	0.32	0.89	0.08	0.03
Theor. PC 3	1	3	0.03	-0.09	-0.04	0.84	-0.14
	3	1	0.02	-0.04	-0.11	0.13	0.83

Table 3.2: Scores of representing the estimated eigenfunctions in the associated theoretic eigenspace. The scores are calculated as described in (3.7).

Scores (theor. by est.)	j	l	Est. PC 1	Est. PC 2	Est. PC 3	Est. PC 4	Est. PC 5
Theor. PC 1	1	1	1.05	-0.03	-0.02	0.06	0.07
Theor. PC 2	1	2	0.01	0.89	-0.34	-0.01	0.01
	2	1	0.01	0.31	0.88	0.08	0.04
Theor. PC 3	1	3	0.07	-0.09	-0.05	0.84	-0.12
	3	1	0.06	-0.05	-0.12	0.13	0.84

Table 3.3: Scores of representing the estimated eigenfunctions in the associated theoretic eigenspace. The scores are calculated as described in (3.8).

For the other direction (representing the theoretic eigenfunctions through the estimated eigenfunctions), the results are given in Table 3.3 and in Figures 3.4 and 3.4. Here the scores are calculated as

$$\hat{\xi}_k(\rho_{j,l}) = \langle \rho_{j,l} - \hat{\mu}, \hat{\rho}_k \rangle. \tag{3.8}$$

The estimated mean function is not exactly zero, hence we have to consider it in the calculations for this direction. Hence, the representations $\tilde{\rho}_{1,2}$ and $\tilde{\rho}_{2,1}$ in the estimated eigenspace for the theoretic eigenfunctions $\rho_{1,2}$ and $\rho_{1,2}$ are calculated according to

$$\tilde{\rho}_{1,2} = \hat{\mu} + \hat{\xi}_2(\rho_{1,2})\hat{\rho}_2 + \hat{\xi}_3(\rho_{1,2})\hat{\rho}_3 \quad \text{and}$$

$$\tilde{\rho}_{2,1} = \hat{\mu} + \hat{\xi}_2(\rho_{2,1})\hat{\rho}_2 + \hat{\xi}_3(\rho_{2,1})\hat{\rho}_3.$$

Again the analog calculations are performed for $\tilde{\rho}_{1,3}$ and $\tilde{\rho}_{3,1}$ using the estimated eigenfunctions $\hat{\rho}_4$ and $\hat{\rho}_5$.

The results confirm the good correspondence of theoretic and estimated eigenspace.

The estimation was also tested using down to 20 realizations and a step width of 0.5. This is a more realistic framework for actual applications. In this case, still the general form of the eigenfunctions to the first and second eigenvalues are recognizable. Furthermore, we successfully performed the full smoothing method on the smaller data set and therefore also have a verification of this method.

To summarize, we verified in this section the estimation of spatial principal components and showed how the case of multi-dimensional eigenspaces can be handled for comparisons through concentrating on the projections on that eigenspace. The calculation method of the theoretic principal components done in this section is generally applicable for processes with a symmetry of dimensions. Hence all these processes have multi-dimensional eigenspaces.

Naturally, in real data applications as in Chapter 5 multi-dimensional eigenspaces are no issue,

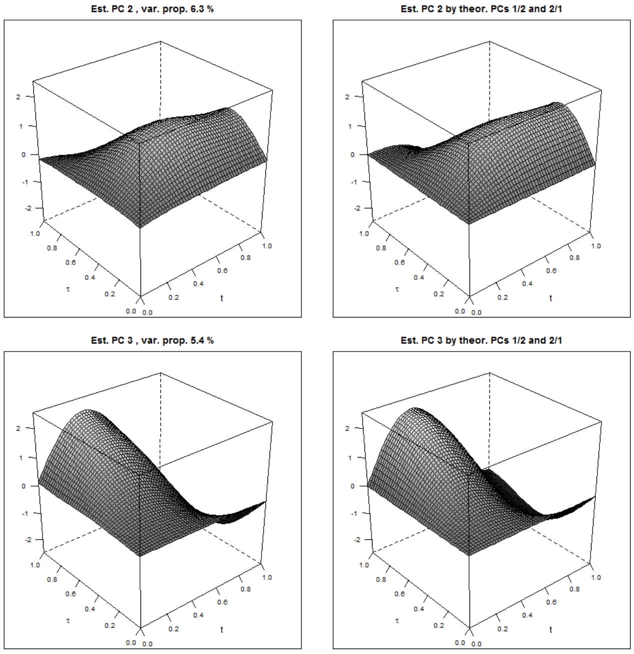


Figure 3.3: Estimated eigenfunctions for the second eigenvalue (left hand side) and their representations in the theoretic second eigenspace (right hand side).

because as already mentioned before, the possibility of having multi-dimensional eigenspaces in estimation is zero.

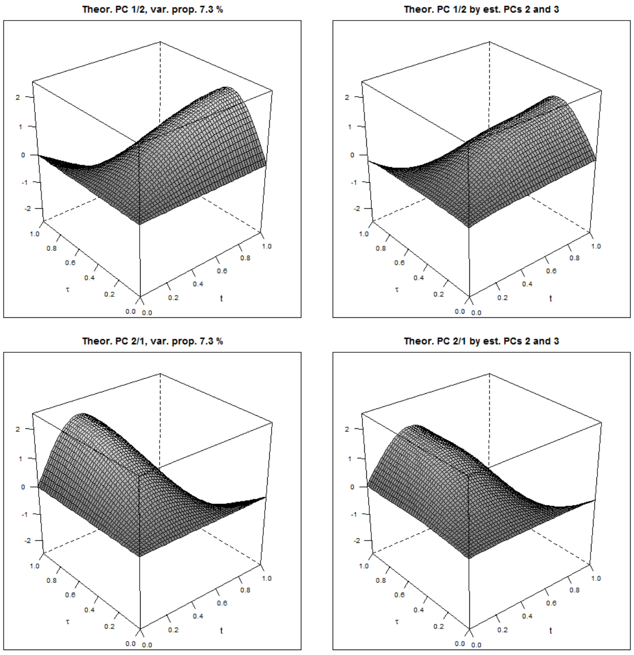


Figure 3.4: True eigenfunctions for the second eigenvalue (left hand side) and their representations in the estimated second eigenspace (right hand side).

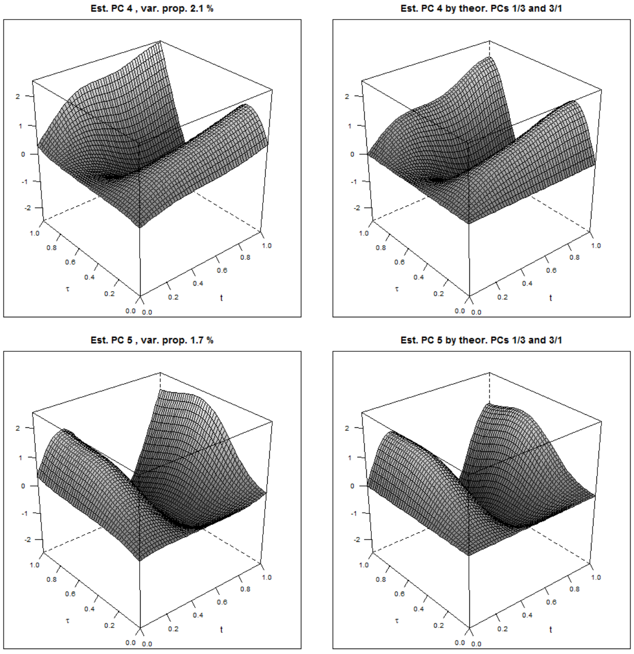


Figure 3.5: Estimated eigenfunctions for the third eigenvalue (left hand side) and their representations in the theoretic third eigenspace (right hand side).

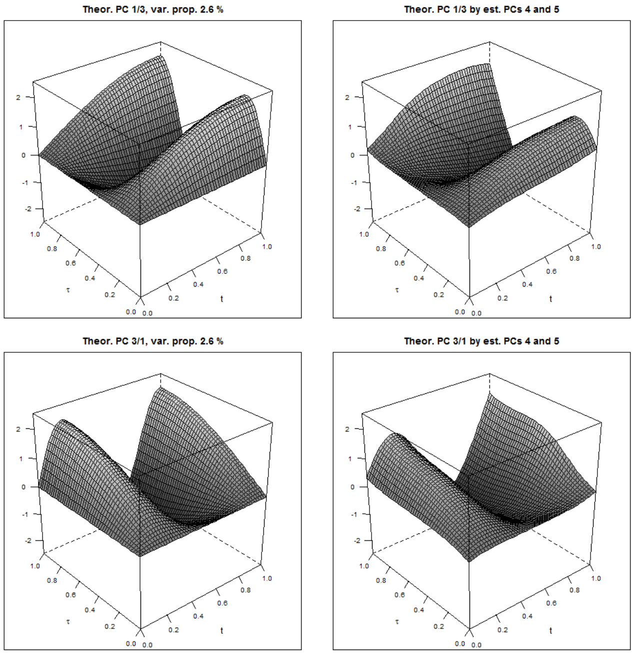


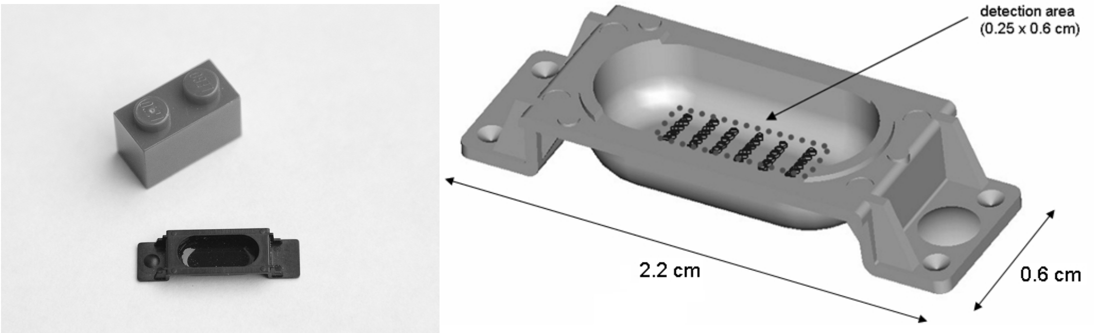
Figure 3.6: True eigenfunctions for the third eigenvalue (left hand side) and their representations in the estimated third eigenspace (right hand side).

4 Application to Diagnostic Data: One-dimensional Analysis

4.1 Introducing the system

The data we analyze are obtained by measurements carried out on a laboratory analysis system. This system can detect several analytes in a blood sample simultaneously. The measurements take place on chips with plane coated surfaces of about 1.3 cm^2 in size (see Figure 4.1). The coating consists of several specific binding points which are arranged in an array of up to 21 columns and 11 rows on the chip. After multiple steps of filling the chip with sample or buffer fluids, incubating, washing and drying it, the amount of specific binding reactions is determined by a camera sensitive to fluorescent light (see Figure 4.2 for a schema of the process). Figure 1.2a shows the image taken of the camera. In applications each column of spots binds specifically one analyte and the measured signal is used to estimate analyte concentrations.

In order to optimize several aspects concerning the performance of this system a special type of chips is used. Instead of coating the surface with binding points for several substances, these chips are coated homogeneously so that one ideally obtains a constant signal on the whole surface. In the introduction we already showed with Figure 1.2b an example for this kind of chips. Due to complex influence effects (e.g. from washing and mixing processes) the ideal situation of a constant signal can hardly be reached. Therefore differences in the basic signal level have to be accounted for in later applications. This is why we will analyze the occurring structures in the following. In order to have an overview as complete as possible the measurement system delivers measurements at a grid of 21 × 11 points per chip.



(a) A measurement chip with a Lego brick for size comparison

(b) A measurement chip with marked detection area

Figure 4.1: The measurement chip

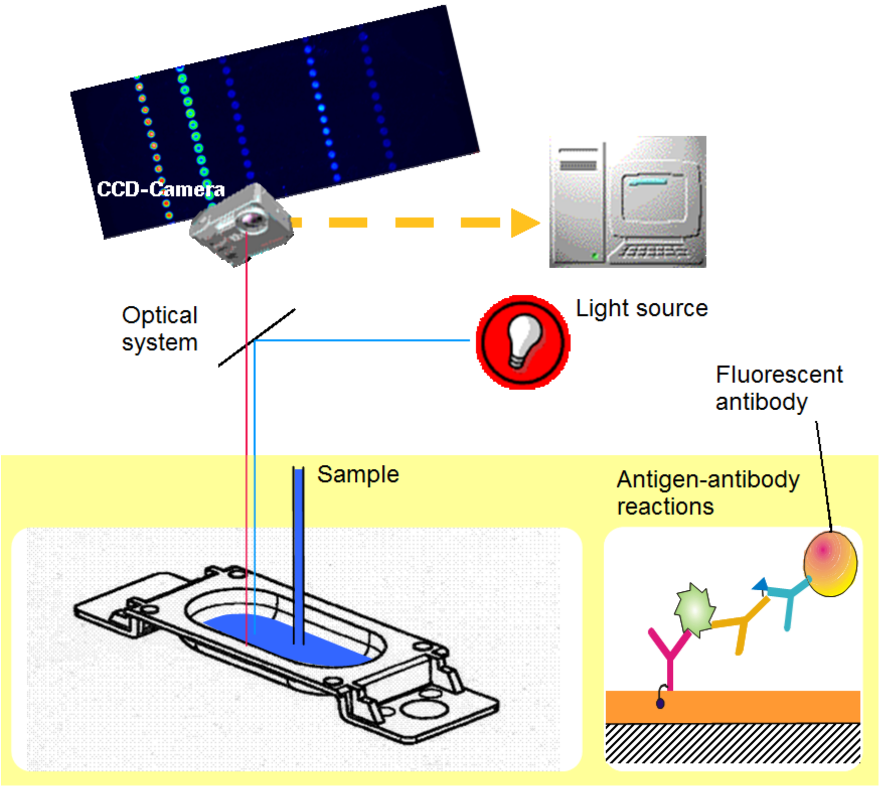


Figure 4.2: The measurement process

4.2 Application 1: Clustering the incubator units

In this section one performance aspect shall be analyzed in detail (Winzenborg et al. [2009]): The system consists of several parallel incubator units to increase the throughput. An incubator unit is an area where a chip stays for about 30 seconds at an ideal temperature so that binding reactions take place. Additionally, each incubator unit has an air jet for sample mixing during the incubation. Due to small differences in the air jets resulting from the fabrication process, the mixing is not equal at each incubator unit. This leads to visible differences in signal patterns of chips incubated at different units. For further evaluations which should not depend on incubator units, it is necessary to select incubator units which are comparable in signal height and course. This task shall be accomplished by a clustering algorithm which groups air jets and thereby incubator units that behave similarly.

A specialty in this clustering task is the fact that the incubator units are clustered indirectly via the signals of the chips they process. It is therefore necessary to deal with multiple measurements per unit in the clustering algorithms.

Instead of executing the cluster algorithms over the two dimensional chip space, column medians (i.e. medians of each 11 signal values per column) are calculated. This procedure for reducing the number of variables is chosen because in laboratory practice also a robust estimation of the column's mean value is evaluated. A robust estimator is necessary due to edge effects which are caused by insufficient washing and mixing power near the edges (the pictures in Figure 1.2 show this effect). The data set we analyze in the following consists of 30 incubator units and three chips measured per unit. Each of the 90 chips is described by a median signal course consisting of 21 values. Because of the mentioned edge effects only the middle 17 medians are analyzed in the following. Figure 4.3 shows the column medians per chip and incubator unit.

We want to perform the clustering via the FPC algorithm presented in Chapter 2 and compare it with standard multivariate procedures like K-means and hierarchical clustering as well as with a parametric method. In the following we present the methods and show results of the data application.

Methods

Treatment of multiple measurements. Since the task is to cluster the incubator units and not the chips it is necessary to find a way to treat multiple measurements on one unit. Two approaches are considered here. In the first approach we average the signals of the same unit and use only the mean functions for clustering. The second approach is to cluster all chip signals and to assign each incubator unit to the cluster where most of the chips processed by this unit belong to. If the incubator unit belongs to two or more clusters with equal weight it is randomly assigned (see Leisch [2007]).

Multivariate cluster analysis comprises essentially two methods that we also consider here: hierarchical clustering and K-means. In the following $y_i \in \mathbb{R}^N$ denotes the ith column median vector ($i = 1, \dots, I$) and $\{S_c\}_{c=1,\dots,C}$ a cluster partition of $\{1, \dots, I\}$.

Agglomerative hierarchical clustering. Hierarchical clustering creates a hierarchy of cluster partitions with increasing fineness. The agglomerative variant starts with each element repre-

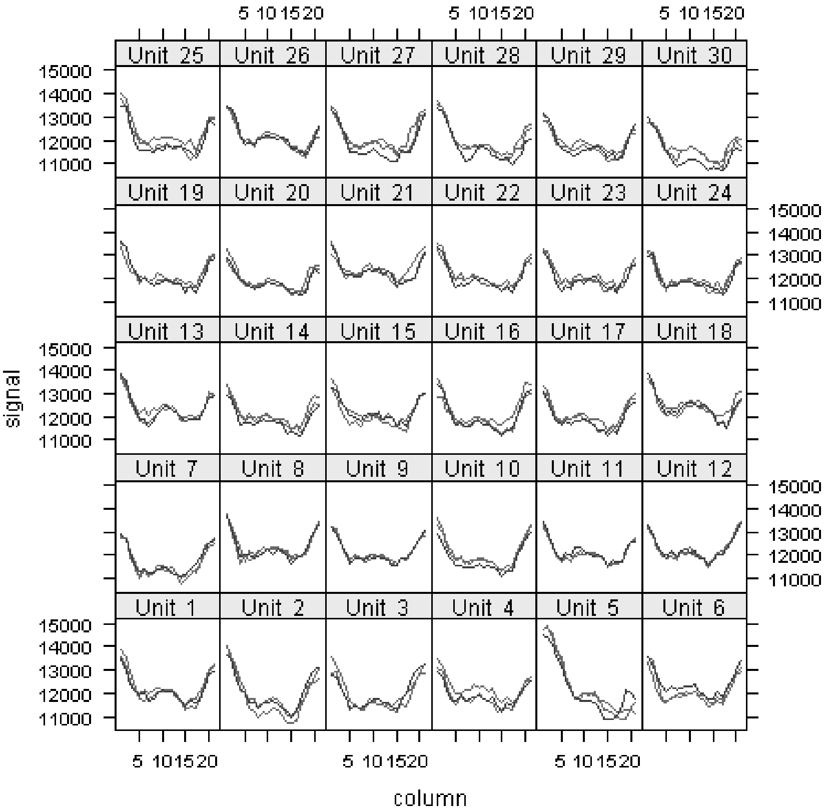


Figure 4.3: Medians per column for each chip (different gray scales) and unit.

senting its own cluster. In each of $I-1$ iterations the closest two clusters are joined into a single cluster so that after $I-1$ iterations only one cluster is left. The resulting hierarchical tree is cut at a specified level so that the desired number of clusters remains.

For determining the two closest clusters we take the complete linkage method, i.e. the maximum distance between elements of each cluster, because we look for compact clusters.

K-means clustering. The aim of K-means clustering is to partition the data in a way that minimizes the within-groups sum of squares $\sum_{c=1}^{C}\sum_{i\in S_c}||y_i-\hat{\mu}^{(c)}||_2^2$ and maximizes the between-groups variance $\sum_{c=1}^{C}||\hat{\mu}^{(c)}-\hat{\mu}||_2^2$ in which $\hat{\mu}^{(c)}$ with $\hat{\mu}^{(c)}(j)=\frac{1}{|S_c|}\sum_{i\in S_c}y_i(j)$ for $j=1,\ldots,J$ is the estimated mean of cluster c and $\hat{\mu}$ the overall mean.

The method starts with a given number of clusters C and an initial (random) distribution of elements to clusters. Then the following steps are carried out:

1. The cluster means $\mu^{(c)}$ for $c=1,\ldots,C$ are calculated.

2. Each element is assigned to the cluster with the nearest centroid, so that one obtains a new partition $\{S_c\}_{c=1,\ldots,C}$.

3. Steps 1 and 2 are repeated until no element is reassigned anymore.

	K-means	**hierarch.**	**param.**	**FPCA**
K-means	1.00	0.83	0.59	0.93
hierarch.	0.83	1.00	0.69	0.83
param.	0.59	0.69	1.00	0.52
FPCA	0.93	0.83	0.52	1.00

Table 4.1: cRate indexes between each pair of cluster partitions.

As distance measure between elements the Euclidean distance $||\cdot||_2$ is applied in all algorithms. We use the standard R functions *kmeans* and *hclust* to perform the clustering.

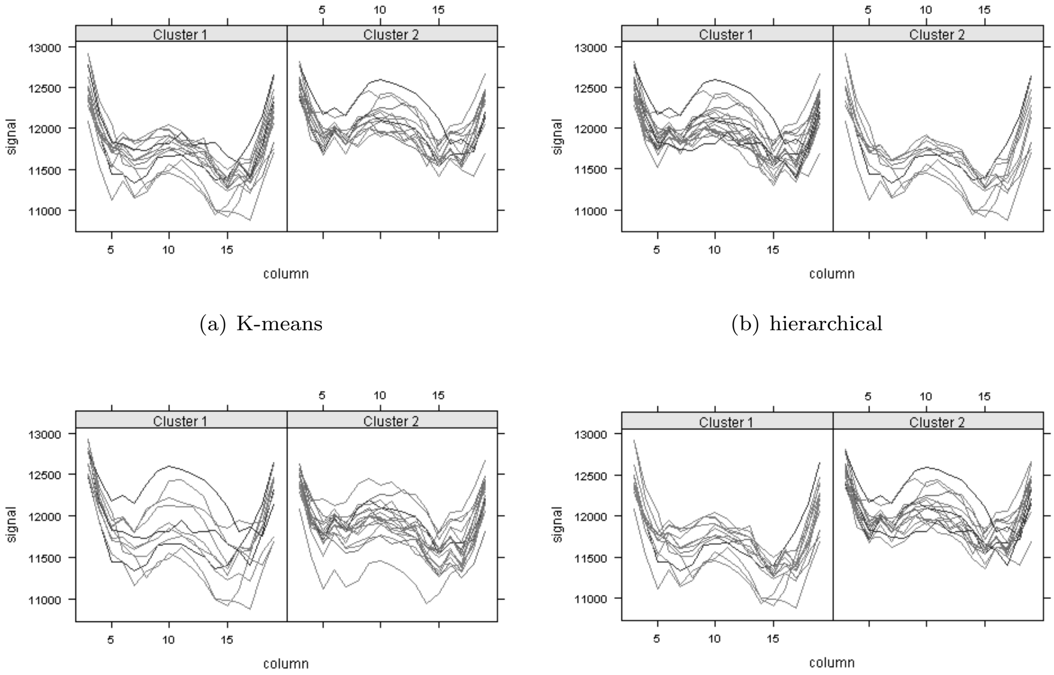
Parametric clustering Figure 4.3 leads to the assumption that the column medians can be well approximated by polynomials of fourth order. Therefore a sensible way of clustering is to fit polynomials to the column median data of each incubator unit and use a multivariate cluster algorithm (for example K-means) to cluster the coefficients afterwards. This method can cope with multiple measurements naturally by using the data of all multiples to fit the polynomial. We used least squares fitting to fit the curves. Parameters are standardized before clustering in order to have equal influence in K-means clustering.

cRate index: Agreement of cluster results The cRate (*correct classification rate*) index can be applied to measure the agreement of two cluster results or, as in Chiou and Li [2007], to compare the results of a cluster algorithm with the correct cluster partition. It expresses the proportion of maximal correspondence between two cluster partitions to the total number of objects. Let G_C be the group of permutation functions on $\{1,\dots,C\}$ and $\{S_c\}_{c=1,\dots,C}$ and $\{S'_{c'}\}_{c'=1,\dots,C}$ two cluster partitions[1].
Then the cRate index is defined by

$$\mathrm{cRate}(\{S_c\},\{S'_{c'}\}) = \max_{g\in G_C} \frac{\#\{k \mid \exists\, c \in \{1,\dots,C\} : k \in S_c \wedge k \in S'_{g(c)}\}}{I}.$$

K-means, hierarchical, parametric and FPCA clustering are applied to the data shown in Figure 4.3. As the signal of unit 5 clearly differs from the rest of the data (see Figure 4.3), all algorithms always form a single cluster out of unit 5 if three clusters are demanded. Hence the results presented here are obtained by forming two clusters of the remaining 29 incubator units. We do not increase the number of clusters, because the cluster sizes otherwise get too small for using only one of the clusters in later analysis. In FPCA we include the first principal components which describe 95 % of variation and use bandwidth 2 for estimating mean and covariance functions. For treatment of multiple measurements only the method of averaging before clustering is presented here because both methods lead to approximately the same results. Agreement of the cluster partitions of the four algorithms are compared by the cRate index. Results are shown in table 4.1.
The multivariate methods (Figures 4.4a, b) yield two clusters which clearly differ in their location. K-means and hierarchical lead here to similar, but not entirely equal results.
Before regarding the parametric results we evaluated the fit. Residual plots demonstrate that not much information is left in the residuals (not shown here). The autocorrelation of the resid-

[1] $\{S_c\}_{c=1,\dots,C}$ and $\{S'_{c'}\}_{c'=1,\dots,C}$ are of equal length. If necessary, they are filled with empty subsets.



(a) K-means

(b) hierarchical

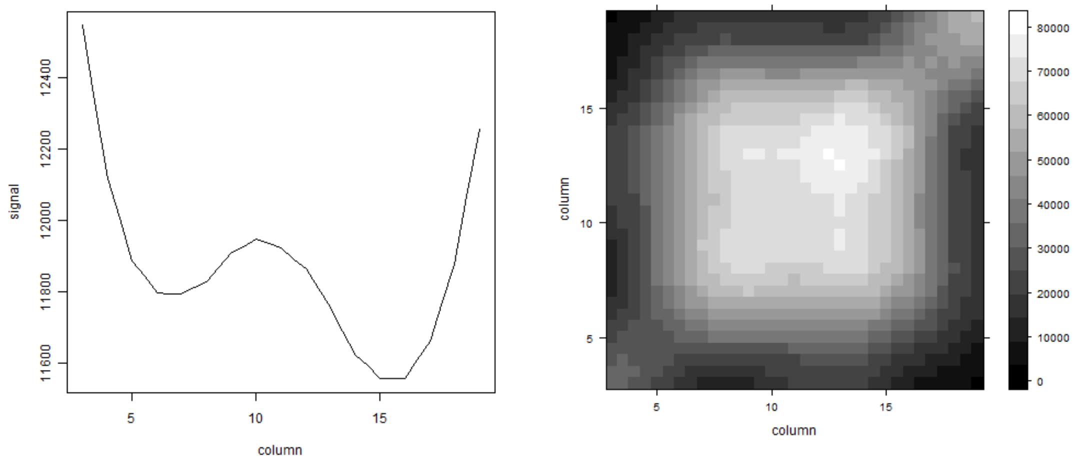
(c) parametric

(d) FPCA

Figure 4.4: Partition of incubator units into clusters for each algorithm. The lines represent mean curves per unit.

uals reveals a normal extent of correlation between joining column medians. The parametric clustering results (Figure 4.4c) are hard to interpret because curves belonging to one cluster show no obvious agreement in location or form.

As the FPCs are calculated based on mean and covariance function of the process, those are shown in Figure 4.5. Especially noticeable is that the covariance function has a high, almost constant level between the inner columns. The first four principal components of the data are shown in Figure 4.6. The first principal component explains already 71 % of variation and describes especially variations of the locations of the inner column medians. The second principal component describes mainly deviations at the right side and has with 20 % still a great influence. The third component with an influence of 8.2 % describes deviations at the left side. Compared to the first three principal components which already describe over 95 % of variation, the rest of the components have a neglectable influence. Here the fourth PC is also presented and it can be observed that this component with an influence of only 1.5 % is difficult to estimate at the boundaries. We decided that 95 % of explained variation, e.g. the first three scores suffice for clustering. The FPCA clustering results (Figure 4.4d) are similar to the K-means results.



(a) Estimated mean function

(b) Estimated covariance function

Figure 4.5: Estimated moments of the process

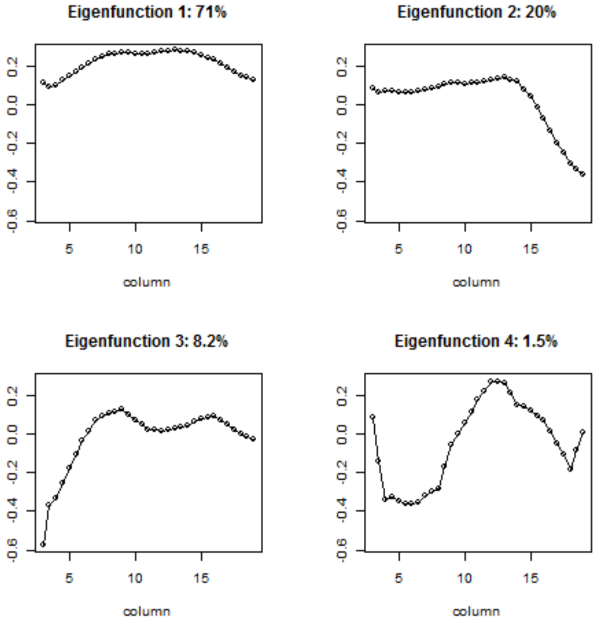


Figure 4.6: Estimated functional principal components of the column medians

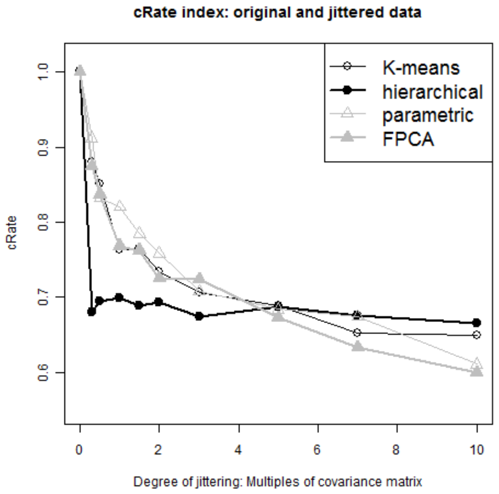


Figure 4.7: Simulation results of robustness analysis. The mean cRate index between cluster results of original and disturbed data is presented.

Robustness analysis

In order to analyze the sensitivity of the algorithms to random signal fluctuations we add a small amount of variation to each data point:

$$\tilde{y_i}(n) = y_i(n) + \epsilon_{in} \quad \text{with} \quad \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iN}) \sim \mathcal{N}(0_N, b\,\hat{\Omega})$$

$$\text{and} \quad \hat{\Omega}(n,n') = \frac{1}{I-1}\sum_{i=1}^{I}(y_i(n) - \hat{\mu}(n))(y_i(n') - \hat{\mu}(n'))$$

for $n, n' = 1\ldots, N$ and $i = 1, \ldots, I$. In this case $N = 17$ denotes the number of spot lines and 0_N the zero vector with N items. $\mathcal{N}$ is the multivariate normal distribution with the given mean vector and covariance matrix parameters. The parameter b controls the extent of disturbance and is varied in irregular distances from 0.3 to 10. For each parameter 25 data sets are constructed. The disturbed data are clustered like the original data. Afterwards the mean cRate indexes between cluster results from original and disturbed data are calculated.

The cRate indexes of K-means, parametric and FPCA clustering show a nearly equal decrease over the disturbance parameter (see Figure 4.7) so that they are similarly robust, whereas the cRate index of hierarchical clustering decays rapidly and stays constant over the rest of the analyzed parameter domain. This observation indicates that hierarchical clustering is less stable in this situation.

Discussion

The intention of this analysis is to examine and compare the K-means, hierarchical, parametric and FPCA clustering algorithms in their behavior in a special data situation.

K-means, hierarchical and FPCA clustering yield to well interpretable results which are mainly separated by location.

The initially astonishing effect that K-means and FPCA lead to similar results, can be explained by the fact that the (most influential) first principal component score mainly describes the location of each curve and location is also the most important criterion in K-means clustering. Due to the low performance of hierarchical clustering within the chosen settings we also tested other settings of metric and linkage, *e.g.* the ward linkage with the correlation coefficient as metric or single linkage with different metrics, but no setting leads to performance improvements. Single linkage settings even decrease the performance.

The advantage of FPCA compared to the multivariate methods is the consideration of the alignment of the columns. The cost is that it is slower than the multivariate algorithms due to the additional calculation of functional principal components.

For the parametric clustering one has to consider that the parameters are standardized before clustering. If not doing so, data are mainly clustered by their intercept, because this location parameter has the biggest variance. To obtain sensible results between the two extremes one could try to assign weights to the parameters.

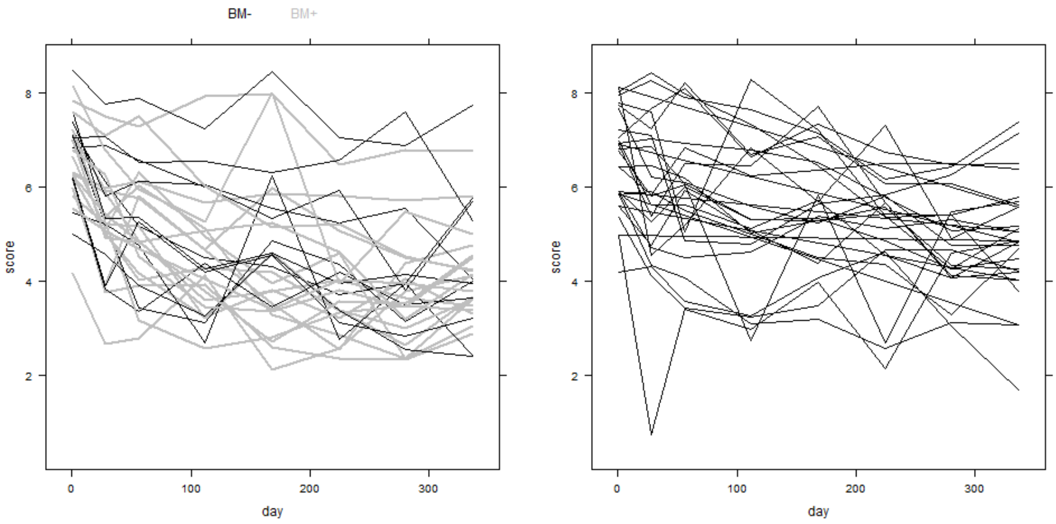
We also tried to use the KCFC algorithm presented in Section 2.4.1, but as already mentioned there, this algorithm does not work for small datasets and this dataset was too small to deliver any sensible results with this method.

In summary, regarding the interpretation of cluster results and the robustness, the algorithms K-means and FPCA are most advisable in the analyzed data situation.

4.3 Application 2: Longitudinal patient data

The main application subject in this thesis is the analysis system described in the foregoing section. But as FPCA has a wide spectrum of possible applications, we want to show the application in one further area: The analysis of longitudinal measurements of clinical parameters. In this study a continuous illness score is measured over time in two groups of patients, a placebo and a treatment group. Here the placebo group receives a standard medication and the treatment group the standard medication plus a new treatment. Measurements are taken regularly over a period of time. One aim of the analysis is to see whether the courses are differentiable between the treatment and the placebo group. Furthermore, biomarkers were measured for each patient at the beginning of the treatment. A biomarker is a protein whose concentration (for example in blood) allows conclusions about the disease state of a patient. Previous analyses have shown that a combination of two biomarkers is able to stratify the patients to a certain extent into a group where the treatment is going to help the patients getting better (biomarker positive group) and into a group where the treatment doesn't work (biomarker negative group). This rule was developed through discriminant analysis based on the disease score at the beginning of the treatment and the disease score at one fixed further point in time. We will use FPCA to see if we can detect differences in both groups.

Furthermore, we use this application to compare our FPCA method (from now on called *non-parametric FPCA* in this section) with a method described by Ramsay and Silverman (Ramsay and Silverman [2006] and Ramsay et al. [2009]). Their approach uses a functional basis in which the data set is represented before the FPCA is performed. Hence we will refer to this method



(a) Patients with treatment

(b) Patients with placebo

Figure 4.8: Exemplary patient score courses of randomly chosen placebo and treatment patients. In the treatment group patients are labeled according to their biomarker status.

as *basis FPCA* in the remainder of this section.

Data

385 patients were included in the analysis, each measured at eight points in time from one to 336 days. 69 of the patients belong to the placebo group, 316 to the treatment group. Of the treatment group 142 patients were biomarker negative, 174 were biomarker positive according to a biomarker rule.

The outcome of the measurements is a score that allows judging the severity of the disease. A patient with a higher score is more ill than a patient with a lower score.

Methods

The nonparametric FPCA is performed like presented in Section 2.2. We use a smoothing parameter of 100 days for both mean and covariance estimation and calculate the results for each 10 days between 0 and 330 days.

The basis FPCA (Ramsay and Silverman [2006]) has another approach: In this method a system of independent functions $\{\phi_l, l = 0, \ldots, L\}$ with $L \in \mathbb{N}$ is defined as a basis and in a first step, each observation is represented in the basis system:

$$y_i \approx \sum_{l=0}^{L} \hat{c}_{il} \phi_l$$

The coefficients $\hat{c}_{il} \in \mathbb{R}$ need to be chosen appropriately for each observation y_i. This is done by minimizing the squared distance of y_i and a linear combination of the basis functions:

$$\hat{c}_i \in \arg \min_{c_i \in \mathbb{R}^{L+1}} \left[\sum_{n=1}^{N} \left(y_i(t_n) - \sum_{l=0}^{L} c_{il}\phi_l(t_n) \right) \right]^2$$

In this application, we approximate each observation through a polynomial, e.g. we use the monomial basis

$$\phi_l(t) = (t-\omega)^l, l = 0, \ldots, L.$$

ω is a shift parameter, usually the middle of the observation interval, in this case $\omega = 165$.
For calculating the FPCA first remember that we have to solve the eigenequation

$$\int G(s,t)\rho(t)\, dt = \lambda\rho(s). \tag{4.1}$$

The Ramsay and Silverman method uses the assumption that the observations can be exactly represented through the basis functions, e.g.

$$y_i(t) = \sum_{l=1}^{L} c_{il}\phi_l(t) \quad \text{for all} \quad t \in \mathbb{R}$$

or in matrix notation:

$$Y = C\Phi \quad \text{with} \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_I \end{pmatrix}, C = \begin{pmatrix} c_{11} & \ldots & c_{1L} \\ \vdots & \ddots & \vdots \\ c_{I1} & \ldots & c_{IL} \end{pmatrix}, \Phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_L \end{pmatrix},$$

The covariance function is in this case given by:

$$G(s,t) = I^{-1}\Phi(s)^T C^T C\Phi(t)$$

assuming that the mean function is subtracted.
It is further assumed that the eigenfunctions can also be represented through the basis functions, e.g. for an eigenfunction ρ:

$$\rho(t) = \sum_{l=1}^{L} b_l\phi_l(t) \quad \text{respectively} \quad \rho(t) = \Phi(t)^T b \tag{4.2}$$

with coefficients $b_l \in \mathbb{R}$ and $b = (b_1, \ldots, b_L)^T$.
If we define $W = \int \Phi\Phi^T$ componentwise, the left side of equation (4.1) can be rewritten as

$$\int G(s,t)\rho(t)\, dt = \int I^{-1}\Phi(s)^T C^T C\Phi(t)\Phi(t)^T b\, dt = \Phi(s)^T I^{-1} C^T CWb$$

and the whole eigenequation as

$$\phi(s)^T I^{-1} C^T CWb = \lambda\phi(s)^T b.$$

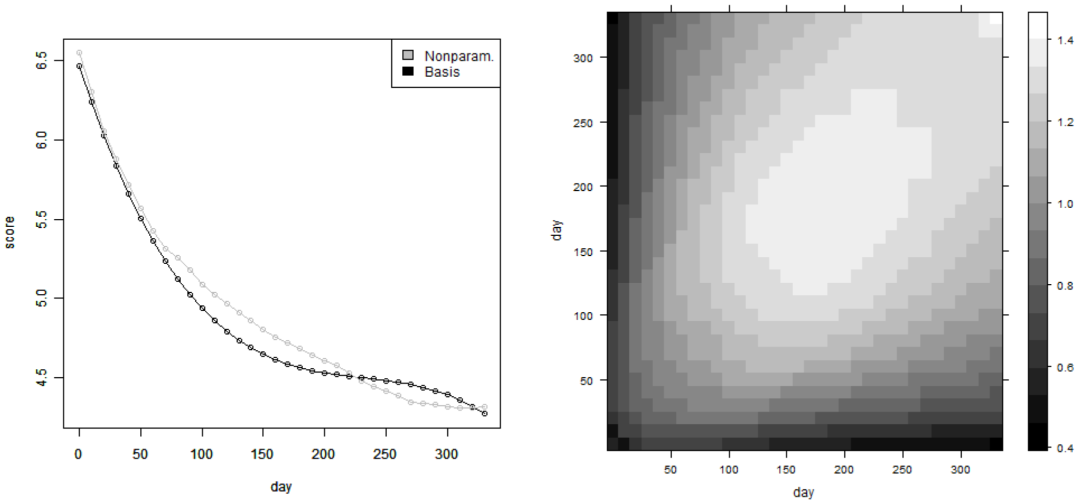
As this equation must hold for every $s \in \mathbb{R}$, this leads to

$$I^{-1}C^TCWb = \lambda b.$$

For $u := W^{\frac{1}{2}}b$ we now obtain an eigenequation in matrix form:

$$I^{-1}W^{\frac{1}{2}}C^TCW^{\frac{1}{2}}u = \lambda u$$

From the condition $||\rho|| = 1$ and $\rho_1 \perp \rho_2$ for $\rho_1 \neq \rho_2$ we also have the conditions $u^Tu = b^TWb = 1$ and $u_1^Tu_2 = b_1^TWb_2 = 0$ for $u_1 \neq u_2$. Finally the coefficients b for the original problem are given by $b = W^{\frac{1}{2}}u$, such that with equation (4.2) we obtain the functional principal components according to Ramsay and Silverman [2006].



(a) Mean functions

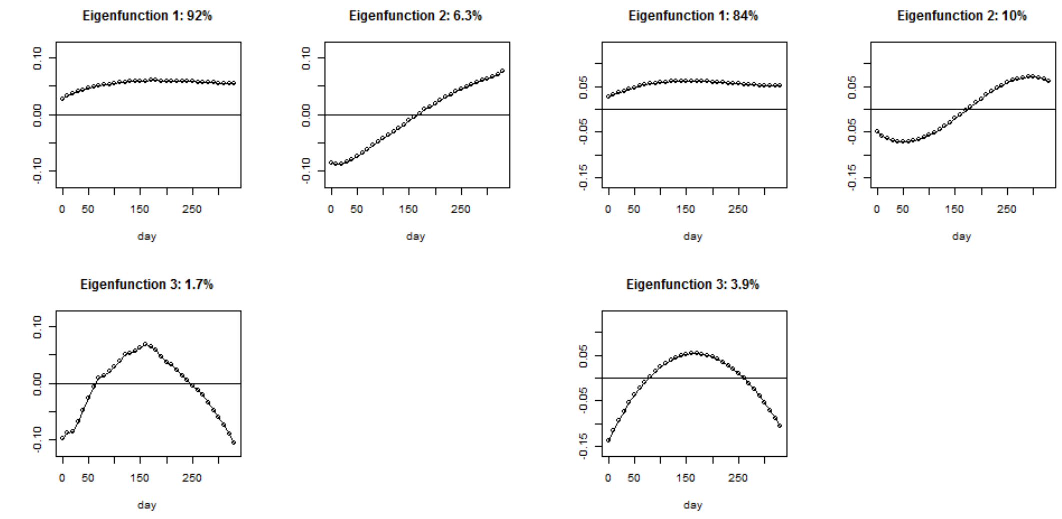
(b) Covariance function of nonparametric method

Figure 4.9: The left graphic shows the mean functions of both approaches in comparison. The right graphic shows the covariance function of the nonparametric approach.

Results

Figure 4.9 shows the mean functions over all patients for both methods as well as the covariance function for the nonparametric method. One can see that the mean score decreases over all patients and that the mean functions calculated by both methods are relatively similar and differ only for the last measurements. The covariance function shows that the measurements taken at later days are more strongly correlated than the first measurements.

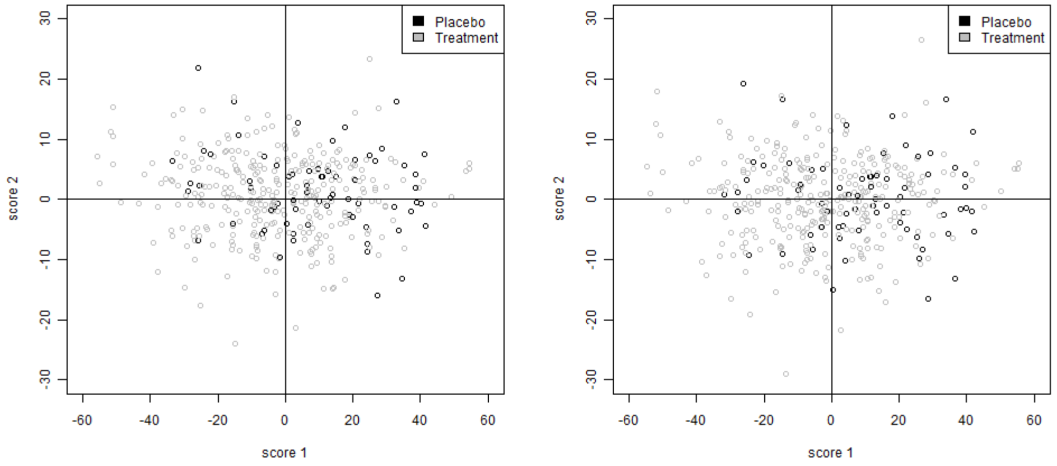
The first three eigenfunctions calculated by both methods are shown in Figure 4.10. As those eigenfunctions already explain nearly 100 % of the variability in the nonparametric case (actually the Figure headings indicate exactly 100 %, but this is due to the rounding of the variance proportion) and nearly 98 % in the basis method case, we do not show further eigenfunctions. One can see that the shape of the eigenfunctions calculated by both methods is similar. The nonparametric eigenfunctions are more variable, but therefore explain the data at hand a bit



(a) Nonparametric method

(b) Basis method

Figure 4.10: The first three estimated eigenfunctions of the process.



(a) Nonparametric method

(b) Basis method

Figure 4.11: The scores belonging to the first and to the second eigenfunction are plotted for each patient and for both methods separately. They are colored according to the treatment method.

better.

For both methods, the first eigenfunction explains mainly the location of the measurements, i.e. a positive score for the first eigenfunction means that the patients score is located higher than in the average patient. The second score describes the increase or decrease over time. A patient with a negative second score therefore has a higher decrease than the average patient. The third score alters the form of the observation, but it is with an influence of barely 2 % for the nonparametric and 4 % for the basis method case only of minor interest.

For this reason, we will concentrate our further analysis on the scores belonging to the first two eigenfunctions. Figure 4.11 shows the scores for both methods colored by treatment. One can see that the distributions look relatively similar and differ only a bit from method to method. As for both methods no clusters of patients are visible and therefore no natural division exists, we will look at the quadrants in which the first two scores lie. Therefore Figure 4.12 shows the patient courses for each quadrant, again colored by treatment and Table 4.2 shows the percentages of patients belonging to each quadrant for the treatment and for the placebo group. The patients who benefit the most from the treatment are those in quadrant 4, because they have a high score at the beginning and a steep decrease afterwards.

Both methods sort more than half of the placebo patients into the first two quadrants (which are the quadrants with substandard decrease), but also 30 respectively 36 % of the patients into quadrant four, which means that these patients have a strong medical effect also without additional treatment. The distribution in the treatment group is even less clear, which leads to the judgment that not all patients profit by the treatment. We could stratify the patients according to the scores into a group where the treatment works and into a group where the treatment does not work. But as it would be helpful to know before starting the treatment which patients are going to benefit from it, the biomarkers were measured additionally and a rule for stratifying the patients was deduced. We calculated the FPCA again, but this time only on the treatment patients. The eigenfunctions and scores are not shown, because the results are similar to the FPCA results on all patients, as they are mainly influenced by the greater group of treatment patients.

Figure 4.13 show the results, this time colored according to the biomarker status. Table 4.3 shows the percentages of the patients in each quadrant by biomarker status. 142 of the treatment patients are biomarker negative and 174 are biomarker positive. The nonparametric FPCA methods sorts 70 % of the biomarker negative patients in quadrants 1 and 2 where the decrease is slower than average and about 54 % of the biomarker positive patients into quadrants 3 and 4. In the basis method case about 65% of the biomarker negative patients are in quadrants 1 and 2 and about 57% of the biomarker positive patients are in quadrants 3 and 4.

		Quadrant 1	Quadrant 2	Quadrant 3	Quadrant 4
Nonparametric	Placebo	33.33	24.64	11.59	30.43
	Treatment	20.89	35.44	20.89	22.78
Basis	Placebo	34.78	17.39	11.59	36.23
	Treatment	25.95	25.95	29.11	18.99

Table 4.2: The distribution of the placebo and the treatment group on the four quadrants (in percent) for both FPCA methods.

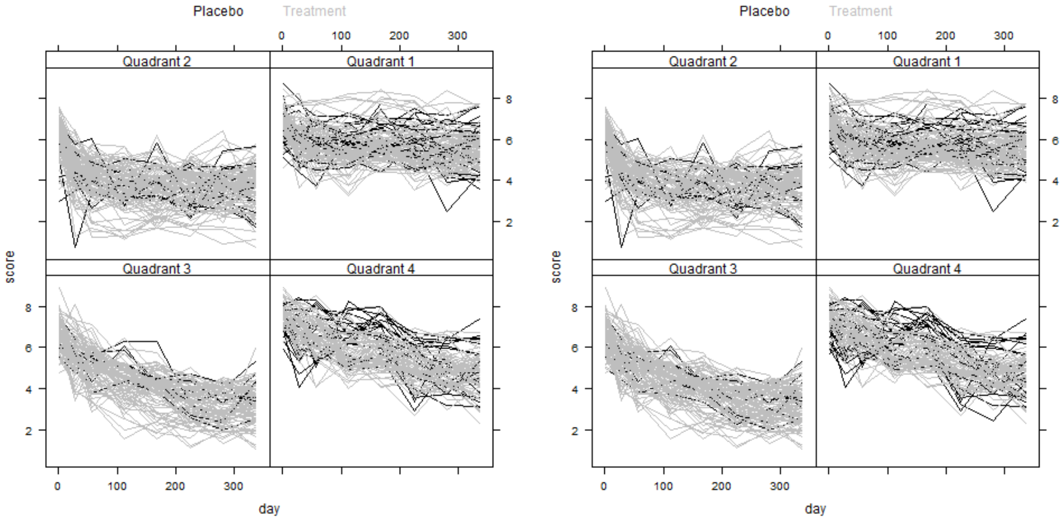
		Quadrant 1	Quadrant 2	Quadrant 3	Quadrant 4
Nonparametric	BM-	35.92	34.51	14.79	14.79
	BM+	18.39	27.59	29.89	24.14
Basis	BM-	37.32	28.17	19.72	14.79
	BM+	20.11	22.99	29.31	27.59

Table 4.3: The distribution of biomarker positive and biomarker negative patients on the four quadrants (in percent) for both FPCA methods.

Discussion

To summarize, one can conclude that the nonparametric and the basis FPCA lead to similar eigenfunctions. The eigenfunctions of the nonparametric FPCA are more variable and explain the data at hand a bit better which is due to the fact that the method has more degrees of freedom than the basis method. The basis method depends very strongly on the chosen basis system such that one difficulty in applying this method is to choose an appropriate basis. Here the system works well for the first and third eigenfunction, but non data-driven side effects occur for the second eigenfunction. We haven't tried alternative basis systems that would perhaps better describe the data at hand.

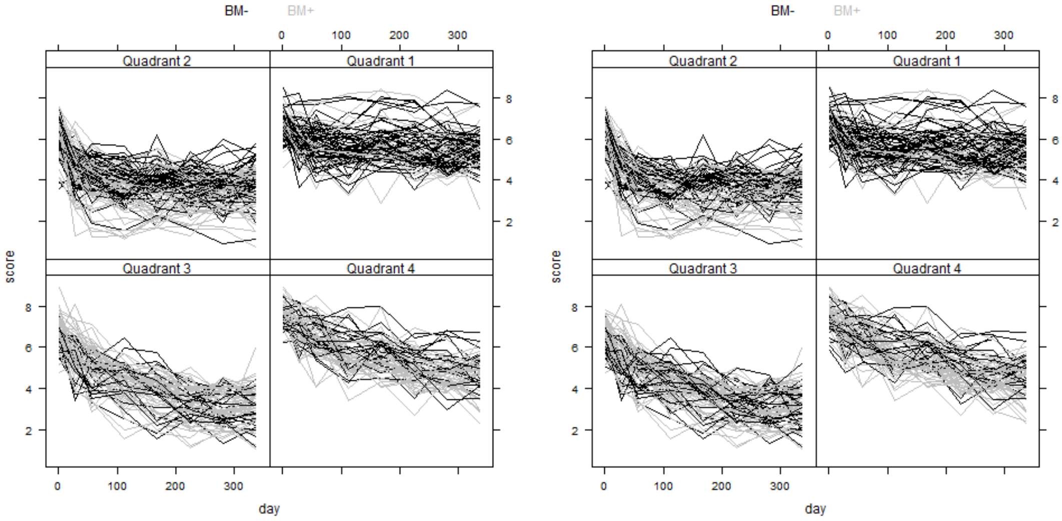
All in all, the FPCA method helps first to obtain a better visualization of the data set. In Figure 4.8 it is hard to see any structure, as the patients have totally different courses, whereas in Figure 4.13 it is possible to see the general groups of courses. One can further to a certain extent verify the results of the biomarker classification through the grouping of the scores. As the biomarker rule was developed by taking only two points in time into consideration and therefore does not include the behavior apart from these two points in time, one could imagine to perform the biomarker classification directly on the FPCA results in order to include the temporal courses into the analysis and hence to perform the classification on a more accurate representation of the actual course of disease of a patient.



(a) Nonparametric method

(b) Basis method

Figure 4.12: The patient courses in the quadrant where the corresponding first two scores belong to colored by placebo or treatment.



(a) Nonparametric method

(b) Basis method

Figure 4.13: The patient courses in the quadrant where the corresponding first two scores belong to colored by biomarker classification.

5 Application to Diagnostic Data: Spatial Analysis

5.1 Application 1: Comparison of one- and two-dimensional methods

In the following, an example which shows the advantages of complementing standard one-dimensional evaluation methods with the newly developed spatial functional principal components analysis, will be presented.

Data were measured on the multi-parameter analysis system introduced in Chapter 4. In this case, so called TSH chips (chips which are designed to bind the thyroid activity marker thyroid-stimulating hormone or thyrotropin) are used. These chips consist of 10 spot lines with 12 spots each. In order to judge the reproducibility over time, measurements were conducted at two points in time 14 days apart with 21 chips measured at the first and 10 chips measured at the second point in time. The measurement settings were identical except for the batch of one ingredient. Spot line 3 was removed from further analysis because the chips are corrupted at this line.

A standard method for evaluating the data is to look only at the spot line means and standard deviations, as presented in Chapter 4. This approach is presented in Figure 5.1, where the spot line means and standard deviations per chip are plotted against the spot line for the first and second run separately. The dots are colored by the measurement chip they belong to.

This graphic allows seeing that there are major differences between measurements taken at the two points in time. Both, mean signals and standard deviations, are higher and more variable

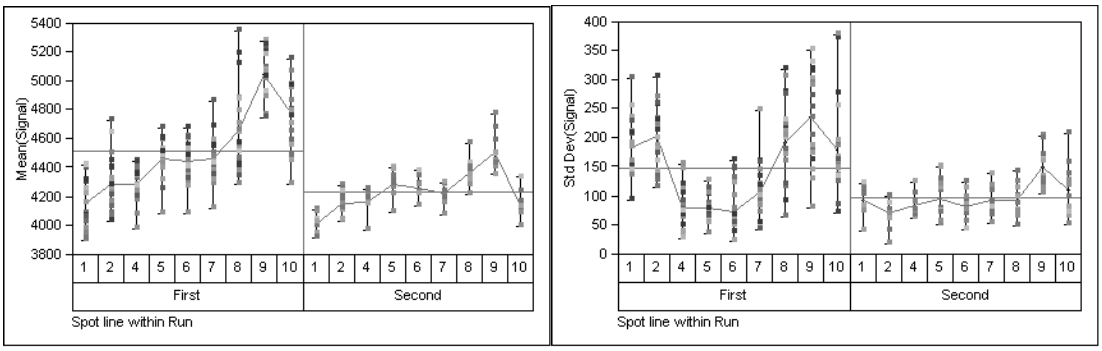
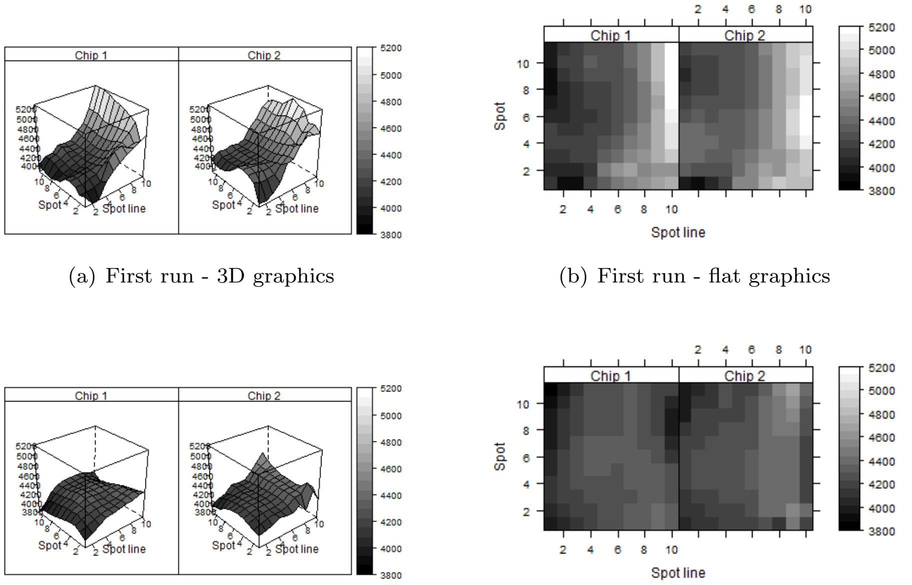


Figure 5.1: A standard approach to look at chip data: Mean and standard deviation of the signal values of each spot line as a dot colored by the chip they belong to. The additional lines connect the spot line means respectively show the overall mean of both runs separately. (Figure produced with SAS JMP 7.0.)



(a) First run - 3D graphics

(b) First run - flat graphics

(c) Second run - 3D graphics

(d) Second run - flat graphics

Figure 5.2: Smoothed surfaces of measurement values of four measurement chips belonging to two different runs. The left graphics ((a) and (c)) show the values as 3D plots and the right graphics as flat color plots. The coloring schema is the same in all plots. In the following only the flat color plots are shown.

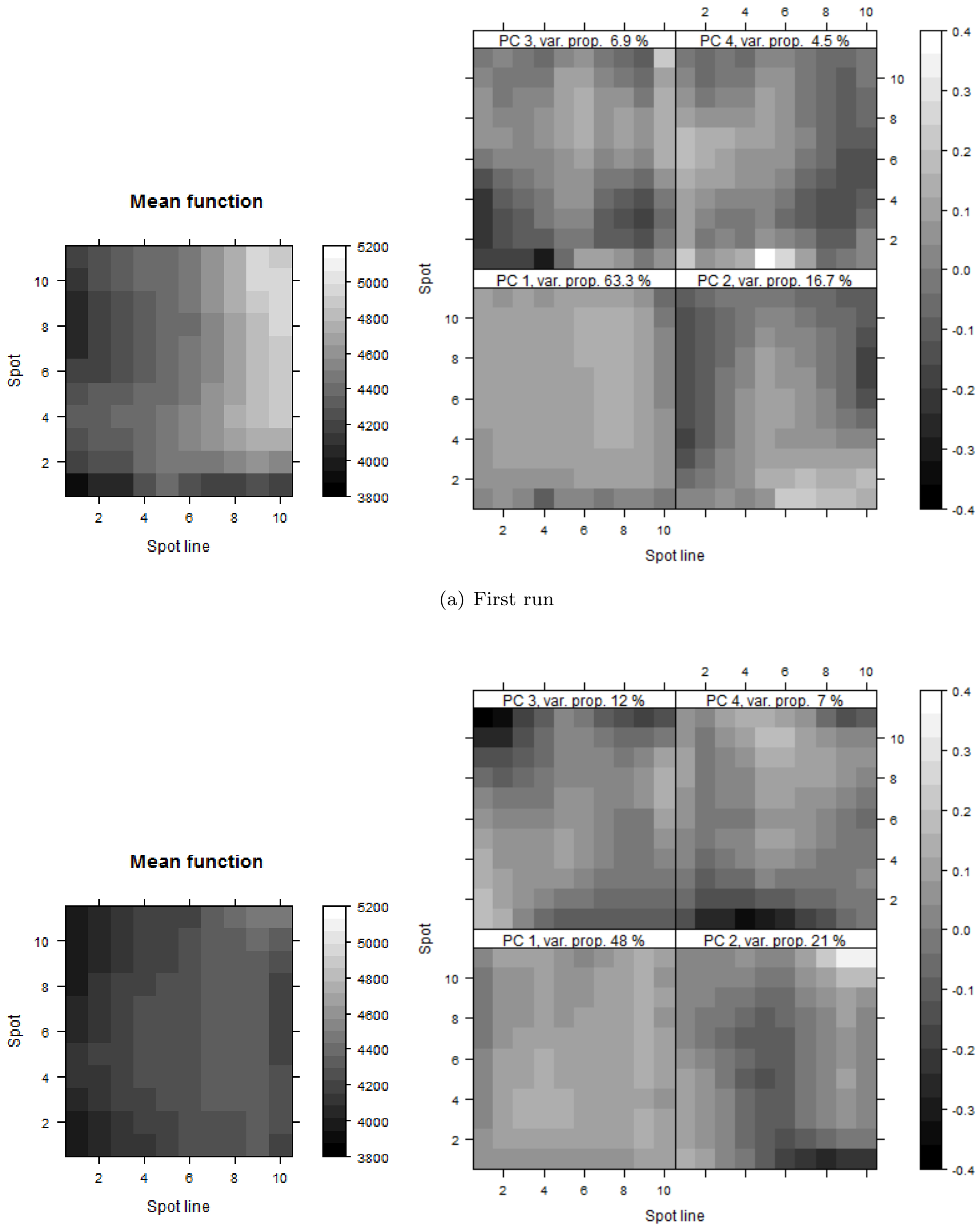
at the first point in time than at the second one. Though we obtain this information, a great disadvantage of this kind of summary is that the spatial information is almost lost and it is impossible to extract where on the chip surface the differences occur.

Therefore the approach of spatial functional principal component analysis and its various means of looking at the data are presented in the following for this data example.

To introduce the kind of data dealt with, Figures 5.2(a) and (c) plot smoothed versions of the signal values as surfaces against spot and spot line values and thereby allow to see the structures occurring on the whole chip. Especially one can see that not only in the direction of the spot lines, but also in the spots' direction structures are visible. An alternative illustration of the same data give Figures 5.2(b) and (d), where the surface is only presented by color. The second illustration method is perhaps less intuitive, but has the crucial advantage that no information is lost when printed in two dimensions. Therefore the following graphics only use the flat illustration.

To summarize the information of the whole dataset while retaining the spatial view, the method of spatial functional principal component analysis was developed (see Chapter 3). First we compare the method applied to both runs separately, second a common FPCA expansion is calculated for both runs so that a further analysis of the FPC scores can be conducted.

Figure 5.3 shows the smoothed mean and the results of the spatial FPCA calculated separately on both runs. In this case we used the approach with smoothed estimation of mean and covariance function as described in Section 3.2 with a bandwidth of 3 for both directions and both, mean and covariance function.



(a) First run

(b) Second run

Figure 5.3: Mean chip surface and first four principal components of both runs analyzed separately (PC n = nth principal component, var. prop.= variance proportion of the nth principal component in percent).

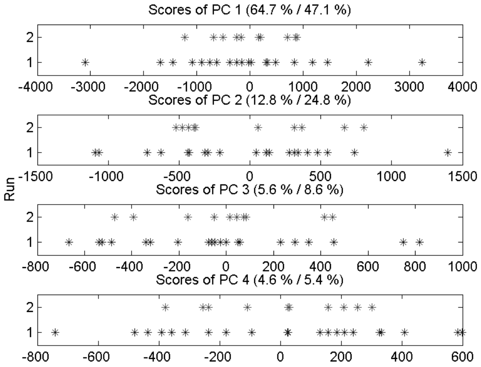


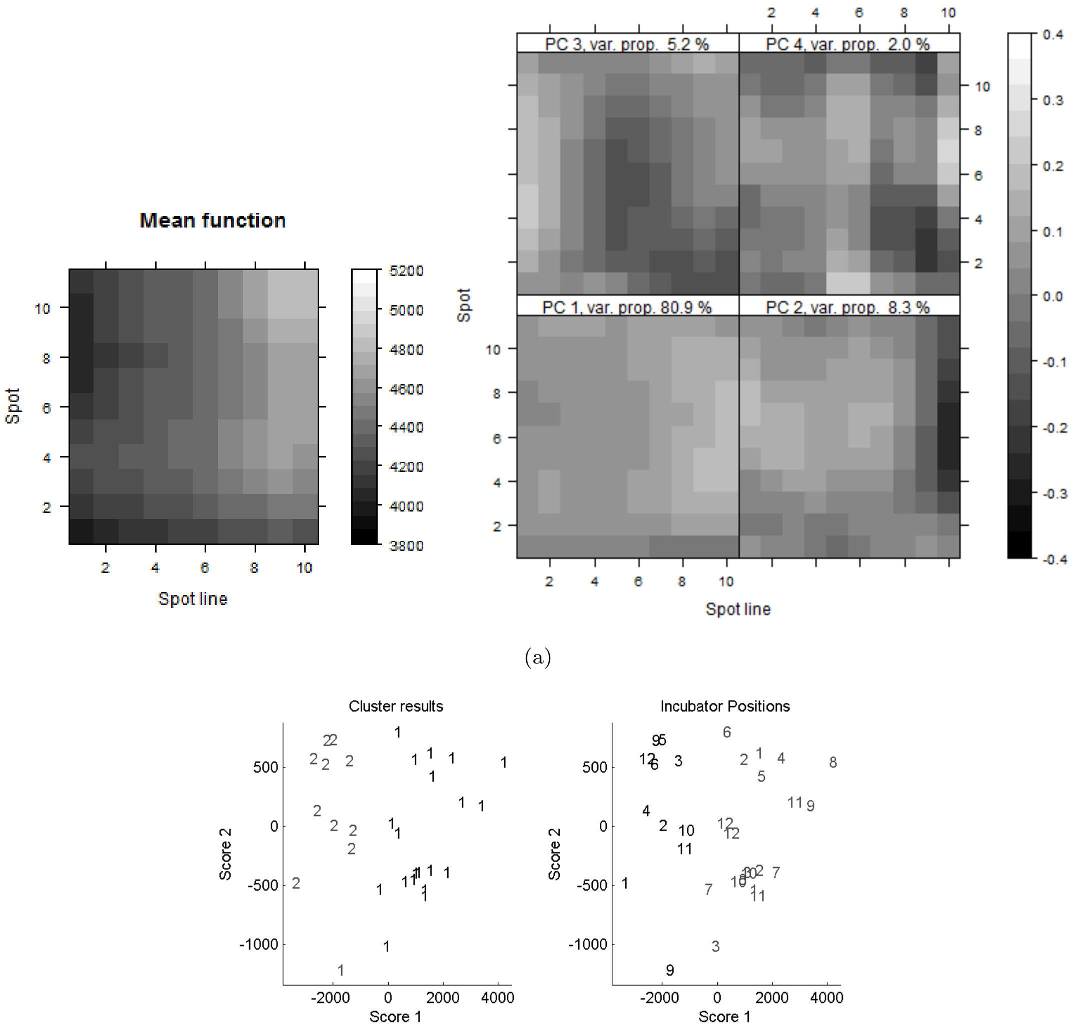
Figure 5.4: Scores of the FPCA calculation. Each dot represents the score of one measurement chip of the indicated principal component and run.

In the first run, the mean signal lies higher and shows more variation over space. The first run also has an influential first principal component with 63 % which is relatively flat. This shows the strong location variability in the first run. In the second run, it is noticeable that the first and the second principal component have a high absolute value in the upper right corner which indicates that most of the variation occurs there. This is a spatial effect that cannot be detected in Figure 5.1.

In Figure 5.4 the scores of each measurement chip (i.e. the coefficients in the linear combination of principal components which approximate the chip signal) are plotted for both runs. One can see that the scores of the first run are more variable for all four principal components. This observation corresponds to the higher standard deviation of the first run in Figure 5.1.

Further the common principal component expansion of both runs is analyzed (see Figure 5.5(a)). Having the same principal components in both runs allows analyzing the scores in more detail. Figures 5.5(b)+(c) plot the first against the second PC score. The numbers in 5.5(b) indicate the run the scores belong to. One can see that the scores of both runs can be separated clearly which underlines the observation that both runs are crucially different. A k-means clustering was performed on the scores and it leads to a nearly perfect separation. As it was seen before that the incubator position of the measurement system could have an influence on the measurement, the plot was reprinted in Figure 5.5(c) with the numbers representing the incubator positions. In this example no strong effect of the incubator position is visible.

The example shows that the spatial examination of the data through mean signal and the first principal components allows seeing structures on the chip surface and chip-to-chip variability in more detail than in the standard spot line summaries.



(a)

(b)

(c)

Figure 5.5: (a) Mean chip surface and first four principal components of both runs combined. The lower graphics show the results of k-means clustering of the scores belonging to the first four principal components presented in the plane of the first two scores. The colors indicate the cluster membership, the number in (c) the run respectively in (d) the incubator position.

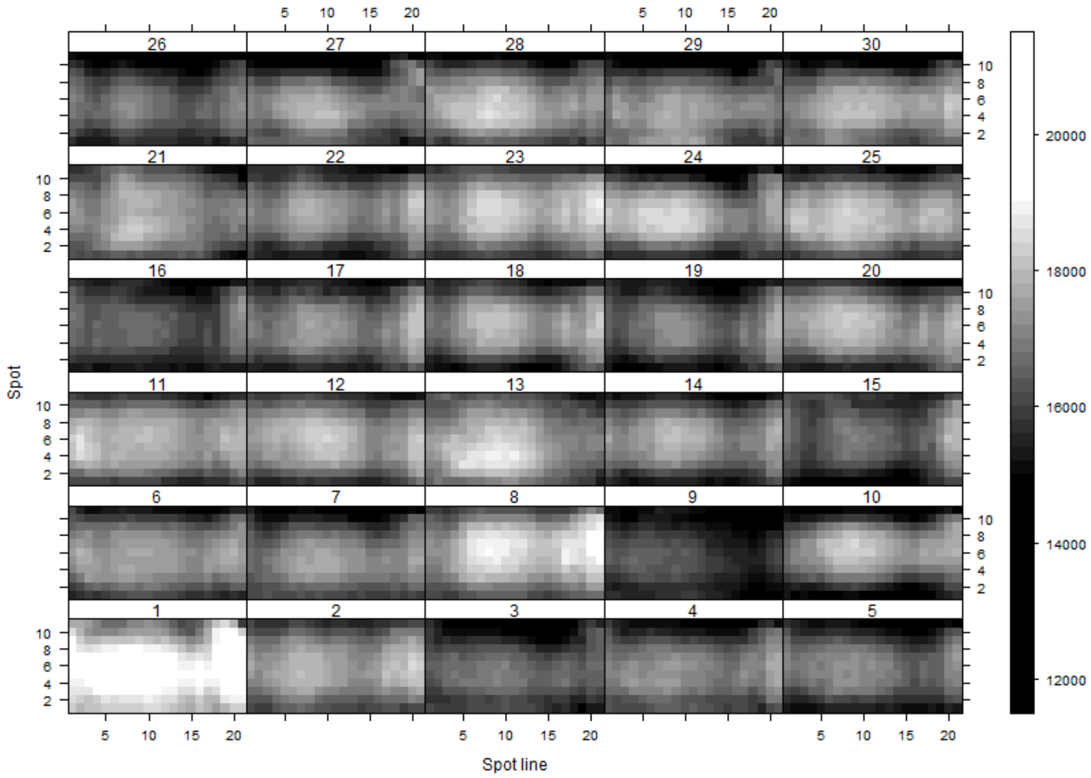


Figure 5.6: 30 measurement chips of one instrument check.

5.2 Application 2: Analysis of system performance over time

In this section we discuss a second example of spatial FPCA applied to the data of the measurement system. This time a series of so-called instrument checks are conducted at regular intervals (every 2 weeks). The aim of these tests is to see if the performance of the system stays constant over time. The standard analysis procedure for these instrument checks is like in the example before through calculating spot line means and coefficients of variation. This example shall show what can be seen using FPCA instead.

We will analyze seven consecutive instrument checks. An instrument check consists of the measurement of 30 chips, each measured at one of 30 incubator positions. The chips are coated homogeneously and are evaluated at 21 times 11 points.

The aim is to analyze the general structure of the process, detect a change of the structure over time and compare measurements taken at different positions.

Figure 5.6 shows an example for one instrument check with 30 chips.

The FPCA can be calculated for each instrument check such that the principal components can be compared over time or the FPCA can be determined over all chips and the single processes can be compared by the scores. We decided this time to use the method with smoothing of observations and eigenfunctions and not to perform the smoothed mean and covariance estimation, because for 210 chips it is very computing time-consuming. The smoothing was done with bandwidth 2 in all directions.

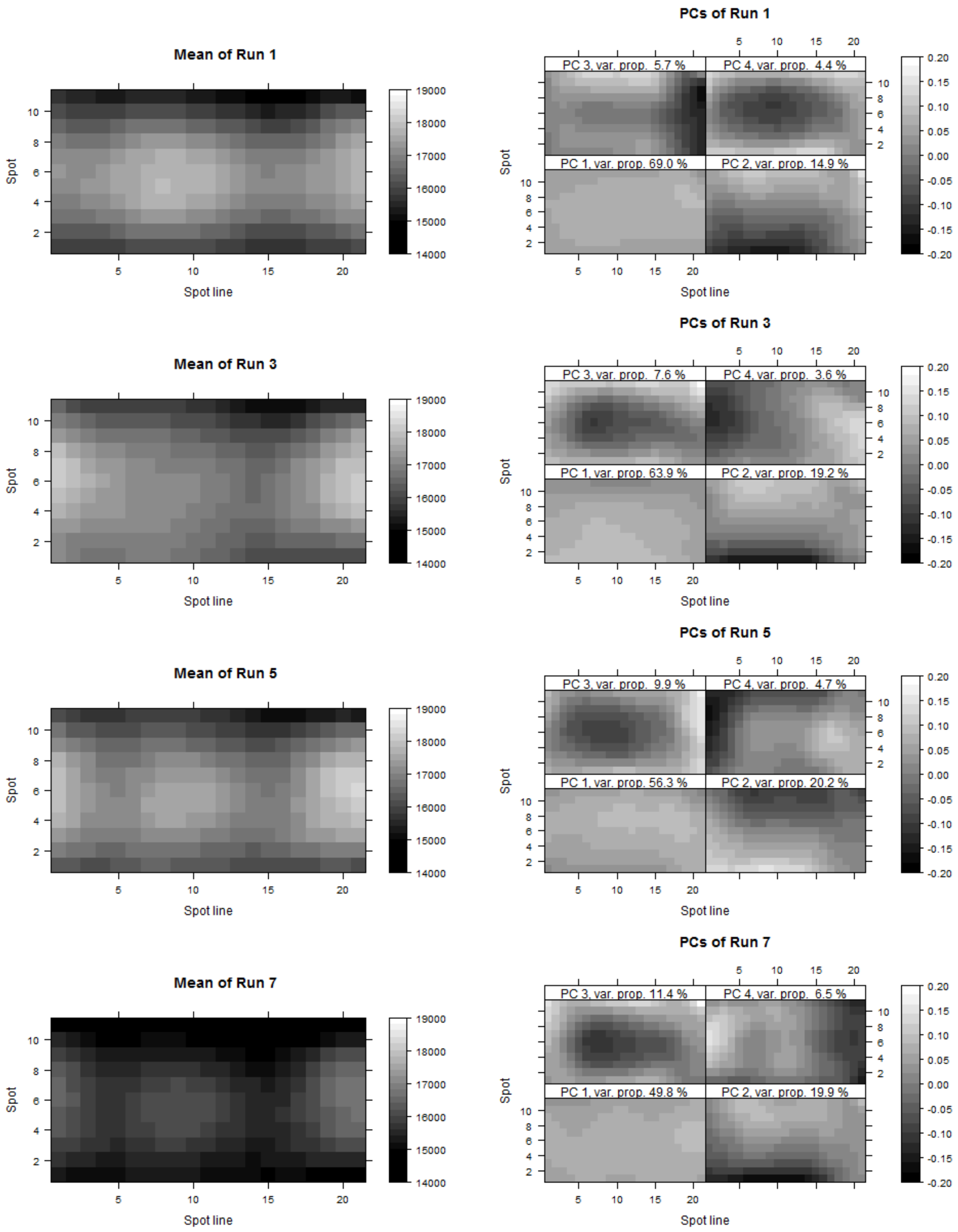


Figure 5.7: Single FPCA analyses of four runs.

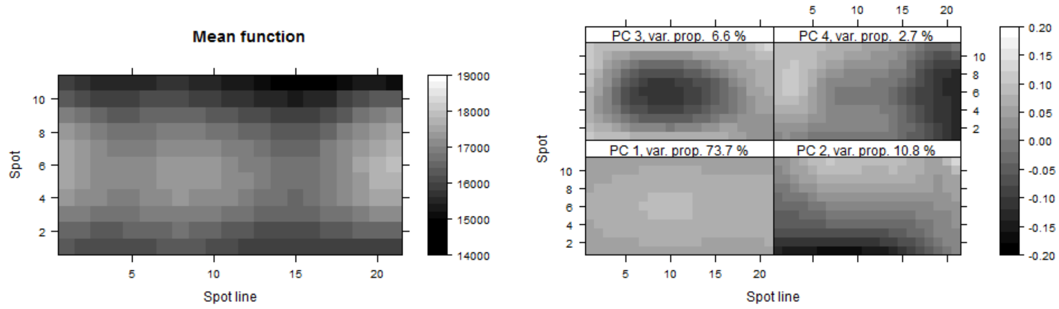


Figure 5.8: Result of the FPCA calculation. The left graphic shows the mean function, whereas the four graphics on the right show the first four principal components.

Figure 5.7 shows the FPCA calculated for four of the seven runs. These graphics allow comparing the general covariance structure between the runs. The first observation is that the mean function seems to change over time concerning the peaks in the middle of the chips and the general height, which above all decreases in run number 7. Further one can see that the first principal component is mainly flat, so that it describes the general location. The second principal component always describes differences between the upper right and the lower left part of the chips. Then we have a component describing differences in the height of the peak in the middle of the chip. For the first run it is the fourth component, for the other three runs it is the third component. The remaining component describes again differences between sides of the chip with varying priorities between the runs. In runs 1 and 3 the first component has a much higher influence as in runs 4 and 7 which can be interpreted that in runs 1 and 3 the chips differ more in location than in structure.

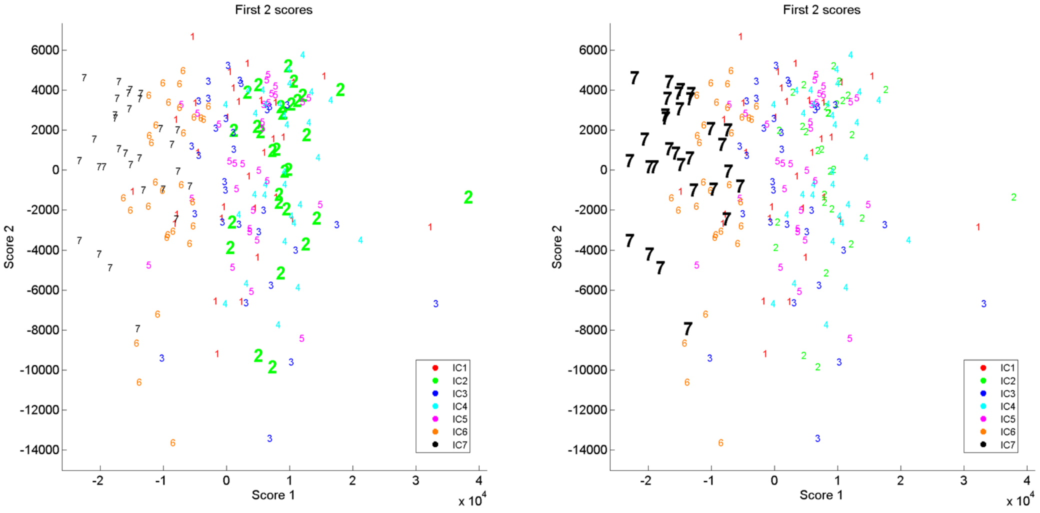
In order to be able to compare the runs through regarding the scores, Figure 5.8 shows the common FPCA over all chips. One can see that the mean function has three maxima in space. Further the first principal component is again mostly flat, whereas the second describes differences in height between the upper right and lower left side of the chip. The third component models like in the single FPCAs differences in the magnitude of the inner maximum and the fourth component describes differences in signal height between the upper left and the lower right side. The first four components describe together over 90 % of variability. We can deduce that the common FPCA is somewhat a summary of the single FPCAs of Figure 5.7.

The scores of the first two principal components of the common FPCA are shown in Figures 5.9 and 5.10. Figure 5.9 is colored and named by the instrument check such that one can see how the structure varies over time. In this case the values for score 1 decrease over time, while the distribution of the values of score 2 stays approximately the same. Additionally one can see the reason why the first principal component has such a strong influence in runs 1 and 3: The one outlying chip in runs 1,2 and 3 on the right sight in 5.9 leads to this phenomenon.

Another point of interest was to detect differences between incubator units. Therefore in a second representation the graph is colored again by the instrument check, but named by the incubator unit. Figure 5.10 shows two versions with different incubator positions highlighted. It can be seen that the first score (the position) varies within one unit, but the second score (the shape) stays almost constant. This is an interesting observation because it means that the shape of the measured signal is influenced mainly by the unit, whereas the position is influenced by other factors, for example the deterioration of ingredients. Here we see that the three outliers were all measured at the first incubator position. The explanation is that after three runs the engineers detected the bad performance of this incubator position and exchanged the unit. Hence in the rest of the runs this position is no outlier anymore.

Another effect which can be seen by analyzing the graphs closer is that the variance in the scores of one incubator unit seems to increase in the last units. This can be explained by the approach of the engineers to put the incubator units with bad performances to the last positions of the system as they are not used as often as the first units.

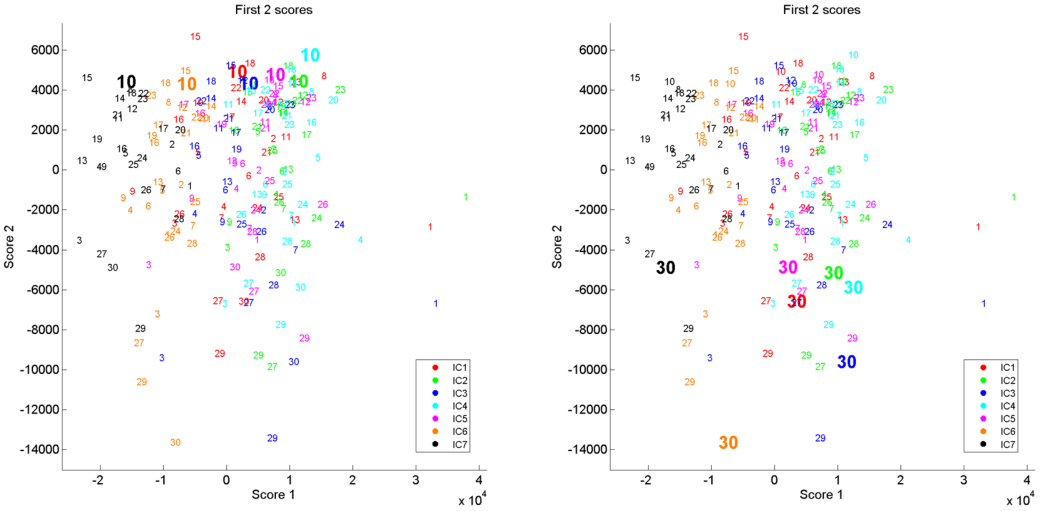
We see that this analysis already tells a lot about the spatial structure of the measurement. One can monitor changes in performance over time as well as the performance of single incubator positions. Additionally one can detect outlying incubator positions. In summary the FPCA method is a suitable tool to monitor the instrument checks.



(a) Instrument check 2 highlighted

(b) Instrument check 7 highlighted

Figure 5.9: The scores of the first two principal components named and colored by instrument check number.



(a) Incubator unit 10 highlighted

(b) Incubator unit 30 highlighted

Figure 5.10: The scores of the first two principal components named by incubator unit and colored by instrument check number.

6 Consistency Results

This chapter shows how accurate the estimation of principal components on data sets works. Therefore consistency is shown in the case of one- as well as two-dimensional functional data sets and convergence rates are calculated. We concentrate on the situations most common in our applications and demand in addition to Sections 2.2 and 3.2 that data are measured at regular points in time and on a regular grid respectively.

As guidelines for the consistency calculations serve the papers Yao et al. [2005] and Yao and Lee [2006]. In Yao et al. [2005] the one-dimensional case with random points in time is presented and proved. Yao and Lee [2006] treats the case of one-dimensional data taken at regular points in time without carrying out the proofs.

Before turning to the consistency results, a short overview about the Landau symbolism to notate convergence rates shall be given.

6.1 Landau symbols

The Landau symbolism allows to notate the orders of convergence in a simple way and makes calculations with them more clearly.

In numerics it is common to notate the speed of convergence of sequences with the (deterministic) *Landau symbols* O and o:

Definition 6. Let $(x_n)_{n\in\mathbb{N}}$ and $(a_n)_{n\in\mathbb{N}} > 0$ be real sequences, then:

$$x_n = O(a_n)$$

$$:\Leftrightarrow \quad \forall\, n\, \exists\, M > 0\text{: } |x_n| \leq M a_n$$

$$x_n = o(a_n)$$

$$:\Leftrightarrow \quad \forall\, \epsilon > 0\, \exists\, n_0 : \; |x_n| \leq \epsilon\, a_n \text{ for } n > n_0$$

There exists a corresponding notation with (stochastic) Landau symbols O_P and o_P for sequences of random vectors (Pollard [1984]):

Definition 7. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random values and $(a_n)_{n\in\mathbb{N}} > 0$ a real sequence, then:

$$\begin{aligned} & X_n = O_p(a_n) \\ :\Leftrightarrow \quad & \forall\, \epsilon > 0 \,\exists\, M > 0 \,\exists\, n_0 > 0 \;:\; P(|X_n| \geq M a_n) \leq \epsilon \,\forall\, n > n_0 \\ & X_n = o_p(a_n) \\ :\Leftrightarrow \quad & \forall\, \epsilon > 0 : P(|X_n| \geq \epsilon\, a_n) \overset{n\to\infty}{\longrightarrow} 0 \end{aligned}$$

Especially this means:

$$X_n = o_p(1) \Leftrightarrow X_n \to 0 \text{ in probability}$$

$$X_n = O_p(1) \Leftrightarrow X_n \text{ is stochastically bounded}$$

Properties 6. *If $X_n = a + O_P(a_n), Y_n = b + O_P(b_n), a, b \in \mathbb{R}, a_n, b_n \to 0$, $a_n = O(b_n)$, we can derive that:*

$$X_n + Y_n = a + b + O_P(b_n)$$

$$X_n Y_n = ab + O_P(b_n)$$

$$\frac{X_n}{Y_n} = \frac{a}{b} + O_P(b_n) \quad \textit{(provided } b \neq 0\textit{)}$$

These properties will be used in the following proofs.

6.2 Consistency for FPCA

In order to derive consistency results for the FPCA estimation, we assume in this chapter i.i.d. processes Y_t^i with $t \in T$ for $i = 1, \ldots, I$ having a density function $g(y;t)$. Further $g_2(y_1, y_2; t_1, t_2)$ shall be the joint density of $Y_{t_1}^i$ and $Y_{t_2}^i$. The points in time t_n^i where the measurements take place are assumed to be fixed, equal for all observations and of the same distance. For notational convenience we further assume $T = [0,1]$ and therewith $t_n^i = t_n = \frac{n}{N}$ and $Y_n^i := Y_{\frac{n}{N}}^i$.
In the proofs of the consistency results for mean and covariance estimators, we will use the fact that the estimators can be composed of more simple kernel estimators. Thats why at first general consistency results for kernel estimators of functions defined on T respectively T^2 are shown in Lemmas 7 and 8.

6.2.1 Lemmata

Lemma 7 (Lemma for mean estimation, see also Yao et al. [2005])**.** *Assume that*

(a) $\mathcal{K}$ is an absolutely integrable kernel, i.e. $\int |\mathcal{K}(t)|\, dt < \infty$, and has an absolutely integrable Fourier transform.

(b) $\mathcal{K}$ is compactly supported of order (ν, l) with $l > \nu$ and $\int \mathcal{K}^2(t)\, dt < \infty$.

(c) The bandwidth h and the data points per observation N depend on the sample size I and fulfill $h \to 0, Ih^{\nu+1} \to \infty, Ih^{2l+2} = O(1), \frac{1}{N} = O(h^{l+2})$ for $I \to \infty$.

(d) $\frac{d^l}{dt^l} g(y;t)$ exists and is continuous on $T \times \mathbb{R}$.

Let further $\psi : \mathbb{R}^2 \to \mathbb{R}$ be a real function with:

(e) ψ is uniformly continuous on $T \times \mathbb{R}$.

(f) $\sup_{t \in T} \int \psi^2(t,y) g(y;t)\, dy < \infty$

Now define a weighted average and its limit expected value via

$$\Psi_I(t) = \frac{1}{Ih^{\nu+1}N} \sum_{i=1}^{I} \sum_{n=1}^{N} \psi\left(\frac{n}{N}, Y_n^i\right) \mathcal{K}\left(\frac{\frac{n}{N} - t}{h}\right),$$

$$\mu(t) = \frac{d^\nu}{dt^\nu} \int \psi(t,y) g(y;t)\, dy$$

and assume that

(g) $\frac{d^l}{dt^l} \mu(t)$ exists and is continuous on T

(h) $\mu^*(t) := \int \psi(t,y) g(y;t)\, dy$ is Lipschitz continuous on T.

Under these assumptions $\Psi_I(t)$ is a consistent estimator for $\mu(t)$ and one can obtain the following uniform consistency rate:

$$\tau_I := \sup_{t \in T} |\Psi_I(t) - \mu(t)| = O_P\left(\frac{1}{\sqrt{I}h^{\nu+1}}\right).$$

Remark. If ψ is a function only in t (i.e. Ψ_I has no random part), the rate is

$$\tau_I = O\left(h^{l-\nu}\right).$$

Proof (see also Yao et al. [2005]). First observe that the term τ_I can be splitted into a variance and a bias term:

$$\mathbb{E}|\tau_I| \leq \underbrace{\mathbb{E} \sup_{t \in T} |\Psi_I(t) - \mathbb{E}(\Psi_I(t))|}_{=:A} + \underbrace{\sup_{t \in T} |\mathbb{E}(\Psi_I(t)) - \mu(t)|}_{=:B}. \tag{6.1}$$

In the following parts A and B are evaluated separately.

Part A: First the kernel is represented in the Fourier space in order to find an expression which can be evaluated easily. Assumption (b) guarantees that the Fourier inversion formula can be applied. Hence we insert the expression

$$\mathcal{K}\left(\frac{\frac{n}{N} - t}{h}\right) = \frac{1}{2\pi} \int e^{iv \frac{\frac{n}{N} - t}{h}} \underbrace{\left[\int e^{-iuv} \mathcal{K}(u)\, du\right]}_{=:\rho(v)} dv$$

into $\Psi_I(t)$ and perform a substitution $v = uh$ afterwards:

$$\begin{aligned}
\Psi_I(t) &= \frac{1}{Ih^{\nu+1}N}\sum_{i=1}^{I}\sum_{n=1}^{N}\psi\left(\frac{n}{N},Y_n^i\right)\frac{1}{2\pi}\int e^{iv\frac{\frac{n}{N}-t}{h}}\rho(v)\,dv \\
&= \frac{1}{2\pi Ih^{\nu}N}\sum_{i=1}^{I}\sum_{n=1}^{N}\psi\left(\frac{n}{N},Y_n^i\right)\int e^{iu(\frac{n}{N}-t)}\rho(uh)\,du \\
&= \frac{1}{2\pi h^{\nu}}\int e^{-iut}\rho(uh)\underbrace{\left[\frac{1}{IN}\sum_{i=1}^{I}\sum_{n=1}^{N}e^{iu\frac{n}{N}}\psi\left(\frac{n}{N},Y_n^i\right)\right]}_{=:\varphi_I(u)}du \\
&= \frac{1}{2\pi h^{\nu}}\int e^{-iut}\rho(uh)\varphi_I(u)\,du
\end{aligned}$$

Therefore we get for the expected value of $\Psi_I(t)$:

$$\mathbb{E}(\Psi_I(t)) = \frac{1}{2\pi h^{\nu}}\int e^{-iut}\rho(uh)\mathbb{E}(\varphi_I(u))\,du$$

and for the whole part A:

$$\mathbb{E}\left(\sup_{t\in T}|\Psi_I(t)-\mathbb{E}(\Psi_I(t))|\right) \leq \frac{1}{2\pi h^{\nu}}\int|\rho(uh)|\,\mathbb{E}|\varphi_I(u)-\mathbb{E}(\varphi_I(u))|\,du.$$

For the evaluation of the term $\mathbb{E}|\varphi_I(u)-\mathbb{E}(\varphi_I(u))|$ we observe that

$$\mathbb{E}|\varphi_I(u)-\mathbb{E}(\varphi_I(u))| \leq \sqrt{\mathbb{E}(\varphi_I(u)-\mathbb{E}(\varphi_I(u)))^2} = \sqrt{\mathrm{Var}(\varphi_I(u))}.$$

A bound for the variance term $\mathrm{Var}(\varphi_I(u))$ can be calculated exploiting that Y_n^i are i.i.d. random variables (in i) and that the variance is smaller than the expected squared value:

$$\begin{aligned}
\mathrm{Var}(\varphi_I(u)) &= \mathrm{Var}\left[\frac{1}{IN}\sum_{i=1}^{I}\sum_{n=1}^{N}e^{iu\frac{n}{N}}\psi\left(\frac{n}{N},Y_n^i\right)\right] \\
&= \frac{1}{IN^2}\mathrm{Var}\left[\sum_{n=1}^{N}e^{iu\frac{n}{N}}\psi\left(\frac{n}{N},Y_n\right)\right] \\
&\leq \frac{1}{IN^2}\mathbb{E}\left[\sum_{n=1}^{N}e^{iu\frac{n}{N}}\psi\left(\frac{n}{N},Y_n\right)\right]^2 \\
&\leq \frac{1}{IN^2}\mathbb{E}\Bigg[\underbrace{\left(\sum_{n=1}^{N}|e^{2iu\frac{n}{N}}|\right)}_{=N}\left(\sum_{n=1}^{N}\psi^2\left(\frac{n}{N},Y_n\right)\right)\Bigg] \\
&\leq \frac{1}{IN}\sum_{n=1}^{N}\mathbb{E}\left(\psi^2\left(\frac{n}{N},Y_n\right)\right) \qquad \text{(using Cauchy-Schwarz inequality)}
\end{aligned}$$

$$\leq \frac{1}{I} \max_{n=1,\dots,N} \mathbb{E}\left(\psi^2\left(\frac{n}{N}, Y_n\right)\right)$$

All in all one obtains for expression A:

$$\mathbb{E}\left(\sup_{t\in T} |\Psi_I(t) - \mathbb{E}(\Psi_I(t))|\right) \leq \frac{1}{2\pi h^\nu}\sqrt{\frac{1}{I} \max_{n=1,\dots,N} \mathbb{E}\left(\psi^2\left(\frac{n}{N}, Y_n\right)\right)} \underbrace{\int |\rho(uh)|\, du}_{=\int \frac{1}{h}|\rho(v)|\, dv}$$

$$= \frac{1}{h^{\nu+1}\sqrt{I}} \underbrace{\frac{1}{2\pi}\sqrt{\max_{n=1,\dots,N} \mathbb{E}\left(\psi^2\left(\frac{n}{N}, Y_n\right)\right)} \int |\rho(v)|\, dv}_{\text{bounded by assumption (f)}}$$

$$= O\left(\frac{1}{h^{\nu+1}\sqrt{I}}\right)$$

Part B: For evaluating B we show that $\mathbb{E}(\Psi_I(t)) = \mu(t) + O(h^{l-\nu}) + O\left(\frac{1}{h^{\nu+2}N}\right)$ uniformly for $t \in T$. In order to obtain this consistency rate the sum expression is approximated by an integral with an error term of order $O\left(\frac{1}{h^{\nu+2}N}\right)$ (for calculation see (6.2) below). Afterwards the integral is evaluated by a substitution $\tau = t + vh$ such that it can be examined via a Taylor expansion afterwards. For notational convenience set $\mu^*(t) := \int \psi\left(t, y\right) g\left(y; t\right)\, dt$.

$$\mathbb{E}(\Psi_I(t)) = \frac{1}{h^{\nu+1}N}\mathbb{E}\left(\sum_{n=1}^{N} \psi\left(\frac{n}{N}, Y_n\right) \mathcal{K}\left(\frac{\frac{n}{N} - t}{h}\right)\right)$$

$$= \frac{1}{h^{\nu+1}N}\sum_{n=1}^{N} \int \psi\left(\frac{n}{N}, y\right) g\left(y; \frac{n}{N}\right) \mathcal{K}\left(\frac{\frac{n}{N} - t}{h}\right)\, dy$$

$$= \frac{1}{h^{\nu+1}N}\sum_{n=1}^{N} \mu^*\left(\frac{n}{N}\right) \mathcal{K}\left(\frac{\frac{n}{N} - t}{h}\right)$$

$$= \frac{1}{h^{\nu+1}} \int \mu^*(\tau)\mathcal{K}\left(\frac{\tau - t}{h}\right)\, d\tau + O\left(\frac{1}{h^{\nu+2}N}\right) \quad \text{(sum replaced by integral)}$$

$$= \frac{1}{h^{\nu}} \int \mu^*\left(t + vh\right)\mathcal{K}(v)\, dv + O\left(\frac{1}{h^{\nu+2}N}\right)$$

$$= \frac{1}{h^{\nu}} \int \left[\sum_{j=0}^{l-1} \mu^{*(j)}(t)\frac{(vh)^j}{j!} + \mu^{*(l)}(\xi_v)\frac{(vh)^l}{l!}\right] \times \mathcal{K}(v)\, dv + O\left(\frac{1}{h^{\nu+2}N}\right)$$

$$\text{for} \quad \xi_v \in [t, t + vh]$$

$$= \frac{h^\nu}{\nu! h^\nu}\mu^{*(\nu)}(t) \underbrace{\int \mathcal{K}(v)v^\nu\, dv}_{=(-1)^\nu \nu!} + \frac{h^l}{l! h^\nu} \underbrace{\mu^{*(l)}(\xi_v)}_{\text{bounded}} \int \mathcal{K}(v)v^l\, dv + O\left(\frac{1}{h^{\nu+2}N}\right)$$

$$= \mu(t) + \underbrace{O(h^{l-\nu}) + O\left(\frac{1}{h^{\nu+2}N}\right)}_{\text{independent of } t\in T}$$

The error of discretization, e.g. the error that occurs when substituting a sum expression by an integral as in the last calculation, is evaluated in the following. Therefore the integral is first inserted artificially in the sum expression and afterwards the term is divided in two parts E_1 and E_2, where conditions of the kernel function and of μ^* can be used.

$$\begin{aligned}
&\frac{1}{h^{\nu+1}N}\sum_{n=1}^{N}\mu^*\left(\frac{n}{N}\right)\mathcal{K}\left(\frac{\frac{n}{N}-t}{h}\right) \qquad (6.2)\\
=&\frac{1}{h^{\nu+1}}\int \mu^*(v)\,\mathcal{K}\left(\frac{v-t}{h}\right)\,dv\\
&+\frac{1}{h^{\nu+1}}\sum_{n=1}^{N}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\mu^*\left(\frac{n}{N}\right)\mathcal{K}\left(\frac{\frac{n}{N}-t}{h}\right)\,dv-\frac{1}{h^{\nu+1}}\sum_{n=1}^{N}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\mu^*(v)\,\mathcal{K}\left(\frac{v-t}{h}\right)\,dv\\
=&\frac{1}{h^{\nu+1}}\int \mu^*(v)\,\mathcal{K}\left(\frac{v-t}{h}\right)\,dv\\
&\underbrace{+\frac{1}{h^{\nu+1}}\sum_{n=1}^{N}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\left[\mu^*\left(\frac{n}{N}\right)-\mu^*(v)\right]\mathcal{K}\left(\frac{\frac{n}{N}-t}{h}\right)\,dv}_{=:E_1}\\
&\underbrace{+\frac{1}{h^{\nu+1}}\sum_{n=1}^{N}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\left[\mathcal{K}\left(\frac{\frac{n}{N}-t}{h}\right)-\mathcal{K}\left(\frac{v-t}{h}\right)\right]\mu^*(v)\,dv}_{=:E_2}
\end{aligned}$$

E_1 is evaluated based on the condition that μ^* is Lipschitz continuous:

$$E_1=\frac{1}{h^{\nu+1}}\sum_{n=1}^{N}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\underbrace{\left[\mu^*\left(\frac{n}{N}\right)-\mu^*(v)\right]}_{\left|\mu^*\left(\frac{n}{N}\right)-\mu^*(v)\right|\le c\left|\frac{n}{N}-v\right|\le\frac{c}{N}}\mathcal{K}\left(\frac{\frac{n}{N}-t}{h}\right)\,dv=O\left(\frac{1}{h^{\nu+1}N}\right)$$

For the evaluation of E_2 the mean value theorem is used:

$$\begin{aligned}
E_2&\le\frac{1}{h^{\nu+1}}\sum_{n=1}^{N}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\left|\left[\mathcal{K}\left(\frac{\frac{n}{N}-t}{h}\right)-\mathcal{K}\left(\frac{v-t}{h}\right)\right]\mu^*(v)\right|\,dv\\
&=\frac{1}{h^{\nu+1}}\sum_{n=1}^{N}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\left|\mathcal{K}'(\xi_n)\left(\frac{\frac{n}{N}-v}{h}\right)\mu^*(v)\right|\,dv\\
&\quad\text{for}\quad \xi_n\in\left[\min\left\{\frac{\frac{n}{N}-t}{h},\frac{v-t}{h}\right\},\max\left\{\frac{\frac{n}{N}-t}{h},\frac{v-t}{h}\right\}\right]
\end{aligned}$$

$$\leq \frac{1}{h^{\nu+1}} \sum_{n=1}^{N} \max_{n=1,\ldots,N} |\mathcal{K}'(\xi_n)| \frac{1}{hN} \int_{\frac{n-1}{N}}^{\frac{n}{N}} |\mu^*(v)| \, dv$$

$$= O\left(\frac{1}{h^{\nu+2}N}\right)$$

Now it is shown that:

Part A: $\mathbb{E}\left(\sup_{t\in T} |\Psi_I(t) - \mathbb{E}(\Psi_I(t))|\right) = O\left(\frac{1}{\sqrt{I}h^{\nu+1}}\right)$

Part B: $\mathbb{E}(\Psi_I(t)) = \mu(t) + O(h^{l-\nu}) + O\left(\frac{1}{h^{\nu+2}N}\right)$

For the combined expression (6.1) we therefore obtain the inequality:

$$\mathbb{E}|\tau_I| = O\left(\frac{1}{\sqrt{I}h^{\nu+1}}\right) + O(h^{l-\nu}) + O\left(\frac{1}{h^{\nu+2}N}\right) = O\left(\frac{1}{\sqrt{I}h^{\nu+1}}\right),$$

because $h^{l-\nu} = \frac{h^{l+1}}{h^{\nu+1}} = \underbrace{\sqrt{Ih^{2l+2}}}_{\text{bounded}} \frac{1}{\sqrt{I}h^{\nu+1}} = O\left(\frac{1}{\sqrt{I}h^{\nu+1}}\right)$ and $\frac{1}{N} = O(h^{l+2})$ (see assumption (c)).
Using Markov's inequality one obtains the result:

$$\mathbb{E}|\tau_I| = O\left(\frac{1}{\sqrt{I}h^{\nu+1}}\right) \Rightarrow \tau_I = O_P\left(\frac{1}{\sqrt{I}h^{\nu+1}}\right)$$

Observe that in case of a non-random function $\Psi_I(t)$ Part A vanishes such that the final rate is only the better rate of Part B.

□

Before applying the proven lemma, we first want to derive a second lemma, which is the two-dimensional variant of Lemma 7. This lemma is necessary in order to derive rates for covariance estimation which leads us to Remark 6.2.1.

Lemma 8 (Lemma for covariance estimation, see also Yao et al. [2005])**.** *Now let $t = (t_1, t_2)$ be an element of T^2. Assume that*

(a) $\mathcal{K}$ is an absolutely integrable kernel, i.e. $\int\int |\mathcal{K}(t_1,t_2)|\, dt_1\, dt_2 < \infty$, and has an absolutely integrable Fourier transform.

(b) $\mathcal{K}$ is compactly supported of order $((\nu_1,\nu_2), l)$ and $\int\int \mathcal{K}^2(t_1,t_2)\, dt_1\, dt_2 < \infty$.

(c) The bandwidth h and the data points per observation N depend on the sample size I and fulfill $h \to 0, Ih^{|\nu|+2} \to \infty, Ih^{2l+4} = O(1), \frac{1}{N} = O(h^{l+3})$ for $I \to \infty$.

(d) All derivatives of g_2 in t up to the lth degree exist and are uniformly continuous.

Let further $\theta : \mathbb{R}^4 \to \mathbb{R}$ be a real function with:

(e) θ is uniformly continuous on $T^2 \times \mathbb{R}^2$.

(f) $\sup_{t\in T^2} \int\int \theta^2(t_1,t_2,y_1,y_2) g_2(y_1,y_2;t_1,t_2)\, dy_1\, dy_2 < \infty$

Now define a weighted average and its limit expected value via

$$\Theta_I(t) = \frac{1}{Ih^{|\nu|+2}N(N-1)} \sum_{i=1}^{I} \sum_{n_1 \neq n_2} \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, Y^i_{n_1}, Y^i_{n_2},\right) \mathcal{K}\left(\frac{t_1 - \frac{n_1}{N}}{h}, \frac{t_2 - \frac{n_2}{N}}{h}\right),$$

$$\gamma(t) = \frac{d^{|\nu|}}{dt_1^{\nu_1} dt_2^{\nu_2}} \int\int \theta(t_1, t_2, y_1, y_2) g_2(y_1, y_2; t_1, t_2)\, dy_1\, dy_2$$

and assume that

(g) all derivatives of $\gamma(t)$ up to 2nd degree exist and are uniformly continuous on T^2

(h) $\gamma^*(t) := \int\int \theta(t_1, t_2, y_1, y_2) g_2(y_1, y_2; t_1, t_2)\, dy_1\, dy_2$ is Lipschitz continuous on T^2.

Then $\Theta_I(t)$ is a consistent estimator for $\gamma(t)$ and

$$\tau_I := \sup_{t \in T^2} |\Theta_I(t) - \gamma(t)| = O_P\left(\frac{1}{\sqrt{I}h^{|\nu|+2}}\right).$$

Remark. If θ is a function only in t (i.e. Θ_I has no random part), the error is

$$\tau_I = O\left(h^{l-|\nu|}\right).$$

Proof. Most parts of the this proof are conceptually equal to the proof of Lemma 7. Again observe at first that

$$\mathbb{E}|\tau_I| \leq \underbrace{\mathbb{E} \sup_{t \in T} |\Theta_I(t) - \mathbb{E}(\Theta_I(t))|}_{=:A} + \underbrace{\sup_{t \in T} |\mathbb{E}(\Theta_I(t)) - \gamma(t)|}_{=:B}. \tag{6.3}$$

Parts A and B are evaluated separately.

Part A: First the kernel is represented in the Fourier space in order to find an expression which can be evaluated easily. Assumption (b) guarantees that the Fourier inversion formula can be applied in both dimensions. Hence we insert the expression

$$\begin{aligned}&\mathcal{K}\left(\frac{\frac{n_1}{N} - t_1}{h}, \frac{\frac{n_2}{N} - t_2}{h}\right) \\ &= \frac{1}{(2\pi)^2} \int\int e^{i\left(v\frac{\frac{n_1}{N}-t_1}{h} + w\frac{\frac{n_2}{N}-t_2}{h}\right)} \underbrace{\left[\int\int e^{-i(uv+u'w)} \mathcal{K}(u, u')\, du\, du'\right]}_{=:\rho(v,w)} dv\, dw\end{aligned}$$

into $\Theta_I(t_1, t_2)$ and perform a substitution $v = uh, w = u'h$ afterwards:

$$\begin{aligned}&\Theta_I(t_1, t_2) \\ =&\frac{1}{Ih^{|\nu|+2}N(N-1)} \sum_{i=1}^{I} \sum_{n_1 \neq n_2} \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, Y^i_{n_1}, Y^i_{n_2}\right) \frac{1}{(2\pi)^2} \int\int e^{i\left(v\frac{\frac{n_1}{N}-t_1}{h} + w\frac{\frac{n_2}{N}-t_2}{h}\right)} \rho(v, w)\, dv\, dw\end{aligned}$$

$$=\frac{1}{(2\pi)^2 h^{|\nu|} IN(N-1)} \sum_{i=1}^{I} \sum_{n_1\neq n_2} \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, Y^i_{n_1}, Y^i_{n_2}\right) \int\int e^{iu(\frac{n_1}{N}-t_1)+iu'(\frac{n_2}{N}-t_2)} \rho(uh, u'h)\, du\, du'$$

$$=\frac{1}{(2\pi)^2 h^{|\nu|}} \int\int e^{-i(ut_1+u't_2)} \rho(uh, u'h)$$

$$* \underbrace{\left[\frac{1}{IN(N-1)} \sum_{i=1}^{I} \sum_{n_1\neq n_2} e^{i(u\frac{n_1}{N}+u'\frac{n_2}{N})} \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, Y^i_{n_1}, Y^i_{n_2}\right)\right]}_{=:\varphi_I(u,u')} du\, du'$$

$$=\frac{1}{(2\pi)^2 h^{|\nu|}} \int\int e^{-i(ut_1+u't_2)} \rho(uh, u'h) \varphi_I(u,u')\, du\, du'$$

Therefore like in Lemma 7 we obtain:

$$\mathbb{E}(\Theta_I(t_1,t_2)) = \frac{1}{(2\pi)^2 h^{|\nu|}} \int\int e^{-i(ut_1+u't_2)} \rho(uh, u'h) \mathbb{E}(\varphi_I(u,u'))\, du\, du'$$

and

$$\mathbb{E}\left(\sup_{t\in T} |\Theta_I(t_1,t_2) - \mathbb{E}(\Theta_I(t_1,t_2))|\right)$$

$$quad \leq \frac{1}{(2\pi)^2 h^{|\nu|}} \int |\rho(uh, u'h)|\, \mathbb{E}|\varphi_I(u,u') - \mathbb{E}(\varphi_I(u,u'))|\, du\, du'.$$

For the evaluation of the term $\mathbb{E}|\varphi_I(u,u') - \mathbb{E}(\varphi_I(u,u'))|$ we observe that

$$\mathbb{E}|\varphi_I(u,u') - \mathbb{E}(\varphi_I(u,u'))| \leq \sqrt{\mathbb{E}(\varphi_I(u,u') - \mathbb{E}(\varphi_I(u,u')))^2} = \sqrt{\mathrm{Var}(\varphi_I(u,u'))}.$$

The variance is again evaluated by using the condition of i.i.d. observations and the Cauchy-Schwarz inequality:

$$\mathrm{Var}(\varphi_I(u,u')) = \mathrm{Var}\left[\frac{1}{IN(N-1)} \sum_{i=1}^{I} \sum_{n_1\neq n_2} e^{i(u\frac{n_1}{N}+u'\frac{n_2}{N})} \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, Y^i_{n_1}, Y^i_{n_2}\right)\right]$$

$$= \frac{1}{IN^2(N-1)^2} \mathrm{Var}\left[\sum_{n_1\neq n_2} e^{i(u\frac{n_1}{N}+u'\frac{n_2}{N})} \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, Y_{n_1}, Y_{n_2}\right)\right]$$

$$\leq \frac{1}{IN^2(N-1)^2} \mathbb{E}\left[\sum_{n_1\neq n_2} e^{i(u\frac{n_1}{N}+u'\frac{n_2}{N})} \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, Y_{n_1}, Y_{n_2}\right)\right]^2$$

$$\leq \frac{1}{IN^2(N-1)^2} \mathbb{E}\Bigg[\underbrace{\left(\sum_{n_1\neq n_2} \left|e^{2i(u\frac{n_1}{N}+u'\frac{n_2}{N})}\right|\right)}_{=N(N-1)} \left(\sum_{n_1\neq n_2} \theta^2\left(\frac{n_1}{N}, \frac{n_2}{N}, Y_{n_1}, Y_{n_2}\right)\right)\Bigg]$$

$$\leq \frac{1}{IN(N-1)} \sum_{n_1 \neq n_2} \mathbb{E}\left(\theta^2\left(\frac{n_1}{N}, \frac{n_2}{N}, Y_{n_1}, Y_{n_2}\right)\right)$$

$$\leq \frac{1}{I} \max_{n_1 \neq n_2} \mathbb{E}\left(\theta^2\left(\frac{n_1}{N}, \frac{n_2}{N}, Y_{n_1}, Y_{n_2}\right)\right)$$

All in all one obtains for part A:

$$\mathbb{E}\left(\sup_{t \in T} |\Theta_I(t_1,t_2) - \mathbb{E}(\Theta_I(t_1,t_2))|\right)$$

$$\leq \frac{1}{(2\pi)^2 h^{|\nu|}} \sqrt{\frac{1}{I} \max_{n_1 \neq n_2} \mathbb{E}\left(\theta^2\left(\frac{n_1}{N}, \frac{n_2}{N}, Y_{n_1}, Y_{n_2}\right)\right)} \underbrace{\int\int |\rho(uh, u'h)|\, du\, du'}_{= \int\int \frac{1}{h^2} |\rho(v,w)|\, dv\, dw}$$

$$= \frac{1}{h^{|\nu|+2}\sqrt{I}} \underbrace{\frac{1}{(2\pi)^2} \sqrt{\max_{n_1 \neq n_2} \mathbb{E}\left(\theta^2\left(\frac{n_1}{N}, \frac{n_2}{N}, Y_{n_1}, Y_{n_2}\right)\right)} \int\int |\rho(v,w)|\, dv\, dw}_{\text{bounded by assumption (f)}}$$

$$= O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)$$

Part B: For evaluating B we show that $\mathbb{E}(\Theta_I(t_1,t_2)) = \gamma(t_1,t_2) + O(h^{l-|\nu|}) + O\left(\frac{1}{h^{|\nu|+3}N}\right)$ uniformly for $t_1, t_2 \in T$. Like in the previous proof the sum expression is approximated by an integral with an error term of order $O\left(\frac{1}{h^{|\nu|+3}N}\right)$ (for calculation see (6.4) below). After a substitution in both dimensions the integral is examined via a multi-dimensional Taylor expansion (see e.g. Heuser [2002]). Set $\gamma^*(t_1,t_2) := \int \theta\left(t_1, t_2, y_1, y_2\right) g_2\left(y_1, y_2; t_1, t_2\right)\, dy$ in order to simplify the notation.

$$\mathbb{E}(\Theta_I(t_1,t_2))$$

$$= \frac{1}{h^{|\nu|+2}N(N-1)} \mathbb{E}\left(\sum_{n_1 \neq n_2} \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, y_1, y_2\right) \mathcal{K}\left(\frac{\frac{n_1}{N} - t_1}{h}, \frac{\frac{n_2}{N} - t_2}{h}\right)\right)$$

$$= \frac{1}{h^{|\nu|+2}N(N-1)} \sum_{n_1 \neq n_2} \int \theta\left(\frac{n_1}{N}, \frac{n_2}{N}, y_1, y_2\right) g_2\left(y_1, y_2; \frac{n_1}{N}, \frac{n_2}{N}\right) \mathcal{K}\left(\frac{\frac{n_1}{N} - t_1}{h}, \frac{\frac{n_2}{N} - t_2}{h}\right) dy$$

$$= \frac{1}{h^{|\nu|+2}N(N-1)} \sum_{n_1 \neq n_2} \gamma^*\left(\frac{n_1}{N}, \frac{n_2}{N}, y_1, y_2\right) \mathcal{K}\left(\frac{\frac{n_1}{N} - t_1}{h}, \frac{\frac{n_2}{N} - t_2}{h}\right)$$

$$= \frac{1}{h^{|\nu|+2}} \int\int \gamma^*\left(t_1', t_2'\right) \mathcal{K}\left(\frac{t_1' - t_1}{h}, \frac{t_2' - t_2}{h}\right) dt_1'\, dt_2' + O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

(sum replaced by integral)

$$= \frac{1}{h^{|\nu|}} \int\int \gamma^*\left(t_1 + vh, t_2 + wh\right) \mathcal{K}(v,w)\, dv\, dw + O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

$$=\frac{1}{h^{|\nu|}}\int\int\left[\sum_{j=0}^{l-1}\sum_{\substack{j_1,j_2\\j_1+j_2=j}}\frac{\partial^{j_1}}{\partial^{j_1}t_1}\frac{\partial^{j_2}}{\partial^{j_2}t_2}\gamma^*(t_1,t_2)\frac{h^j v^{j_1}w^{j_2}}{j_1!j_2!}+\sum_{\substack{j_1,j_2\\j_1+j_2=l}}\frac{\partial^{j_1}}{\partial^{j_1}t_1}\frac{\partial^{j_2}}{\partial^{j_2}t_2}\gamma^*(\xi_1,\xi_2)\frac{h^l v^{j_1}w^{j_2}}{j_1!j_2!}\right]$$

$$\times\mathcal{K}(v,w)\,dv\,dw+O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

$$=\frac{1}{h^{|\nu|}}\left[\frac{\partial^{\nu_1}}{\partial^{\nu_1}t_1}\frac{\partial^{\nu_2}}{\partial^{\nu_2}t_2}\gamma^*(t_1,t_2)\frac{h^{|\nu|}}{|\nu|!}\underbrace{\int\int v^{\nu_1}w^{\nu_2}\mathcal{K}(v,w)\,dv\,dw}_{(-1)^{|\nu|}|\nu|!}\right.$$

$$\left.+\sum_{\substack{j_1,j_2\\j_1+j_2=l}}\underbrace{\frac{\partial^{j_1}}{\partial^{j_1}t_1}\frac{\partial^{j_2}}{\partial^{j_2}t_2}\gamma^*(\xi_1,\xi_2)}_{\text{bounded}}\frac{h^l}{j_1!j_2!}\underbrace{\int\int v^{j_1}w^{j_2}\mathcal{K}(v,w)\,dv\,dw}_{\text{at least one}\neq 0}\right]+O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

$$=\gamma(t_1,t_2)+\underbrace{O(h^{l-|\nu|})+O\left(\frac{1}{h^{|\nu|+3}N}\right)}_{\text{independent of }t\in T^2}$$

The error of discretization is evaluated like in the Lemma before, e.g. the integral is first inserted into the sum expression and afterwards the term is divided into two parts E_1 and E_2, that are evaluated using the properties of the kernel function and of γ^*:

$$\frac{1}{h^{|\nu|+2}N(N-1)}\sum_{n_1\neq n_2}\gamma^*\left(\frac{n_1}{N},\frac{n_2}{N}\right)\mathcal{K}\left(\frac{\frac{n_1}{N}-t_1}{h},\frac{\frac{n_2}{N}-t_2}{h}\right)\tag{6.4}$$

$$=\frac{1}{h^{|\nu|+2}}\int\int\gamma^*(t_1',t_2')\mathcal{K}\left(\frac{t_1'-t_1}{h},\frac{t_2'-t_2}{h}\right)dt_2'\,dt_1'$$

$$+\frac{1}{h^{|\nu|+2}}\sum_{n_1\neq n_2}\int_{\frac{n_1-1}{N}}^{\frac{n_1}{N}}\int_{\frac{n_2-1}{N}}^{\frac{n_2}{N}}\gamma^*\left(\frac{n_1}{N},\frac{n_2}{N}\right)\mathcal{K}\left(\frac{\frac{n_1}{N}-t_1}{h},\frac{\frac{n_2}{N}-t_2}{h}\right)dt_2'\,dt_1'$$

$$-\frac{1}{h^{|\nu|+2}}\sum_{n_1\neq n_2}\int_{\frac{n_1-1}{N}}^{\frac{n_1}{N}}\int_{\frac{n_2-1}{N}}^{\frac{n_2}{N}}\gamma^*(t_1',t_2')\mathcal{K}\left(\frac{t_1'-t_1}{h},\frac{t_2'-t_2}{h}\right)dt_2'\,dt_1'$$

$$=\frac{1}{h^{|\nu|+2}}\int\int\gamma^*(t_1',t_2')\mathcal{K}\left(\frac{t_1'-t_1}{h},\frac{t_2'-t_2}{h}\right)dt_2'\,dt_1'$$

$$+\underbrace{\frac{1}{h^{|\nu|+2}}\sum_{n_1\neq n_2}\int_{\frac{n_1-1}{N}}^{\frac{n_1}{N}}\int_{\frac{n_2-1}{N}}^{\frac{n_2}{N}}\left[\gamma^*\left(\frac{n_1}{N},\frac{n_2}{N}\right)-\gamma^*(t_1',t_2)\right]\mathcal{K}\left(\frac{\frac{n_1}{N}-t_1}{h},\frac{\frac{n_2}{N}-t_2}{h}\right)dt_2'\,dt_1'}_{=:E_1}$$

$$+\underbrace{\frac{1}{h^{|\nu|+2}}\sum_{n_1\neq n_2}\int_{\frac{n_1-1}{N}}^{\frac{n_1}{N}}\int_{\frac{n_2-1}{N}}^{\frac{n_2}{N}}\left[\mathcal{K}\left(\frac{\frac{n_1}{N}-t_1}{h},\frac{\frac{n_2}{N}-t_2}{h}\right)-\mathcal{K}\left(\frac{t_1'-t_1}{h},\frac{t_2'-t_2}{h}\right)\right]\gamma^*(t_1',t_2')\,dt_2'\,dt_1'}_{=:E_2}$$

(6.5)

E_1 is evaluated based on the condition that γ^* is Lipschitz continuous and therefore

$$\left|\gamma^*\left(\frac{n_1}{N},\frac{n_2}{N}\right)-\gamma^*(t_1',t_2')\right| \leq c||(\frac{n_1}{N},\frac{n_2}{N})-(t_1',t_2')|| \leq \frac{c}{N} \text{ (assumption (c))} :$$

$$E_1 = \frac{1}{h^{|\nu|+2}}\sum_{n_1\neq n_2}\int_{\frac{n_1-1}{N}}^{\frac{n_1}{N}}\int_{\frac{n_2-1}{N}}^{\frac{n_2}{N}}\left[\gamma^*\left(\frac{n_1}{N},\frac{n_2}{N}\right)-\gamma^*(t_1',t_2')\right]\mathcal{K}\left(\frac{\frac{n_1}{N}-t_1}{h},\frac{\frac{n_2}{N}-t_2}{h}\right)dt_2'\,dt_1'$$

$$= O\left(\frac{1}{h^{|\nu|+2}N}\right)$$

For the evaluation of E_2 the mean value theorem in several variables is used (with $\nabla\mathcal{K}$ being the gradient of $\mathcal{K}$):

E_2

$$\leq \frac{1}{h^{|\nu|+2}}\sum_{n_1\neq n_2}\int_{\frac{n_1-1}{N}}^{\frac{n_1}{N}}\int_{\frac{n_2-1}{N}}^{\frac{n_2}{N}}\left|\left[\mathcal{K}\left(\frac{\frac{n_1}{N}-t_1}{h},\frac{\frac{n_2}{N}-t_2}{h}\right)-\mathcal{K}\left(\frac{t_1'-t_1}{h},\frac{t_2'-t_2}{h}\right)\right]\gamma^*(t_1',t_2')\right|dt_2'\,dt_1'$$

$$\leq \frac{1}{h^{|\nu|+2}}\sum_{n_1\neq n_2}\int_{\frac{n_1-1}{N}}^{\frac{n_1}{N}}\int_{\frac{n_2-1}{N}}^{\frac{n_2}{N}}\left|\nabla\mathcal{K}(\xi_{nm})\left(\frac{\frac{n_1}{N}-t_1'}{h},\frac{\frac{n_2}{N}-t_2'}{h}\right)^T\gamma^*(t_1',t_2')\right|dt_2'\,dt_1'$$

$$\text{for}\quad \xi_{nm}\in\left\{\left(\frac{\frac{n_1}{N}-t_1}{h},\frac{\frac{n_2}{N}-t_2}{h}\right)+\eta\left(\frac{\frac{n_1}{N}-t_1'}{h},\frac{\frac{n_2}{N}-t_2'}{h}\right),\eta\in[0,1]\right\}$$

$$\leq \frac{1}{h^{|\nu|+2}}\sum_{n_1\neq n_2}\max_{n_1\neq n_2}||\nabla\mathcal{K}(\xi_{mn})||\frac{2}{hN}\int_{\frac{n_1-1}{N}}^{\frac{n_1}{N}}\int_{\frac{n_2-1}{N}}^{\frac{n_2}{N}}|\gamma^*(t_1',t_2')|\,dt_2'\,dt_1'$$

$$= O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

Now it is shown that:

Part A: $\mathbb{E}\left(\sup_{t\in T}|\Theta_I(t_1,t_2)-\mathbb{E}(\Theta_I(t_1,t_2))|\right)=O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)$

Part B: $\mathbb{E}(\Theta_I(t_1,t_2))=\gamma(t_1,t_2)+O(h^{l-|\nu|})+O(\frac{1}{h^{|\nu|+3}N})$

For (6.3) we therefore obtain the inequality:

$$\mathbb{E}|\tau_I| = O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)+O(h^{l-|\nu|})+O\left(\frac{1}{h^{|\nu|+3}N}\right)=O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right),$$

because $h^{l-|\nu|}=\underbrace{\sqrt{Ih^{2l+4}}}_{\text{bounded}}\frac{1}{h^{|\nu|+2}\sqrt{I}}=O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)$ and $\frac{1}{N}=O(h^{l+3})$ (assumption (c)).

Using Markov's inequality one obtains:

$$\mathbb{E}|\tau_I| = O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)\Rightarrow\tau_I=O_P\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)$$

Again, in case of a non-random function $\Theta_I(t)$ Part A vanishes such that the final rate is only the better rate of Part B.

□

Based on Lemmas 7 and 8 consistency results for mean and covariance functions can be obtained:

6.2.2 Mean, covariance and principal components

Theorem 9 (Theorem for mean estimation, see also Yao et al. [2005]). *Under assumptions (a),(b) and (d) of Lemma 7 with a kernel of order* $(\nu, l) = (0, 2)$ *and*

(c) $h \to 0, Ih^2 \to \infty, Ih^6 = O(1), \frac{1}{N} = O(h^5)$ *for* $I \to \infty$,

the estimator for the mean function as defined in Section 2.2 fulfills the following uniform convergence rate:

$$\sup_{t \in T} |\hat{\mu}(t) - \mu(t)| = O_P\left(\frac{1}{\sqrt{I}h}\right) \tag{6.6}$$

Proof (see also Yao et al. [2005]). The local linear estimator $\hat{\mu}(t)$ can explicitly be written as (with $w_n = \mathcal{K}\left(\frac{t-\frac{n}{N}}{h}\right)$)

$$\hat{\mu}(t) = \hat{\beta}_0(t) = \frac{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n Y_n^i}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n} - \frac{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n}\hat{\beta}_1(t),$$

where

$$\hat{\beta}_1(t) = \frac{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)Y_n^i - \frac{\left(\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)\right)\left(\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n Y_n^i\right)}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n}}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)^2 - \frac{\left(\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)\right)^2}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n}}.$$

We use Lemma 7 to evaluate the terms in $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$. Therefore we define the derivative kernels $\mathcal{K}_t(t) = \frac{-t\mathcal{K}(t)}{\sigma_t^2}$ with order $(1, 3)$ and $\mathcal{K}_{t^2}(t) = \frac{t^2\mathcal{K}(t)}{\sigma_t^2}$ with order $(0, 2)$. σ_t is a scaling factor.

Using Lemma 7 one can now calculate convergence rates for the single terms. In order to demonstrate the calculations, the lemma is applied for two terms step-by-step:

Set $\psi_1(t, y) = 1$ and use $\mathcal{K}$ of order $(0, 2)$ as kernel function, then

$$\mu_1(t) = \int 1\, g(y;t)\, dy = 1 \text{ and } \Psi_{1I}(t) = \frac{1}{IhN}\sum_{i=1}^{I}\sum_{n=1}^{N} 1\, w_n$$

and therefore, as this term has no random parts,

$$\sup_{t \in T}\left|\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n - 1\right| = O(h^2).$$

As another example set $\psi_2(t,y) = y$ and use the kernel $\mathcal{K}_t$ of order $(1,3)$. Then

$$\mu_2(t) = \int y g(y;t)\,dy = \mu'(t) \text{ and}$$

$$\Psi_{2I}(t) = \frac{1}{Ih^2N}\sum_{i=1}^{I}\sum_{n=1}^{N} Y_n^i \left(\frac{\frac{n}{N}-t}{h\sigma_t^2}\right) w_n = \frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)Y_n^i.$$

In this case the lemma says that

$$\sup_{t\in T}\left|\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)Y_n^i - \mu'(t)\right| = O_P\left(\frac{1}{\sqrt{I}h^2}\right).$$

In summary we obtain the following convergence rates for the single terms:

$$\sup_{t\in T}\left|\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n - 1\right| = O(h^2) \text{ (using kernel } \mathcal{K})$$

$$\sup_{t\in T}\left|\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t) - 0\right| = O(h^2) \text{ (using kernel } \mathcal{K}_t)$$

$$\sup_{t\in T}\left|\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)^2 - 1\right| = O(h^2) \text{ (using kernel } \mathcal{K}_{t^2})$$

$$\sup_{t\in T}\left|\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n Y_n^i - \mu(t)\right| = O_P\left(\frac{1}{\sqrt{Ih}}\right) \quad \text{(using kernel } \mathcal{K})$$

$$\sup_{t\in T}\left|\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)Y_n^i - \mu'(t)\right| = O_P\left(\frac{1}{\sqrt{I}h^2}\right) \quad \text{(using kernel } \mathcal{K}_t)$$

Now $\hat\beta_1(t)$ can be rewritten in order to directly apply the single term rates. We obtain

$$\hat\beta_1(t) = \frac{\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)Y_n^i - \frac{\left(\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)\right)\left(\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n Y_n^i\right)}{\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n}}{\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)^2 - \frac{h^2\sigma_t^2\left(\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n(\frac{n}{N}-t)\right)^2}{\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N} w_n}}$$

$$= \frac{\left[\mu'(t) + O_P\left(\frac{1}{\sqrt{I}h^2}\right)\right] - \frac{\left[0+O(h^2)\right]\left[\mu(t)+O_P\left(\frac{1}{\sqrt{Ih}}\right)\right]}{[1+O(h^2)]}}{[1+O(h^2)] - \frac{h^2\sigma_t^2[0+O(h^2)]^2}{[1+O(h^2)]}}$$

$$= \mu'(t) + O_P\left(\frac{1}{\sqrt{I}h^2}\right) + O(h^2) = \mu'(t) + O_P\left(\frac{1}{\sqrt{I}h^2}\right) \text{ by assumption (c)}$$

uniformly in t, i.e.

$$\sup_{t\in T}|\hat{\beta}_1(t)-\mu'(t)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right)$$

Now $\hat{\beta}_0(t)$ can be evaluated in the same manner:

$$\begin{aligned}
\hat{\beta}_0(t) &= \frac{\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N}w_n Y_n^i}{\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N}w_n} - h^2\sigma_t^2\frac{\frac{1}{Ih^3\sigma_t^2}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N}w_n(\frac{n}{N}-t)}{\frac{1}{Ih}\sum_{i=1}^{I}\frac{1}{N}\sum_{n=1}^{N}w_n}\hat{\beta}_1(t) \\
&= \frac{\left[\mu(t)+O_P\left(\frac{1}{\sqrt{Ih}}\right)\right]}{[1+O(h^2)]} - h^2\sigma_t^2\frac{[0+O(h^2)]}{[1+O(h^2)]}\left[\mu'(t)+O_P\left(\frac{1}{\sqrt{I}h^2}\right)\right] \\
&= \mu(t)+O_P\left(\frac{1}{\sqrt{Ih}}\right)+h^4 O_P\left(\frac{1}{\sqrt{I}h^2}\right)
\end{aligned}$$

uniformly in t and therefore we obtain the final result

$$\sup_{t\in T}|\hat{\mu}(t)-\mu(t)| = O_P\left(\frac{1}{\sqrt{Ih}}\right).$$

□

Theorem 10 (Theorem for covariance estimation, see also Yao et al. [2005])**.** *Assume that (a),(b) and (d) of Lemma 8 are valid here as well and that*

(c) $h\to 0, Ih^4\to\infty, Ih^8 = O(1), \frac{1}{N} = O(h^7)$ *for* $I\to\infty$.

Further $\mathcal{K}$ *is assumed to be a two-dimensional product kernel of a kernel* $\mathcal{K}^*$ *of order* $(0,2)$, *e.g.* $\mathcal{K}(t_1,t_2) = \mathcal{K}^*(t_1)\mathcal{K}^*(t_2)$ *for* $t_1,t_2\in T$. $\mathcal{K}$ *is therefore of order* $((0,0),2)$ *(see calculation* (2.7)*). The estimator for the covariance function as defined in Section 2.2 fulfills the following uniform convergence rate:*

$$\sup_{t_1,t_2\in T}|\hat{G}(t_1,t_2)-G(t_1,t_2)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right) \tag{6.7}$$

Proof. As the estimation of the covariance function is based on the raw covariances with estimated mean function $G_i(\frac{n_1}{N},\frac{n_2}{N}) = (Y_{n_1}^i-\hat{\mu}(\frac{n_1}{N}))(Y_{n_2}^i-\hat{\mu}(\frac{n_2}{N}))$ instead of $\tilde{G}_i(\frac{n_1}{N},\frac{n_2}{N}) = (Y_{n_1}^i-\mu(\frac{n_1}{N}))(Y_{n_2}^i-\mu(\frac{n_2}{N}))$, we will first show that $G_i(\frac{n_1}{N},\frac{n_2}{N})$ is asymptotically equivalent to $\tilde{G}_i(\frac{n_1}{N},\frac{n_2}{N})$ and work with $\tilde{G}_i(\frac{n_1}{N},\frac{n_2}{N})$ in the further proof.
This can be seen observing that (for $t_1,t_2\in T$):

$$\begin{aligned}
G_i(t_1,t_2) = \tilde{G}_i(t_1,t_2) &+ \left(Y_{t_1}^i-\mu(t_1)\right)(\mu(t_2)-\hat{\mu}(t_2)) \\
&+ \left(Y_{t_2}^i-\mu(t_2)\right)(\mu(t_1)-\hat{\mu}(t_1)) \\
&+ (\mu(t_1)-\hat{\mu}(t_1))(\mu(t_2)-\hat{\mu}(t_2))
\end{aligned}$$

As $\mathrm{Var}(Y_t)$ is finite for all $t \in T$, $\left(Y_{t_1}^i - \mu\left(t_1\right)\right)$ and $\left(Y_{t_2}^i - \mu\left(t_2\right)\right)$ are stochastically bounded. Further $\sup_{t\in T} |\mu(t) - \hat{\mu}(t)| = O_P\left(\frac{1}{\sqrt{I}h}\right)$ such that

$$\sup_{t_1,t_2\in T} \left| G_i\left(t_1,t_2\right) - \tilde{G}_i\left(t_1,t_2\right) \right| = O_P\left(\frac{1}{\sqrt{I}h}\right)$$

which shows the asymptotic equivalence of G_i and $\tilde{G}_i$.
The local linear estimator $\hat{G}(t_1,t_2)$ can explicitly be written as

$$\begin{aligned}\hat{G}(t_1,t_2) = \hat{\beta}_0(t_1,t_2) = & \frac{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2} \tilde{G}\left(\frac{n_1}{N}, \frac{n_2}{N}\right)}{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}} \\ & - \frac{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t_1)}{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}} \hat{\beta}_{11}(t_1,t_2) \\ & - \frac{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_2}{N} - t_2)}{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}} \hat{\beta}_{12}(t_1,t_2)\end{aligned}$$

where $w_{n_1n_2} = \mathcal{K}\left(\frac{t_1-\frac{n_1}{N}}{h}, \frac{t_2-\frac{n_2}{N}}{h}\right)$ and

$$\hat{\beta}_{11}(t_1,t_2) = D_1 + E_1\hat{\beta}_{12}(t_1,t_2)$$

$$\hat{\beta}_{12}(t_1,t_2) = D_2 + E_2\hat{\beta}_{11}(t_1,t_2)$$

with

$$\begin{aligned} D_1 = & \frac{1}{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t_1)^2 - \frac{\left(\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t_1)\right)^2}{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}}} \\ & * \left[\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t_1)\tilde{G}\left(\frac{n_1}{N}, \frac{n_2}{N}\right)\right. \\ & \left. - \frac{\left(\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t_1)\right)\left(\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}\tilde{G}\left(\frac{n_1}{N}, \frac{n_2}{N}\right)\right)}{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}}\right] \end{aligned}$$

$$\begin{aligned} E_1 = & \frac{1}{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t_1)^2 - \frac{\left(\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t)\right)^2}{\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}}} \\ & * \left[\left(\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t_1)\right)\left(\frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_2}{N} - t_2)\right)\right. \\ & \left. - \frac{1}{I}\sum_i \frac{1}{N(N-1)} \sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N} - t_1)(\frac{n_2}{N} - t_2)\right]. \end{aligned}$$

D_2 and E_2 are composed analogously exchanging t_1 with t_2 and n_1 with n_2.

We can rewrite:

$$\hat{\beta}_{11}(t_1,t_2) = \frac{D_1 + E_1 D_2}{1 - E_1 E_2}$$

$$\hat{\beta}_{12}(t_1,t_2) = \frac{D_2 + E_2 D_1}{1 - E_1 E_2}$$

Hence we first evaluate the terms D_1, D_2, E_1 and E_2 and afterwards conclude the convergence rates of $\hat{\beta}_{11}(t_1,t_2)$ and $\hat{\beta}_{12}(t_1,t_2)$.

In order to evaluate the terms we define the derivative kernels $\mathcal{K}_{t_1}(t_1,t_2) = \frac{-t_1\mathcal{K}(t_1,t_2)}{\sigma_{t_1}^2}$ of order $((1,0),3)$, $\mathcal{K}_{t_1{}^2}(t_1,t_2) = \frac{t_1^2\mathcal{K}(t_1,t_2)}{\sigma_{t_1}^2}$ of order $((0,0),2)$ (as well as the analogue kernels for t_2) and $\mathcal{K}_{t_1,t_2}(t_1,t_2) = \frac{t_1 t_2\mathcal{K}(t_1,t_2)}{\sigma_{t_1,t_2}^2}$ of order $((1,1),4)$.

Using Lemma 8 we can obtain the convergence rates for the single terms. For example let $\theta_1(t_1,t_2,y_1,y_2) := 1$, then using the kernel $\mathcal{K}$ of order $((0,0),2)$,

$$\theta_1(t) = \int\int 1\, g_2(y_1,y_2;t_1,t_2)\,dy_1\,dy_2 = 1 \text{ and } \Theta_{1I}(t) = \frac{1}{Ih^2 N(N-1)}\sum_i \sum_{n_1\neq n_2} 1 w_{n_1 n_2}.$$

According to the remark of Lemma 8 we therefore obtain: ´

$$\sup_{t_1,t_2\in T} |\frac{1}{Ih^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1 n_2} - 1| = O(h^2)$$

The following overview shows only terms in t_1 and mixed terms (the analogue terms in t_2 have the same convergence rates). The kernels used are $\mathcal{K}$, $\mathcal{K}_{t_1}$, $\mathcal{K}_{t_1,t_2}$, $\mathcal{K}_{t_1^2}$, $\mathcal{K}$ and $\mathcal{K}_{t_1}$:

$$\sup_{t_1,t_2\in T}\left|\frac{1}{Ih^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1 n_2} - 1\right| = O(h^2)$$

$$\sup_{t_1,t_2\in T}\left|\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1 n_2}(\frac{n_1}{N} - t_1) - 0\right| = O(h^2)$$

$$\sup_{t_1,t_2\in T}\left|\frac{1}{Ih^6\sigma_{t_1,t_2}}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1 n_2}(\frac{n_1}{N} - t_1)(\frac{n_2}{N} - t_2) - 0\right| = O(h^2)$$

$$\sup_{t_1,t_2\in T}\left|\frac{1}{Ih^4\sigma_{t_1}^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1 n_2}(\frac{n_1}{N} - t_1)^2 - 1\right| = O(h^2)$$

$$\sup_{t_1,t_2\in T}\left|\frac{1}{Ih^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1 n_2}\tilde{G}\left(\frac{n_1}{N},\frac{n_2}{N}\right) - \mu(t_1,t_2)\right| = O_P\left(\frac{1}{\sqrt{I}h^2}\right)$$

$$\sup_{t_1,t_2\in T}\left|\frac{1}{Ih^4\sigma_{t_1}^2}\sum_i\frac{1}{N(N-1)}\sum_{n_1\neq n_2}w_{n_1n_2}(\frac{n_1}{N}-t_1)\tilde{G}\left(\frac{n_1}{N},\frac{n_2}{N}\right)-\mu_t(t_1,t_2)\right|=O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

D_1 therefore has the following rate:

$$D_1=\frac{\left[G_{t_1}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right]-\frac{\left[0+O(h^2)\right]\left[G(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^2}\right)\right]}{[1+O(h^2)]}}{[1+O(h^2)]-\frac{h^2\sigma_t^2[0+O(h^2)]^2}{[1+O(h^2)]}}$$

$$=G_{t_1}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

uniformly in t_1,t_2, i.e.

$$\sup_{t_1,t_2\in T}|D_1-G_{t_1}(t_1,t_2)|=O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

and analogously

$$\sup_{t_1,t_2\in T}|D_2-G_{t_2}(t_1,t_2)|=O_P\left(\frac{1}{\sqrt{I}h^3}\right).$$

For E_1 we obtain:

$$E_1=\frac{h^4\sigma_{t_2}^2\left[0+O(h^2)\right]\left[0+O(h^2)\right]-h^2\frac{\sigma_{t_1t_2}^2}{\sigma_{t_1}^2}\left[0+O(h^2)\right]}{[1+O(h^2)]-\frac{h^2\sigma_{t_1}^2[0+O(h^2)]^2}{[1+O(h^2)]}}=O(h^4)$$

uniformly in t_1,t_2, i.e.

$$\sup_{t_1,t_2\in T}|E_1|=O(h^4)\quad\text{and analogously}\quad\sup_{t_1,t_2\in T}|E_2|=O(h^4).$$

With this information, $\hat{\beta}_{11}(t_1,t_2)$ and $\hat{\beta}_{12}(t_1,t_2)$ can be evaluated as follows (uniformly in t_1,t_2):

$$\hat{\beta}_{11}(t_1,t_2)=\frac{\left[G_{t_1}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right]+\left[0+O(h^4)\right]\left[G_{t_2}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right]}{1-[0+O(h^4)]\,[0+O(h^4)]}$$

$$=G_{t_1}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)+O(h^4)=G_{t_1}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

$$\hat{\beta}_{12}(t_1,t_2)=G_{t_2}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

For $\hat{\beta}_0(t_1,t_2)$ we now obtain:

$$\begin{aligned}
\hat{\beta}_0(t_1,t_2) =& \frac{\frac{1}{Ih^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1n_2}\tilde{G}\left(\frac{n_1}{N},\frac{n_2}{N}\right)}{\frac{1}{Ih^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1n_2}} \\
&- \frac{h^2\sigma_t^2\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_1}{N}-t_1)}{\frac{1}{Ih^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1n_2}}\hat{\beta}_{11}(t_1,t_2) \\
&- \frac{h^2\sigma_{t_2}^2\frac{1}{Ih^4\sigma_{t_2}^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1n_2}(\frac{n_2}{N}-t_2)}{\frac{1}{Ih^2}\sum_i \frac{1}{N(N-1)}\sum_{n_1\neq n_2} w_{n_1n_2}}\hat{\beta}_{12}(t_1,t_2) \\
=& \frac{\left[G(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^2}\right)\right]}{[1+O(h^2)]} - \frac{h^2\left[0+O(h^2)\right]}{[1+O(h^2)]}\left[G_{t_1}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right] \\
&- \frac{h^2\left[0+O(h^2)\right]}{[1+O(h^2)]}\left[G_{t_2}(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right] \\
=& G(t_1,t_2)+O_P\left(\frac{1}{\sqrt{I}h^2}\right)+O(h^4)+h^4O_P\left(\frac{1}{\sqrt{I}h^3}\right)
\end{aligned}$$

uniformly in t_1, t_2, such that we obtain the final uniform convergence rate

$$\sup_{t_1,t_2\in T}|\hat{G}(t_1,t_2)-G(t_1,t_2)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right).$$

□

For proving the rates for eigenvalues and eigenfunctions, we introduce Weyl's lemma first (compare Heuser [2006, Chapter 5]):

Lemma 11 (Weyl's eigenvalue inequality). *Let X_1, X_2 and Z be symmetric, compact operators with $Z = X_1 + X_2$ and let $\lambda_k^{X_1}$, $\lambda_k^{X_2}$ and λ_k^Z be positive, monotonically decreasing eigenvalues of X_1, X_2 and Z. Then the following inequality holds:*

$$\lambda_{k+l-1}^Z \leq \lambda_k^{X_1} + \lambda_l^{X_2} \tag{6.8}$$

Proof. Weyl's eigenvalue inequality is a consequence of Courant's minimax principle (Heuser [2006, Chapter 5]), which states under the same conditions as Lemma 11 that

$$\lambda_k^Z = \min_S \sup_{0\neq x\in S^\perp} \frac{\langle Zx, x\rangle}{\langle x, x\rangle} \tag{6.9}$$

for all $(k-1)$ dimensional subspaces S of $L_2(T)$. The minimum is reached for $S := [\rho_1, \ldots, \rho_{k-1}]$. Let $\rho_k^{X_1}$ and $\rho_k^{X_2}$ be the eigenfunctions of X_1 and X_2 and furthermore, define the spans of eigenfunctions $S_1 = [\rho_1^{X_1}, \ldots, \rho_{k-1}^{X_1}]$, $S_2 = [\rho_1^{X_2}, \ldots, \rho_{l-1}^{X_2}]$ and $S_{12} = [\rho_1^{X_1}, \ldots, \rho_{k-1}^{X_1}, \rho_1^{X_2}, \ldots, \rho_{l-1}^{X_2}]$.

Then inequality (6.8) follows with

$$\lambda^Z_{k+l-1} \leq \sup_{x \in S_{12}^\perp} \langle Zx, x\rangle \leq \sup_{x \in S_{12}^\perp} \langle X_1 x, x\rangle + \sup_{x \in S_{12}^\perp} \langle X_2 x, x\rangle$$

$$\leq \sup_{x \in S_1^\perp} \langle X_1 x, x\rangle + \sup_{x \in S_2^\perp} \langle X_2 x, x\rangle = \lambda_k^{X_1} + \lambda_l^{X_2}.$$

The first inequality follows due to Courants' principle (6.9), the second simply because $Z = X_1 + X_2$. For the third inequality note that $S_1 \subset S_{12}$ and therefore $S_{12}^\perp \subset S_1^\perp$ (and the same argumentation for S_2). The last equality follows again due to Courants' principle or the variance maximization criterion Theorem 4. □

Now we proceed with the Theorem with the rates for the eigenanalysis estimation:

Theorem 12 (Theorem for eigenanalysis estimation, see also Yao et al. [2005])**.** *Under the assumptions of Theorems 9 and 10 we obtain the following consistency rate for eigenvalues ($k \in \mathbb{N}$):*

$$|\hat{\lambda}_k - \lambda_k| = O_P\left(\frac{1}{\sqrt{I}h^2}\right). \tag{6.10}$$

If λ_k has multiplicity 1, the appropriately standardized estimator for the eigenfunction $\hat{\rho}_k$ has the following convergence rate in the L_2 sense

$$||\hat{\rho}_k - \rho_k||_{L_2} = O_P\left(\frac{1}{\sqrt{I}h^2}\right) \tag{6.11}$$

and fulfills furthermore the uniform consistency rate

$$\sup_{t \in T} |\hat{\rho}_k(t) - \rho_k(t)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right). \tag{6.12}$$

Proof. We consider the Hilbert space $H = \{f : T \to \mathbb{R} | \int |f(t)|^2\, dt < \infty\}$ with the inner product $\langle f, g\rangle = \int |f(t)g(t)|\, dt$ and norm $||f||_{L_2} = \sqrt{\int |f(t)|^2\, dt}$.
We further define the space $\sigma_2(H)$ of Hilbert-Schmidt operators on H.
An operator $T : H \to H$ is called Hilbert-Schmidt operator on H, if for an orthonormal system $\{e_i\}$ of H, the Hilbert-Schmidt norm is finite, e.g. $\sum_i ||Te_i||^2 < \infty$. The inner product on $F := \sigma_2(H)$ for $T_1, T_2 \in \sigma_2(H)$ is given by $\langle T_1, T_2\rangle_F = \sum_j \langle T_1 e_j, T_2 e_j\rangle_H$. We further consider the operator $f \otimes g : H \to H$ with $(f \otimes g)(h) = \langle f, h\rangle_H g$.
One can show that the real and estimated covariance operators

$$G(f) = \int G(t_1, t_2) f(t_1)\, dt_1$$

$$\hat{G}(f) = \int \hat{G}(t_1, t_2) f(t_1)\, dt_1$$

have a finite Hilbert-Schmidt norm and therefore are Hilbert-Schmidt operators (see Heuser [2006, Chapter 12]).

Further we have the following relationship between the supremum norm (for which we know the order of convergence from Theorem 10) and the Hilbert-Schmidt norm (for the first equality see Heuser [2006, Chapter 12]):

$$||\hat{G} - G||_F = ||\hat{G} - G||_{L_2} = \sqrt{\int\int \left|\hat{G}(t_1,t_2) - G(t_1,t_2)\right|^2 dt_1\, dt_2}$$

$$\leq \sup_{t_1,t_2 \in T} \left|\hat{G}(t_1,t_2) - G(t_1,t_2)\right| \underbrace{\sqrt{\int\int 1\, dt_1\, dt_2}}_{|T|} = O_P\left(\frac{1}{\sqrt{I}h^2}\right)$$

In order to derive the rate for the eigenvalue estimation (6.10), we apply Weyl's eigenvalue inequality (Lemma 11). We chose this approach varying to Yao et al. [2005], because the proof of Yao et al. [2005] already needs for the eigenvalue rate the assumption that the eigenvalues have multiplicity one, which is not necessary.

To apply the lemma, observe that G has eigenvalues $\lambda_k \geq 0$ and $\hat{G}$ eigenvalues $\hat{\lambda}_k \geq 0$. Now define $R_1 := \hat{G} - G$ and $R_2 := G - \hat{G}$, such that $\hat{G} = G + R_1$ and $G = \hat{G} + R_2$. Further let λ_1^1 be the greatest (positive) eigenvalue of R_1 and λ_1^2 the greatest (positive) eigenvalue of R_2. Hence Lemma 11 tells us that

$$\hat{\lambda}_k \leq \lambda_k + \lambda_1^1$$

$$\lambda_k \leq \hat{\lambda}_k + \lambda_1^2.$$

Equation (6.10) follows with

$$|\lambda_k - \hat{\lambda}_k| \leq \max(\lambda_1^1, \lambda_1^2) \leq ||\hat{G} - G||_{L_2} = O_P\left(\frac{1}{\sqrt{I}h^2}\right).$$

Now define I_k as the set of indices with the same eigenvalue ($I_k := \{j : \lambda_j = \lambda_k\}$) and I' as the set of indices of eigenvalues with multiplicity 1 ($I' = \{k : |I_k| = 1\}$) .

P_k and $\hat{P}_k$ are the theoretic respectively estimated projections of operators on the space spanned by $\{\rho_j | j \in I_k\}$:

$$P_k = \sum_{j \in I_k} \rho_j \otimes \rho_j \tag{6.13}$$

$$\hat{P}_k = \sum_{j \in I_k} \hat{\rho}_j \otimes \hat{\rho}_j$$

In order to calculate the order of convergence of the eigenvalues and eigenfunctions we define circles around the eigenvalues in the space of complex numbers which contain no other eigenvalues. For a constant $0 < \varsigma < \min\{|\lambda_j - \lambda_k| : j \notin I_k\}$ let $\Delta_{\varsigma,k} := \{z \in \mathbb{C} : |z - \lambda_k| = \varsigma\}$.

Now we regard the resolvents of G and $\hat{G}$: $R(z) = (G - zI)^{-1}$ and $\hat{R}(z) = (\hat{G} - zI)^{-1}$.

They have the relationship $\hat{R}(z) = R(z)[I + (\hat{G} - G)R(z)]^{-1}$ as can be seen by the following calculation:

$$R(z) = \hat{R}(z)[\underbrace{(\hat{G} - zI)}_{\hat{R}(z)^{-1}} R(z)] = \hat{R}(z)[(G - zI + \hat{G} - G)R(z)]$$

$$= \hat{R}(z)[(R(z)^{-1} + \hat{G} - G)R(z)] = \hat{R}(z)[R(z)^{-1}R(z) + (\hat{G} - G)R(z)]$$

$$= \hat{R}(z)[I + (\hat{G} - G)R(z)]$$

$$\Rightarrow \hat{R}(z) = R(z)[I - (G - \hat{G})R(z)]^{-1} = R(z)\sum_{l=0}^{\infty}[(G - \hat{G})R(z)]^l$$

$$\Rightarrow \hat{R}(z) - R(z) = R(z)\sum_{l=1}^{\infty}\Big((\hat{G} - G)R(z)\Big)^l$$

We can use the sum formula for the geometric sequence, because $||(G - \hat{G})R(z)||_F < 1$ for sufficient large I. Therefore we can deduce (again using the sum formula):

$$||\hat{R}(z) - R(z)||_F \leq ||R(z)||_F \sum_{l=1}^{\infty} ||(\hat{G} - G)||_F^l ||R(z)||_F^l$$

$$= ||R(z)||_F \sum_{l=0}^{\infty} ||(\hat{G} - G)||_F^l ||R(z)||_F^l - ||R(z)||_F$$

$$= \frac{||R(z)||_F}{1 - ||(\hat{G} - G)||_F ||R(z)||_F} - ||R(z)||_F$$

$$= \frac{||\hat{G} - G||_F ||R(z)||_F^2}{1 - ||(\hat{G} - G)||_F ||R(z)||_F}$$

This result can be applied by using the following relationship between projections and resolvents (see Heuser [2002]):

$$P_k = \frac{1}{2\pi i}\int_{\Delta_{\varsigma,k}} R(z)\, dz \tag{6.14}$$

$$\hat{P}_k = \frac{1}{2\pi i}\int_{\Delta_{\varsigma,k}} \hat{R}(z)\, dz.$$

Further define $M_{\varsigma_k} := \sup_{z \in \Delta_{\varsigma,k}} ||R(z)||$. It is $M_{\varsigma_k} < \infty$ because of the definition of $\Delta_{\varsigma,k}$ and because $k \in I'$. Therewith:

$$||\hat{P}_k - P_k||_F \leq \frac{1}{2\pi}\int_{\Delta_{\varsigma,k}} ||\hat{R}(z) - R(z)||_F\, dz = ||\hat{R}(z) - R(z)||_F \underbrace{\frac{1}{2\pi}\int_{\Delta_{\varsigma,k}} 1\, dz}_{\varsigma}$$

$$\leq \varsigma \frac{||\hat{G} - G||M_{\varsigma_k}}{1 - ||\hat{G} - G||M_{\varsigma_k}} = O_P\left(\frac{1}{\sqrt{I}h^2}\right)$$

which shows consistency in the L_2 sense. This step is how far we get if an eigenspace is multi-dimensional, e.g. we can provide rates for the eigenvalue and the projection to the multi-dimensional eigenspace.

For finally showing equation (6.12) consider ρ_k for $k \in I'$, choose $\hat{\rho}_k$ such that $\langle \rho_k, \hat{\rho}_k \rangle_H > 0$ (standardization).

Then

$$\begin{aligned} ||\hat{\rho}_k - \rho_k||_H^2 &= \int (\hat{\rho}_k(s) - \rho_k(s))^2 \, ds \\ &= \int \hat{\rho}_k^2(s) \, ds + \int \rho_k^2(s) \, ds - 2 \int \hat{\rho}_k(s) \rho_k(s) \, ds \\ &= 2(1 - \langle \hat{\rho}_k, \rho_k \rangle_H) \leq 2(1 - \langle \hat{\rho}_k, \rho_k \rangle_H^2) \\ &= 2\left(1 - \sum_j \langle \hat{\rho}_k, \rho_j \rangle_H \langle \rho_k, \rho_j \rangle_H \langle \hat{\rho}_k, \rho_k \rangle_H \right) \qquad (6.15) \\ &= 2\left(1 - \sum_j \langle \langle \hat{\rho}_k, \rho_j \rangle_H \hat{\rho}_k, \langle \rho_k, \rho_j \rangle_H \rho_k \rangle_H \right) \\ &= 2\left(1 - \sum_j \langle (\hat{\rho}_k \otimes \hat{\rho}_k)(\rho_j), (\rho_k \otimes \rho_k)(\rho_j) \rangle_H \right) \\ &= 2\left(1 - \langle \hat{\rho}_k \otimes \hat{\rho}_k, \rho_k \otimes \rho_k \rangle_F \right) \\ &= ||\hat{P}_k - P_k||_F^2 \end{aligned}$$

and therefore equation (6.11) follows.

In order to prove equation (6.12), we first observe using the Cauchy-Schwarz inequality, theorem 10 and equation (6.11) that:

$$\begin{aligned} |\hat{\lambda}_k \hat{\rho}_k(t_2) - \lambda_k \rho_k(t_2)| &= \left| \int \hat{G}(t_1, t_2) \hat{\rho}_k(t_1) \, dt_1 - \int G(t_1, t_2) \rho_k(t_1) \, dt_1 \right| \qquad (6.16) \\ &\leq \int |\hat{G}(t_1, t_2) - G(t_1, t_2)| |\hat{\rho}_k(t_1)| \, dt_1 + \int |G(t_1, t_2)| |\hat{\rho}_k(t_1) - \rho_k(t_1)| \, dt_1 \\ &\leq \sqrt{\int (\hat{G}(t_1, t_2) - G(t_1, t_2))^2 \, dt_1} + \sqrt{\int G^2(t_1, t_2) \, dt_1} ||\hat{\rho}_k - \rho_k||_H \\ &= O_P\left(\frac{1}{\sqrt{I}h^2}\right). \end{aligned}$$

In (6.16), it follows that

$$\sup_{t\in T} |\frac{\hat{\lambda}_k \hat{\rho}_k(t)}{\lambda_k} - \rho_k(t)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right)$$

and using (6.10), we also obtain

$$\sup_{t\in T} |\hat{\rho}_k(t) - \rho_k(t)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right).$$

□

Please note that we didn't carry out the discretization error occuring in the eigenfunction estimation in Section 2.2.4. To be absolutely correct, we had to perform similar calculations as done for example in (6.2).

6.3 Consistency for spatial FPCA

Like in the one-dimensional case, we want to show consistency results also for the spatial case. Many steps of the one-dimensional proofs can be transferred to the two-dimensional case as well, but in some situations the generalization must be done carefully.

As in the one-dimensional case, we also will assume here that the measurements per observations are conducted on a regular grid. The measurement points are assumed to be fixed, equal for all observations and of the same distance. For notational convenience we further assume $T \times \mathcal{T} = [0,1]^2$. Therefore we do not use the notation introduced in section 3.2, but the points in space are named as $(\frac{n}{N}, \frac{m}{M})$ ($n = 1, \ldots, N$ and $m = 1, \ldots, M$) with observations Y^i_{nm} for $i = 1, \ldots, I$.

Again the proofs are structured that first lemmata are provided that afterwards allow to calculate rates for mean and covariance estimation.

6.3.1 Lemmata

The following Lemma is a spatial generalization of the one-dimensional Lemma 7 and will later allow to obtain a consistency result for the estimation of the mean function.

Lemma 13 (Lemma for mean estimation)**.** *Assume that*

(a) $\mathcal{K} : \mathbb{R}^2 \to \mathbb{R}$ *is an absolutely integrable two-dimensional kernel, i.e.* $\int\int |\mathcal{K}(t,\tau)|\, dt\, d\tau < \infty$*, and has an absolutely integrable Fourier transform.*

(b) $\mathcal{K}$ *is compactly supported of order* $((\nu_1, \nu_2), l)$.

(c) The bandwidth h *and the data points per observation* $N \times M$ *depend on the sample size* I *and fulfill* $M \geq N, h \to 0, Ih^{|\nu|+2} \to \infty, Ih^{2l+4} = O(1), \frac{1}{N} = O\left(h^{l+3}\right)$ *for* $I \to \infty$.

Let further $\psi : \mathbb{R}^3 \to \mathbb{R}$ *be a real function with:*

(d) ψ *is uniformly continuous on* $T \times \mathcal{T} \times \mathbb{R}$.

(e) $\sup_{t\in T, \tau\in\mathcal{T}} \int \psi^2(t,\tau,y) g(y;t,\tau)\, dy < \infty$

Now define a weighted average and its limit expected value via

$$\Psi_I(t,\tau) = \frac{1}{Ih^{|\nu|+2}NM}\sum_{i=1}^{I}\sum_{n=1}^{N}\sum_{m=1}^{M}\psi\left(\frac{n}{N},\frac{m}{M},Y_{nm}^i\right)\mathcal{K}\left(\frac{t-\frac{n}{N}}{h},\frac{\tau-\frac{m}{M}}{h}\right),$$

$$\mu(t,\tau) = \frac{\partial^{\nu_1}}{\partial^{\nu_1}t}\frac{\partial^{\nu_2}}{\partial^{\nu_2}\tau}\int\psi(t,\tau,y)g(y;t,\tau)\,dy$$

and assume that

(f) all derivatives of $\mu(t,\tau)$ up to 2nd degree exist and are uniformly continuous on $T\times\mathcal{T}$

(g) $\mu^*(t,\tau) := \int\psi(t,\tau,y)g(y;t,\tau)\,dy$ is Lipschitz continuous.

Then one can obtain the following error estimation:

$$\tau_I := \sup_{t\in T,\tau\in\mathcal{T}}|\Psi_I(t,\tau)-\mu(t,\tau)| = O_P\left(\frac{1}{\sqrt{I}h^{|\nu|+2}}\right).$$

Remark. If ψ is a function only in t and τ (i.e. Ψ_I has no random part), the error is

$$\tau_I = O\left(h^{l-|\nu|}\right).$$

Proof. First observe that

$$\mathbb{E}|\tau_I| \le \underbrace{\mathbb{E}\left(\sup_{t\in T,\tau\in\mathcal{T}}|\Psi_I(t,\tau)-\mathbb{E}(\Psi_I(t,\tau))|\right)}_{=:A} + \underbrace{\sup_{t\in T,\tau\in\mathcal{T}}|\mathbb{E}(\Psi_I(t,\tau))-\mu(t,\tau)|}_{=:B}. \tag{6.17}$$

Now parts A and B are evaluated separately.

Part A: First the kernel is represented in the Fourier space in order to find an expression which can be evaluated easily. Assumption (b) guarantees that the Fourier inversion formula can be applied. Hence we insert the expression

$$\mathcal{K}\left(\frac{\frac{n}{N}-t}{h},\frac{\frac{m}{M}-\tau}{h}\right) = \frac{1}{(2\pi)^2}\int\int e^{i\left(v\frac{\frac{n}{N}-t}{h}+w\frac{\frac{m}{M}-\tau}{h}\right)}\underbrace{\left[\int\int e^{-i(uv+u'w)}\mathcal{K}(u,u')\,du\,du'\right]}_{=:\rho(v,w)}dv\,dw$$

into $\Psi_I(t,\tau)$ and perform a substitution $v=uh, w=u'h$ afterwards:

$\Psi_I(t,\tau)$

$$= \frac{1}{Ih^{|\nu|+2}NM}\sum_{i=1}^{I}\sum_{n=1}^{N}\sum_{m=1}^{M}\psi\left(\frac{n}{N},\frac{m}{M},Y_{nm}^i\right)\frac{1}{(2\pi)^2}\int\int e^{i\left(v\frac{\frac{n}{N}-t}{h}+w\frac{\frac{m}{M}-\tau}{h}\right)}\rho(v,w)\,dv\,dw$$

$$= \frac{1}{(2\pi)^2h^{|\nu|}INM}\sum_{i=1}^{I}\sum_{n=1}^{N}\sum_{m=1}^{M}\psi\left(\frac{n}{N},\frac{m}{M},Y_{nm}^i\right)\int\int e^{iu(\frac{n}{N}-t)+iu'(\frac{m}{M}-\tau)}\rho(uh,u'h)\,du\,du'$$

$$= \frac{1}{(2\pi)^2 h^{|\nu|}} \int\int e^{-i(ut+u'\tau)} \rho(uh, u'h) \underbrace{\left[\frac{1}{INM} \sum_{i=1}^{I}\sum_{n=1}^{N}\sum_{m=1}^{M} e^{i(u\frac{n}{N}+u'\frac{m}{M})} \psi\left(\frac{n}{N}, \frac{m}{M}, Y_{nm}^{i}\right)\right]}_{=:\varphi_I(u,u')} du\, du'$$

$$= \frac{1}{(2\pi)^2 h^{|\nu|}} \int\int e^{-i(ut+u'\tau)} \rho(uh, u'h) \varphi_I(u,u')\, du\, du'$$

Therefore we obtain for the expected value of $\Psi_I(t,\tau)$

$$\mathbb{E}(\Psi_I(t,\tau)) = \frac{1}{(2\pi)^2 h^{|\nu|}} \int\int e^{-i(ut+u'\tau)} \rho(uh, u'h) \mathbb{E}(\varphi_I(u,u'))\, du\, du'$$

and for the whole part A:

$$\mathbb{E}\left(\sup_{t\in T, \tau\in\mathcal{T}} |\Psi_I(t,\tau) - \mathbb{E}(\Psi_I(t,\tau))|\right)$$

$$\leq \frac{1}{(2\pi)^2 h^{|\nu|}} \int |\rho(uh, u'h)|\, \mathbb{E}|\varphi_I(u,u') - \mathbb{E}(\varphi_I(u,u'))|\, du\, du'.$$

For the evaluation of the term $\mathbb{E}|\varphi_I(u,u') - \mathbb{E}(\varphi_I(u,u'))|$ we observe that

$$\mathbb{E}|\varphi_I(u,u') - \mathbb{E}(\varphi_I(u,u'))| \leq \sqrt{\mathbb{E}(\varphi_I(u,u') - \mathbb{E}(\varphi_I(u,u')))^2} = \sqrt{\mathrm{Var}(\varphi_I(u,u')).}$$

Using the assumption that the random variables Y_{nm}^{i} are i.i.d. (in i), we can further evaluate $\mathrm{Var}(\varphi_I(u,u'))$:

$$\mathrm{Var}(\varphi_I(u,u')) = \mathrm{Var}\left[\frac{1}{INM} \sum_{i=1}^{I}\sum_{n=1}^{N}\sum_{m=1}^{M} e^{i(u\frac{n}{N}+u'\frac{m}{M})} \psi\left(\frac{n}{N}, \frac{m}{M}, Y_{nm}^{i}\right)\right]$$

$$= \frac{1}{IN^2M^2} \mathrm{Var}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} e^{i(u\frac{n}{N}+u'\frac{m}{M})} \psi\left(\frac{n}{N}, \frac{m}{M}, Y_{nm}\right)\right]$$

$$\leq \frac{1}{IN^2M^2} \mathbb{E}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} e^{i(u\frac{n}{N}+u'\frac{m}{M})} \psi\left(\frac{n}{N}, \frac{m}{M}, Y_{nm}\right)\right]^2$$

$$\leq \frac{1}{IN^2M^2} \mathbb{E}\left[\underbrace{\left(\sum_{n=1}^{N}\sum_{m=1}^{M} \left|e^{2i(u\frac{n}{N}+u'\frac{m}{M})}\right|\right)}_{=NM} \left(\sum_{n=1}^{N}\sum_{m=1}^{M} \psi^2\left(\frac{n}{N}, \frac{m}{M}, Y_{nm}\right)\right)\right]$$

$$\leq \frac{1}{INM} \sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{E}\left(\psi^2\left(\frac{n}{N}, \frac{m}{M}, Y_{nm}\right)\right)$$

$$\leq \frac{1}{I} \max_{\substack{n=1,\ldots,N \\ m=1,\ldots,M}} \mathbb{E}\left(\psi^2\left(\frac{n}{N}, \frac{m}{M}, Y_{nm}\right)\right)$$

using the Cauchy-Schwarz inequality. All in all one obtains

$$\mathbb{E}\left(\sup_{t\in T,\tau\in\mathcal{T}} |\Psi_I(t,\tau) - \mathbb{E}(\Psi_I(t,\tau))|\right)$$

$$\leq \frac{1}{(2\pi)^2 h^{|\nu|}} \sqrt{\frac{1}{I} \max_{\substack{n=1,\ldots,N\\ m=1,\ldots,M}} \mathbb{E}\left(\psi^2\left(\frac{n}{N},\frac{m}{M},Y_{nm}\right)\right)} \underbrace{\int\int |\rho(uh,u'h)|\,du\,du'}_{=\int\int\frac{1}{h^2}|\rho(v,w)|\,dv\,dw}$$

$$= \frac{1}{h^{|\nu|+2}\sqrt{I}} \underbrace{\frac{1}{(2\pi)^2}\sqrt{\max_{\substack{n=1,\ldots,N\\ m=1,\ldots,M}} \mathbb{E}\left(\psi^2\left(\frac{n}{N},\frac{m}{M},Y_{nm}\right)\right)} \int\int |\rho(v,w)|\,dv\,dw}_{\text{bounded by assumption (e)}}$$

$$= O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)$$

for the first part.

Part B: For evaluating B we show that $\mathbb{E}(\Psi_I(t,\tau)) = \mu(t,\tau) + O(h^{l-|\nu|}) + O\left(\frac{1}{h^{|\nu|+3}N}\right)$ uniformly for $t \in T$ and $\tau \in \mathcal{T}$. In order to obtain this error estimation the sum expression is approximated by an integral with an error term of order $O\left(\frac{1}{h^{|\nu|+3}N}\right)$ (for calculation see (6.18) below). After a substitution we can use the Taylor expansion to evaluate the term. Set $\mu^*(t,\tau) := \int \psi\left(t,\tau,y\right) g\left(y;t,\tau\right)\,dy$ for notational convenience.

$$\mathbb{E}(\Psi_I(t,\tau))$$

$$= \frac{1}{h^{|\nu|+2}NM}\mathbb{E}\left(\sum_{n=1}^{N}\sum_{m=1}^{M}\psi\left(\frac{n}{N},\frac{m}{M},Y_{nm}\right)\mathcal{K}\left(\frac{\frac{n}{N}-t}{h},\frac{\frac{m}{M}-\tau}{h}\right)\right)$$

$$= \frac{1}{h^{|\nu|+2}NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\int\psi\left(\frac{n}{N},\frac{m}{M},y_1,y_2\right)g\left(y_1,y_2;\frac{n}{N},\frac{m}{M}\right)\mathcal{K}\left(\frac{\frac{n}{N}-t}{h},\frac{\frac{m}{M}-\tau}{h}\right)\,dy$$

$$= \frac{1}{h^{|\nu|+2}NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\mu^*\left(\frac{n}{N},\frac{m}{M},y_1,y_2\right)\mathcal{K}\left(\frac{\frac{n}{N}-t}{h},\frac{\frac{m}{M}-\tau}{h}\right)$$

$$= \frac{1}{h^{|\nu|+2}}\int\int\mu^*\left(t',\tau'\right)\mathcal{K}\left(\frac{t'-t}{h},\frac{\tau'-\tau}{h}\right)\,dt'\,d\tau' + O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

$$= \frac{1}{h^{|\nu|}}\int\int\mu^*\left(t+vh,\tau+wh\right)\mathcal{K}(v,w)\,dv\,dw + O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

$$= \frac{1}{h^{|\nu|}}\int\int\left[\sum_{j=0}^{l-1}\sum_{\substack{j_1,j_2\\ j_1+j_2=j}}\frac{\partial^{j_1}}{\partial^{j_1}t}\frac{\partial^{j_2}}{\partial^{j_2}\tau}\mu^*(t,\tau)\frac{h^j v^{j_1}w^{j_2}}{j_1!j_2!} + \sum_{\substack{j_1,j_2\\ j_1+j_2=l}}\frac{\partial^{j_1}}{\partial^{j_1}t}\frac{\partial^{j_2}}{\partial^{j_2}\tau}\mu^*(\xi_1,\xi_2)\frac{h^l v^{j_1}w^{j_2}}{j_1!j_2!}\right]$$

$$\times\,\mathcal{K}(v,w)\,dv\,dw + O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

$$= \frac{1}{h^{|\nu|}} \left[\frac{\partial^{\nu_1}}{\partial^{\nu_1} t}\frac{\partial^{\nu_2}}{\partial^{\nu_2}\tau}\mu^*(t,\tau)\frac{h^{|\nu|}}{|\nu|!}\underbrace{\int\int v^{\nu_1}w^{\nu_2}\mathcal{K}(v,w)\,dv\,dw}_{(-1)^{|\nu|}|\nu|!} \right.$$

$$\left. + \sum_{\substack{j_1,j_2\\ j_1+j_2=l}} \underbrace{\frac{\partial^{j_1}}{\partial^{j_1}t}\frac{\partial^{j_2}}{\partial^{j_2}\tau}\mu^*(\xi_1,\xi_2)}_{\text{bounded}} \frac{h^l}{j_1!j_2!}\underbrace{\int\int v^{j_1}w^{j_2}\mathcal{K}(v,w)\,dv\,dw}_{\text{at least one}\neq 0}\right] + O\left(\frac{1}{h^{|\nu|+3}N}\right)$$

$$= \mu(t,\tau) + \underbrace{O(h^{l-|\nu|}) + O\left(\frac{1}{h^{|\nu|+3}N}\right)}_{\text{independent of } t\in T,\tau\in\mathcal{T}}$$

The error of discretization, e.g. the error that occurs when substituting a sum expression by an integral as in the last calculation, is evaluated in the following. Therefore the integral is first inserted artificially in the sum expression and afterwards the term is divided in two parts E_1 and E_2, where conditions of the kernel function and of μ^* can be used.

$$\frac{1}{h^{|\nu|+2}NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\mu^*\left(\frac{n}{N},\frac{m}{M}\right)\mathcal{K}\left(\frac{\frac{n}{N}-t}{h},\frac{\frac{m}{M}-\tau}{h}\right) \tag{6.18}$$

$$\begin{aligned}
=&\frac{1}{h^{|\nu|+2}}\int\int\mu^*(t',\tau')\mathcal{K}\left(\frac{t'-t}{h},\frac{\tau'-\tau}{h}\right)d\tau'\,dt'\\
&+\frac{1}{h^{|\nu|+2}}\sum_{n=1}^{N}\sum_{m=1}^{M}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\int_{\frac{m-1}{M}}^{\frac{m}{M}}\mu^*\left(\frac{n}{N},\frac{m}{M}\right)\mathcal{K}\left(\frac{\frac{n}{N}-t}{h},\frac{\frac{m}{M}-\tau}{h}\right)d\tau'\,dt'\\
&-\frac{1}{h^{|\nu|+2}}\sum_{n=1}^{N}\sum_{m=1}^{M}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\int_{\frac{m-1}{M}}^{\frac{m}{M}}\mu^*(t',\tau')\mathcal{K}\left(\frac{t'-t}{h},\frac{\tau'-\tau}{h}\right)d\tau'\,dt'\\
=&\frac{1}{h^{|\nu|+2}}\int\int\mu^*(t',\tau')\mathcal{K}\left(\frac{t'-t}{h},\frac{\tau'-\tau}{h}\right)d\tau'\,dt'\\
&+\underbrace{\frac{1}{h^{|\nu|+2}}\sum_{n=1}^{N}\sum_{m=1}^{M}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\int_{\frac{m-1}{M}}^{\frac{m}{M}}\left[\mu^*\left(\frac{n}{N},\frac{m}{M}\right)-\mu^*(t',\tau)\right]\mathcal{K}\left(\frac{\frac{n}{N}-t}{h},\frac{\frac{m}{M}-\tau}{h}\right)d\tau'\,dt'}_{=:E_1}\\
&+\underbrace{\frac{1}{h^{|\nu|+2}}\sum_{n=1}^{N}\sum_{m=1}^{M}\int_{\frac{n-1}{N}}^{\frac{n}{N}}\int_{\frac{m-1}{M}}^{\frac{m}{M}}\left[\mathcal{K}\left(\frac{\frac{n}{N}-t}{h},\frac{\frac{m}{M}-\tau}{h}\right)-\mathcal{K}\left(\frac{t'-t}{h},\frac{\tau'-\tau}{h}\right)\right]\mu^*(t',\tau')\,d\tau'\,dt'}_{=:E_2}
\end{aligned} \tag{6.19}$$

Evaluation of E_1 (based on the condition that μ^* is Lipschitz continuous):

$$E_1 = \frac{1}{h^{|\nu|+2}} \sum_{n=1}^{N} \sum_{m=1}^{M} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \int_{\frac{m-1}{M}}^{\frac{m}{M}} \left[\mu^*\left(\frac{n}{N}, \frac{m}{M}\right) - \mu^*(t', \tau')\right] \mathcal{K}\left(\frac{\frac{n}{N} - t}{h}, \frac{\frac{m}{M} - \tau}{h}\right) d\tau'\, dt$$

$$= O\left(\frac{1}{h^{|\nu|+2} N}\right)$$

because $\left|\mu^*\left(\frac{n}{N}, \frac{m}{M}\right) - \mu^*(t', \tau')\right| \leq c||(\frac{n}{N}, \frac{m}{M}) - (t', \tau')|| \leq \frac{c}{N}$ (assumption (c)).

For evaluation of E_2 a multi-dimensional version of the mean value theorem is used (see e.g. Heuser [2002]) with $\nabla\mathcal{K}$ being the gradient of $\mathcal{K}$:

$$E_2 \leq \frac{1}{h^{|\nu|+2}} \sum_{n=1}^{N} \sum_{m=1}^{M} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \int_{\frac{m-1}{M}}^{\frac{m}{M}} \left|\left[\mathcal{K}\left(\frac{\frac{n}{N} - t}{h}, \frac{\frac{m}{M} - \tau}{h}\right) - \mathcal{K}\left(\frac{t' - t}{h}, \frac{\tau' - \tau}{h}\right)\right] \mu^*(t', \tau')\right| d\tau'\, dt'$$

$$\leq \frac{1}{h^{|\nu|+2}} \sum_{n=1}^{N} \sum_{m=1}^{M} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \int_{\frac{m-1}{M}}^{\frac{m}{M}} \left|\nabla\mathcal{K}(\xi_{nm}) \left(\frac{\frac{n}{N} - t'}{h}, \frac{\frac{m}{M} - \tau'}{h}\right)^T \mu^*(t', \tau')\right| d\tau'\, dt'$$

$$\text{for} \quad \xi_{nm} \in \left\{\left(\frac{\frac{n}{N} - t}{h}, \frac{\frac{m}{M} - \tau}{h}\right) + \eta\left(\frac{\frac{n}{N} - t'}{h}, \frac{\frac{m}{M} - \tau'}{h}\right), \eta \in [0, 1]\right\}$$

$$\leq \frac{1}{h^{|\nu|+2}} \sum_{n=1}^{N} \sum_{m=1}^{M} \max_{\substack{n=1,\ldots,N \\ m=1,\ldots,M}} ||\mathcal{K}'(\xi_{mn})|| \frac{2}{hN} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \int_{\frac{m-1}{M}}^{\frac{m}{M}} |\mu^*(t', \tau')|\, d\tau'\, dt'$$

$$= O\left(\frac{1}{h^{|\nu|+3} N}\right)$$

Now it is shown that:

Part A: $\mathbb{E}\left(\sup_{t \in T, \tau \in \mathcal{T}} |\Psi_I(t, \tau) - \mathbb{E}(\Psi_I(t, \tau))|\right) = O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)$

Part B: $\mathbb{E}(\Psi_I(t, \tau)) = \mu(t, \tau) + O(h^{l-|\nu|}) + O(\frac{1}{h^{|\nu|+3}N})$

For (6.17) we therefore obtain the inequality:

$$\mathbb{E}|\tau_I| = O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right) + O(h^{l-|\nu|}) + O\left(\frac{1}{h^{|\nu|+3}N}\right) = O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right),$$

because $h^{l-|\nu|} = \underbrace{\sqrt{Ih^{2l+4}}}_{\text{bounded}} \frac{1}{h^{|\nu|+2}\sqrt{I}} = O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)$ and $\frac{1}{N} = O(h^{l+3})$ (assumption (c)).

Using Markov's inequality one obtains:

$$\mathbb{E}|\tau_I| = O\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right) \Rightarrow \tau_I = O_P\left(\frac{1}{h^{|\nu|+2}\sqrt{I}}\right)$$

As before, if $\Psi_I(t, \tau)$ is a non-random function, Part A vanishes such that the final rate is only the better rate of Part B. □

The previous lemma also needs to be adapted to the situation of the estimation of the covariance function in the spatial case:

Lemma 14 (Lemma for covariance estimation)**.** *Now let $\vec{t} = (t_1, t_2)$ be an element of T^2 and $\vec{\tau} = (\tau_1, \tau_2)$ be an element of $\mathcal{T}^2$. Assume that*

(a) $\mathcal{K}$ is an absolutely integrable kernel, i.e. $\int\int |\mathcal{K}(\vec{t}, \vec{\tau})| \, d\vec{t} d\vec{\tau} < \infty$, and has an absolutely integrable Fourier transform.

(b) $\mathcal{K}$ is compactly supported of order (ν, l) with $\nu \in \mathbb{N}^4$ and $\int\int \mathcal{K}^2(\vec{t}, \vec{\tau}) \, d\vec{t} d\vec{\tau} < \infty$.

(c) The bandwidth h and the data points per observation N, M depend on the sample size I and fulfill $M \geq N$, $h \to 0, Ih^{|\nu|+4} \to \infty, Ih^{2l+8} = O(1), \frac{1}{N} = O(h^{l+5})$ for $I \to \infty$.

Let further $\theta : \mathbb{R}^6 \to \mathbb{R}$ be a real function with:

(d) θ is uniformly continuous on $T^2 \times \mathcal{T}^2 \times \mathbb{R}^2$.

(e) $\sup_{\substack{\vec{t}\in T^2 \\ \vec{\tau}\in\mathcal{T}^2}} \int \theta^2(\vec{t}, \vec{\tau}, y) g(y; \vec{t}, \vec{\tau}) \, dy < \infty$

Now define a weighted average and its limit expected value via

$$\Theta_I(\vec{t}, \vec{\tau})$$

$$= \frac{1}{Ih^{|\nu|+4} N(N-1)M(M-1)} \sum_{i=1}^{I} \sum_{n_1 \neq n_2} \sum_{m_1 \neq m_2} \theta\left(\frac{\vec{n}}{N}, \frac{\vec{m}}{M}, Y_1^i, Y_2^i\right) \mathcal{K}\left(\frac{\vec{t} - \frac{\vec{n}}{N}}{h}, \frac{\vec{\tau} - \frac{\vec{m}}{M}}{h}\right),$$

$$\gamma(\vec{t}, \vec{\tau}) = \frac{d^{|\nu|}}{d(\vec{t}, \vec{\tau})^\nu} \int\int \theta(\vec{t}, \vec{\tau}, y_1, y_2) g(y_1, y_2; \vec{t}, \vec{\tau}) \, dy_1 \, dy_2$$

and assume that

(f) all derivatives of $\gamma(\vec{t}, \vec{\tau})$ up to 2nd degree exist and are uniformly continuous on $T^2 \times \mathcal{T}^2$

(g) $\gamma(\vec{t}, \vec{\tau}) := \int\int \theta(\vec{t}, \vec{\tau}, y_1, y_2) g(y_1, y_2; \vec{t}, \vec{\tau}) \, dy_1 \, dy_2$ is Lipschitz continuous.

Then one can obtain the following error estimation:

$$\tau_I := \sup_{\substack{\vec{t}\in T^2 \\ \vec{\tau}\in\mathcal{T}^2}} |\Theta_I(\vec{t}, \vec{\tau}) - \gamma(\vec{t}, \vec{\tau})| = O_P\left(\frac{1}{\sqrt{I} h^{|\nu|+4}}\right).$$

Remark. If θ is a function only in t and τ (i.e. Θ_I has no random part), the error is

$$\tau_I = O\left(h^{l-|\nu|}\right).$$

Proof. First observe that

$$\mathbb{E}|\tau_I| \leq \underbrace{\mathbb{E} \sup_{\substack{\vec{t}\in T^2 \\ \vec{\tau}\in\mathcal{T}^2}} |\Theta_I(\vec{t}, \vec{\tau}) - \mathbb{E}(\Theta_I(\vec{t}, \vec{\tau}))|}_{=:A} + \underbrace{\sup_{\substack{\vec{t}\in T^2 \\ \vec{\tau}\in\mathcal{T}^2}} |\mathbb{E}(\Theta_I(\vec{t}, \vec{\tau})) - \gamma(\vec{t}, \vec{\tau})|}_{=:B}. \tag{6.20}$$

Now parts A and B are evaluated separately.

Part A: First the kernel is represented in the Fourier space in order to find an expression which can be evaluated easily. Assumption (b) guarantees that the Fourier inversion formula can be applied. Hence we insert the expression

$$\begin{aligned}
&\mathcal{K}\left(\frac{\vec{t}-\frac{\vec{n}}{N}}{h},\frac{\vec{\tau}-\frac{\vec{m}}{M}}{h}\right)\\
&=\frac{1}{(2\pi)^4}\int\int e^{i\left(\vec{v}^T\frac{\frac{\vec{n}}{N}-\vec{t}}{h}+\vec{w}^T\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)}\underbrace{\left[\int\int e^{-i(\vec{u}^T\vec{v}+\vec{u}'^T\vec{w})}\mathcal{K}(\vec{u}^T,\vec{u}'^T)\,d\vec{u}\,d\vec{u}'\right]}_{=:\rho(\vec{v},\vec{w})}d\vec{v}\,d\vec{w}
\end{aligned}$$

into $\Theta_I(\vec{t},\vec{\tau})$ and perform a substitution $\vec{v}=\vec{u}h,\vec{w}=\vec{u}'h$ afterwards. Let R:=N(N-1)M(M-1) and $Y_l:=Y_{n_l,m_l}$ for $l=1,2$.

$$\begin{aligned}
&\Theta_I(\vec{t},\vec{\tau})\\
&=\frac{1}{Ih^{|\nu|+2}R}\sum_{i=1}^{I}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\theta\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1^i,Y_2^i\right)\frac{1}{(2\pi)^2}\int\int e^{i\left(\vec{v}^T\frac{\frac{\vec{n}}{N}-\vec{t}}{h}+\vec{w}^T\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)}\rho(\vec{v},\vec{w})\,d\vec{v}\,d\vec{w}\\
&=\frac{1}{(2\pi)^4h^{|\nu|}IR}\sum_{i=1}^{I}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\theta\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1^i,Y_2^i\right)\int\int e^{i\vec{u}^T(\frac{\vec{n}}{N}-\vec{t})+i\vec{u}'^T(\frac{\vec{m}}{M}-\vec{\tau})}\rho(\vec{u}h,\vec{u}'h)\,d\vec{u}\,d\vec{u}'\\
&=\frac{1}{(2\pi)^4h^{|\nu|}}\int\int e^{-i(\vec{u}^T\vec{t}+\vec{u}'^T\vec{\tau})}\rho(\vec{u}h,\vec{u}'h)\\
&\qquad\times\underbrace{\left[\frac{1}{IR}\sum_{i=1}^{I}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}e^{i(\vec{u}^T\frac{\vec{n}}{N}+\vec{u}'^T\frac{\vec{m}}{M})}\theta\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1^i,Y_2^i\right)\right]}_{=:\varphi_I(\vec{u},\vec{u}')}d\vec{u}\,d\vec{u}'\\
&=\frac{1}{(2\pi)^4h^{|\nu|}}\int\int e^{-i(\vec{u}^T\vec{t}+\vec{u}'^T\vec{\tau})}\rho(\vec{u}h,\vec{u}'h)\varphi_I(\vec{u},\vec{u}')\,d\vec{u}\,d\vec{u}'
\end{aligned}$$

Therewith

$$\mathbb{E}(\Theta_I(\vec{t},\vec{\tau}))=\frac{1}{(2\pi)^4h^{|\nu|}}\int\int e^{-i(\vec{u}^T\vec{t}+\vec{u}'\vec{\tau})}\rho(\vec{u}h,\vec{u}'h)\mathbb{E}(\varphi_I(\vec{u},\vec{u}'))\,d\vec{u}\,d\vec{u}'$$

and

$$\mathbb{E}\left(\sup_{\substack{\vec{t}\in\mathcal{T}^2\\ \vec{\tau}\in\mathcal{T}^2}}|\Theta_I(\vec{t},\vec{\tau})-\mathbb{E}(\Theta_I(\vec{t},\vec{\tau}))|\right)\leq\frac{1}{(2\pi)^4h^{|\nu|}}\int|\rho(\vec{u}h,\vec{u}'h)|\,\mathbb{E}|\varphi_I(\vec{u},\vec{u}')-\mathbb{E}(\varphi_I(\vec{u},\vec{u}'))|\,d\vec{u}\,d\vec{u}'.$$

For the evaluation of the term $\mathbb{E}|\varphi_I(\vec{u},\vec{u}')-\mathbb{E}(\varphi_I(\vec{u},\vec{u}'))|$ we observe that

$$\mathbb{E}|\varphi_I(\vec{u},\vec{u}')-\mathbb{E}(\varphi_I(\vec{u},\vec{u}'))|\leq\sqrt{\mathbb{E}(\varphi_I(\vec{u},\vec{u}')-\mathbb{E}(\varphi_I(\vec{u},\vec{u}')))^2}=\sqrt{\mathrm{Var}(\varphi_I(\vec{u},\vec{u}'))}.$$

The variance is again evaluated by using the condition of i.i.d. observations and the Cauchy-Schwarz inequality:

$$
\begin{aligned}
&\mathrm{Var}(\varphi_I(\vec{u},\vec{u}'))\\
=&\mathrm{Var}\left[\frac{1}{IR}\sum_{i=1}^{I}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2} e^{i(\vec{u}^T\frac{\vec{n}}{N}+\vec{u}'^T\frac{\vec{m}}{M})}\theta\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1^i,Y_2^i\right)\right]\\
=&\frac{1}{IR^2}\mathrm{Var}\left[\sum_{n_1\neq n_2}\sum_{m_1\neq m_2} e^{i(\vec{u}^T\frac{\vec{n}}{N}+\vec{u}'^T\frac{\vec{m}}{M})}\theta\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1,Y_2\right)\right]\\
\leq&\frac{1}{IR^2}\mathbb{E}\left[\sum_{n_1\neq n_2}\sum_{m_1\neq m_2} e^{i(\vec{u}^T\frac{\vec{n}}{N}+\vec{u}'^T\frac{\vec{m}}{M})}\theta\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1,Y_2\right)\right]^2\\
\leq&\frac{1}{IR^2}\mathbb{E}\Bigg[\underbrace{\left(\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\left|e^{2i(\vec{u}^T\frac{\vec{n}}{N}+\vec{u}'^T\frac{\vec{m}}{M})}\right|\right)}_{=R}\left(\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\theta^2\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1,Y_2\right)\right)\Bigg]\\
\leq&\frac{1}{IR}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\mathbb{E}\left(\theta^2\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1,Y_2\right)\right)\\
\leq&\frac{1}{I}\max_{\substack{n_1\neq n_2\\ m_1\neq m_2}}\mathbb{E}\left(\theta^2\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1,Y_2\right)\right)
\end{aligned}
$$

using the Cauchy-Schwarz inequality. All in all one obtains

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}}|\Theta_I(\vec{t},\vec{\tau})-\mathbb{E}(\Theta_I(\vec{t},\vec{\tau}))|\right)\\
&\leq\frac{1}{(2\pi)^4h^{|\nu|}}\sqrt{\frac{1}{I}\max_{\substack{n_1\neq n_2\\ m_1\neq m_2}}\mathbb{E}\left(\theta^2\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1,Y_2\right)\right)}\underbrace{\int\int|\rho(\vec{u}h,\vec{u}'h)|\,d\vec{u}\,d\vec{u}'}_{=\int\int\frac{1}{h^4}|\rho(\vec{v},\vec{w})|\,d\vec{v}\,d\vec{w}}\\
&=\frac{1}{h^{|\nu|+4}\sqrt{I}}\underbrace{\frac{1}{(2\pi)^4}\sqrt{\max_{\substack{n_1\neq n_2\\ m_1\neq m_2}}\mathbb{E}\left(\theta^2\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1,Y_2\right)\right)}\int\int|\rho(\vec{v},\vec{w})|\,d\vec{v}\,d\vec{w}}_{\text{bounded by assumption (e)}}\\
&=O\left(\frac{1}{h^{|\nu|+4}\sqrt{I}}\right)
\end{aligned}
$$

for the first part.

Part B: For evaluating B we show that $\mathbb{E}(\Theta_I(\vec{t},\vec{\tau}))=\gamma(\vec{t},\vec{\tau})+O(h^{l-|\nu|})+O\left(\frac{1}{h^{|\nu|+5}N}\right)$ uniformly for $(\vec{t},\vec{\tau})\in T^2\times\mathcal{T}^2$. In order to obtain this error estimation the sum expression is substituted by an integral with an error term of order $O\left(\frac{1}{h^{|\nu|+4}N}\right)$ (for calculation see (6.21) below) and is examined via a Taylor expansion after a substitution in each dimension. Set $\gamma^*(\vec{t},\vec{\tau}) := \int\int \theta\left(\vec{t},\vec{\tau},y_1,y_2\right) g\left(y_1,y_2;\vec{t},\vec{\tau}\right) dy_1\, dy_2$ for notational convenience.

$$\begin{aligned}
\mathbb{E}(\Theta_I(\vec{t},\vec{\tau})) &= \frac{1}{h^{|\nu|+4}R}\mathbb{E}\left(\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\theta\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},Y_1,Y_2\right)\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)\right)\\
&= \frac{1}{h^{|\nu|+4}R}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\int\theta\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M},y\right)g\left(y;\frac{\vec{n}}{N},\frac{\vec{m}}{M}\right)\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)dy\\
&= \frac{1}{h^{|\nu|+4}R}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\gamma^*\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M}\right)\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)\\
&= \frac{1}{h^{|\nu|+4}}\int\int\gamma^*\left(\vec{t}',\vec{\tau}'\right)\mathcal{K}\left(\frac{\vec{t}'-\vec{t}}{h},\frac{\vec{\tau}'-\vec{\tau}}{h}\right)dt'\,d\tau'+O\left(\frac{1}{h^{|\nu|+4}N}\right)\\
&= \frac{1}{h^{|\nu|}}\int\int\gamma^*\left(\vec{t}+\vec{v}h,\vec{\tau}+\vec{w}h\right)\mathcal{K}(\vec{v},\vec{w})\,d\vec{v}\,d\vec{w}+O\left(\frac{1}{h^{|\nu|+4}N}\right)\\
&= \frac{1}{h^{|\nu|}}\int\int\left[\sum_{j=0}^{l-1}\sum_{\substack{\vec{j_1},\vec{j_2}\\|\vec{j_1}|+|\vec{j_2}|=j}}\frac{\partial^{\vec{j_1}}}{\partial^{\vec{j_1}}\vec{t}}\frac{\partial^{\vec{j_2}}}{\partial^{\vec{j_2}}\vec{\tau}}\gamma^*(\vec{t},\vec{\tau})\frac{h^j\vec{v}^{\vec{j_1}}\vec{w}^{\vec{j_2}}}{\vec{j_1}!\vec{j_2}!}\right.\\
&\qquad\left.+\sum_{\substack{\vec{j_1},\vec{j_2}\\|\vec{j_1}|+|\vec{j_2}|=l}}\frac{\partial^{\vec{j_1}}}{\partial^{\vec{j_1}}\vec{t}}\frac{\partial^{\vec{j_2}}}{\partial^{\vec{j_2}}\vec{\tau}}\gamma^*(\xi_1,\xi_2)\frac{h^l\vec{v}^{\vec{j_1}}\vec{w}^{\vec{j_2}}}{\vec{j_1}!\vec{j_2}!}\right]\times\mathcal{K}(\vec{v},\vec{w})\,d\vec{v}\,d\vec{w}+O\left(\frac{1}{h^{|\nu|+4}N}\right)\\
&= \gamma(\vec{t},\vec{\tau})+\underbrace{O(h^{l-|\nu|})+O\left(\frac{1}{h^{|\nu|+4}N}\right)}_{\text{independent of }\vec{t}\in T^2,\vec{\tau}\in\mathcal{T}^2}
\end{aligned}$$

The error of discretization is evaluated like in the lemma before, e.g. the integral is first inserted into the sum expression and afterwards the term is divided into two parts E_1 and E_2, that are evaluated using the properties of the kernel function and of γ^*:

$$\begin{aligned}
&\frac{1}{h^{|\nu|+4}R}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\gamma^*\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M}\right)\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right) \qquad (6.21)\\
=&\frac{1}{h^{|\nu|+4}}\int\int\gamma^*\left(\vec{t}',\vec{\tau}'\right)\mathcal{K}\left(\frac{\vec{t}'-\vec{t}}{h},\frac{\vec{\tau}'-\vec{\tau}}{h}\right)d\vec{\tau}'\,d\vec{t}'\\
&+\frac{1}{h^{|\nu|+4}}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\int_{\frac{\vec{n}-1}{N}}^{\frac{\vec{n}}{N}}\int_{\frac{\vec{m}-1}{M}}^{\frac{\vec{m}}{M}}\gamma^*\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M}\right)\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)d\vec{\tau}'\,d\vec{t}'
\end{aligned}$$

$$-\frac{1}{h^{|\nu|+4}}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\int_{\frac{\vec{n}-1}{N}}^{\frac{\vec{n}}{N}}\int_{\frac{\vec{m}-1}{M}}^{\frac{\vec{m}}{M}}\gamma^*\left(\vec{t}',\vec{\tau}'\right)\mathcal{K}\left(\frac{\vec{t}'-\vec{t}}{h},\frac{\vec{\tau}'-\vec{\tau}}{h}\right)d\vec{\tau}'\,d\vec{t}'$$

$$=\frac{1}{h^{|\nu|+4}}\int\int\gamma^*\left(\vec{t}',\vec{\tau}'\right)\mathcal{K}\left(\frac{\vec{t}'-\vec{t}}{h},\frac{\vec{\tau}'-\vec{\tau}}{h}\right)d\vec{\tau}'\,d\vec{t}'$$

$$\underbrace{+\frac{1}{h^{|\nu|+4}}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\int_{\frac{\vec{n}-1}{N}}^{\frac{\vec{n}}{N}}\int_{\frac{\vec{m}-1}{M}}^{\frac{\vec{m}}{M}}\left[\gamma^*\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M}\right)-\gamma^*\left(\vec{t},\vec{\tau}\right)\right]\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)d\vec{\tau}'\,d\vec{t}'}_{=:E_1}$$

$$\underbrace{+\frac{1}{h^{|\nu|+4}}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\int_{\frac{\vec{n}-1}{N}}^{\frac{\vec{n}}{N}}\int_{\frac{\vec{m}-1}{M}}^{\frac{\vec{m}}{M}}\left[\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)-\mathcal{K}\left(\frac{\vec{t}'-\vec{t}}{h},\frac{\vec{\tau}'-\vec{\tau}}{h}\right)\right]\gamma^*\left(\vec{t}',\vec{\tau}'\right)d\vec{\tau}'\,d\vec{t}'}_{=:E_2}$$

E_1 is evaluated based on the condition that γ^* :

$$E_1=\frac{1}{h^{|\nu|+4}}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\int_{\frac{\vec{n}-1}{N}}^{\frac{\vec{n}}{N}}\int_{\frac{\vec{m}-1}{M}}^{\frac{\vec{m}}{M}}\left[\gamma^*\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M}\right)-\gamma^*(\vec{t}',\vec{\tau}')\right]\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)d\vec{\tau}'\,d\vec{t}'$$

$$=O\left(\frac{1}{h^{|\nu|+4}N}\right)$$

because $\left|\gamma^*\left(\frac{\vec{n}}{N},\frac{\vec{m}}{M}\right)-\gamma^*(\vec{t}',\vec{\tau}')\right|\leq c||(\frac{\vec{n}}{N},\frac{\vec{m}}{M})-(\vec{t}',\vec{\tau}')||\leq\frac{c}{N}$ (assumption (c)).

For the evaluation of E_2 the mean value theorem in several variables is used (with $\nabla\mathcal{K}$ being the gradient of $\mathcal{K}$:

$$E_2$$

$$\leq\frac{1}{h^{|\nu|+4}}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\int_{\frac{\vec{n}-1}{N}}^{\frac{\vec{n}}{N}}\int_{\frac{\vec{m}-1}{M}}^{\frac{\vec{m}}{M}}\left|\left[\mathcal{K}\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)-\mathcal{K}\left(\frac{\vec{t}'-\vec{t}}{h},\frac{\vec{\tau}'-\vec{\tau}}{h}\right)\right]\gamma^*(\vec{t}',\vec{\tau}')\right|d\vec{\tau}'\,d\vec{t}'$$

$$\leq\frac{1}{h^{|\nu|+4}}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\int_{\frac{\vec{n}-1}{N}}^{\frac{\vec{n}}{N}}\int_{\frac{\vec{m}-1}{M}}^{\frac{\vec{m}}{M}}\left|\mathcal{K}'(\xi_{nm})\left(\frac{\frac{\vec{n}}{N}-\vec{t}'}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}'}{h}\right)^T\gamma^*(\vec{t}',\vec{\tau}')\right|d\vec{\tau}'\,d\vec{t}'$$

$$\text{for}\quad\xi_{nm}\in\left\{\left(\frac{\frac{\vec{n}}{N}-\vec{t}}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}}{h}\right)+\eta\left(\frac{\frac{\vec{n}}{N}-\vec{t}'}{h},\frac{\frac{\vec{m}}{M}-\vec{\tau}'}{h}\right),\eta\in[0,1]\right\}$$

$$\leq\frac{1}{h^{|\nu|+4}}\sum_{n_1\neq n_2}\sum_{m_1\neq m_2}\max_{\substack{n_1\neq n_2\\ m_1\neq m_2}}||\nabla\mathcal{K}(\xi_{mn})||\,|\frac{4}{hN}\int_{\frac{\vec{n}-1}{N}}^{\frac{\vec{n}}{N}}\int_{\frac{\vec{m}-1}{M}}^{\frac{\vec{m}}{M}}\left|\gamma^*(\vec{t}',\vec{\tau}')\right|d\vec{\tau}'\,d\vec{t}'$$

$$=O\left(\frac{1}{h^{|\nu|+5}N}\right)$$

Now it is shown that:

Part A: $\mathbb{E}\left(\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}} |\Theta_I(\vec{t},\vec{\tau}) - \mathbb{E}(\Theta_I(\vec{t},\vec{\tau}))|\right) = O\left(\frac{1}{h^{|\nu|+4}\sqrt{I}}\right)$

Part B: $\mathbb{E}(\Theta_I(\vec{t},\vec{\tau})) = \gamma(\vec{t},\vec{\tau}) + O(h^{l-|\nu|}) + O(\frac{1}{h^{|\nu|+5}N})$

For (6.20) we therefore obtain the inequality:

$$\mathbb{E}|\tau_I| = O\left(\frac{1}{h^{|\nu|+4}\sqrt{I}}\right) + O(h^{l-|\nu|}) + O\left(\frac{1}{h^{|\nu|+5}N}\right) = O\left(\frac{1}{h^{|\nu|+4}\sqrt{I}}\right),$$

because $h^{l-|\nu|} = O\left(\frac{1}{h^{|\nu|+4}\sqrt{I}}\right)$ and $\frac{1}{N} = O(h^{l+5})$ (assumption (c)).
Using Markov's inequality one obtains:

$$\mathbb{E}|\tau_I| = O\left(\frac{1}{h^{|\nu|+4}\sqrt{I}}\right) \Rightarrow \tau_I = O_P\left(\frac{1}{h^{|\nu|+4}\sqrt{I}}\right)$$

As before, if $\Theta_I(\vec{t},\vec{\tau})$ is a non-random function, Part A vanishes such that the final rate is only the better rate of Part B. □

6.3.2 Mean, covariance and principal components

Using the results of Lemmas 13 and 14 we can now provide consistency results for mean and covariance functions.

Theorem 15 (Theorem for mean estimation)**.** *Assume that (a), (b) and (d) of Lemma 13 are valid here as well and that*

(c) $h \to 0, Ih^4 \to \infty, Ih^8 = O(1), \frac{1}{N} = O(h^7), M \geq N$ *for* $I \to \infty$*.*

Further $\mathcal{K}$ *is assumed to be a two-dimensional product kernel of a kernel* $\mathcal{K}^*$ *of order* $(0,2)$*, e.g.* $\mathcal{K}(t,\tau) = \mathcal{K}^*(t)\mathcal{K}^*(\tau)$ *for* $t \in T, \tau \in \mathcal{T}$*.* $\mathcal{K}$ *is therefore of order* $((0,0),2)$ *(see calculation* (2.7)*). The mean function fulfills the following uniform convergence rate:*

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} |\hat{\mu}(t,\tau) - \mu(t,\tau)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right)$$

Proof. The local linear estimator $\hat{\mu}(t,\tau)$ can explicitly be written as

$$\begin{aligned}\hat{\mu}(t,\tau) = \hat{\beta}_0(t,\tau) = &\frac{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}Y^i_{nm}}{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\\ &- \frac{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)}{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\hat{\beta}_{11}(t,\tau)\\ &- \frac{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{m}{M}-\tau)}{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\hat{\beta}_{12}(t,\tau),\end{aligned}$$

where $w_{nm} = \mathcal{K}\left(\frac{t-\frac{n}{N}}{h}, \frac{\tau-\frac{m}{M}}{h}\right)$, and

$$\hat{\beta}_{11}(t,\tau) = D_1 + E_1\hat{\beta}_{12}(t,\tau)$$

$$\hat{\beta}_{12}(t,\tau) = D_2 + E_2\hat{\beta}_{11}(t,\tau)$$

with

$$D_1 = \Bigg[\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)Y^i_{nm} - \frac{\left(\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)\right)\left(\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}Y^i_{nm}\right)}{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\Bigg]$$
$$\times\left[\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)^2 - \frac{\left(\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)\right)^2}{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\right]^{-1}$$

$$E_1 = \Bigg[\left(\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)\right)\left(\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{m}{M}-\tau)\right) - \frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)(\frac{m}{M}-\tau)\Bigg]$$
$$\times\left[\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)^2 - \frac{\left(\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)\right)^2}{\frac{1}{I}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\right]^{-1}$$

and D_2 and E_2 analogous exchanging t with τ and n with m.

We can rewrite

$$\hat{\beta}_{11}(t,\tau) = \frac{D_1 + E_1 D_2}{1 - E_1 E_2}$$

$$\hat{\beta}_{12}(t,\tau) = \frac{D_2 + E_2 D_1}{1 - E_1 E_2}$$

so that we first evaluate the terms D_1, D_2, E_1 and E_2 and afterwards conclude the convergence rates of $\hat{\beta}_{11}(t,\tau)$ and $\hat{\beta}_{12}(t,\tau)$.

In order to evaluate the terms we define the derivative kernels $\mathcal{K}_t(t,\tau) = \frac{-t\mathcal{K}(t,\tau)}{\sigma_t^2}$ with order $((1,0),3)$, $\mathcal{K}_{t^2}(t,\tau) = \frac{t^2\mathcal{K}(t,\tau)}{\sigma_t^2}$ with order $((0,0),2)$ (as well as the analogue kernels for τ) and $\mathcal{K}_{t,\tau}(t,\tau) = \frac{t\tau\mathcal{K}(t,\tau)}{\sigma_{t,\tau}^2}$ with order $((1,1),4)$.

Using Lemma 13 we can obtain the convergence rates for the single terms. For example let $\psi_1(t,\tau,y) := 1$, then using the kernel $\mathcal{K}$ of order $((0,0),2)$,

$$\psi_1(t) = \int 1\, g(y;t,\tau)\, dy = 1 \text{ and } \Psi_{1I}(t,\tau) = \frac{1}{Ih^2NM}\sum_i\sum_n\sum_m 1w_{nm}.$$

According to the remark of Lemma 13 we therefore obtain: ´

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} |\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm} - 1| = O(h^2)$$

The following overview shows only terms in t and mixed terms (the analogue terms in τ have the same convergence rates):

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} \left|\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm} - 1\right| = O(h^2) \text{ (using kernel } \mathcal{K}\text{)}$$

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} \left|\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t) - 0\right| = O(h^2) \text{ (using kernel } \mathcal{K}_t\text{)}$$

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} \left|\frac{1}{Ih^6\sigma_{t,\tau}}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)(\frac{m}{M}-\tau) - 0\right| = O(h^2) \text{ (using kernel } \mathcal{K}_{t,\tau}\text{)}$$

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} \left|\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)^2 - 1\right| = O(h^2) \text{ (using kernel } \mathcal{K}_{t^2}\text{)}$$

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} \left|\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}Y^i_{nm} - \mu(t,\tau)\right| = O_P\left(\frac{1}{\sqrt{I}h^2}\right) \text{ (using kernel } \mathcal{K}\text{)}$$

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} \left|\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)Y^i_{nm} - \mu_t(t,\tau)\right| = O_P\left(\frac{1}{\sqrt{I}h^3}\right) \text{ (using kernel } \mathcal{K}_t\text{)}$$

D_1 therewith has the following rate:

$$\begin{aligned}
D_1 = &\left[\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)Y^i_{nm}\right.\\
&\left.\qquad - \frac{\left(\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)\right)\left(\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}Y^i_{nm}\right)}{\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\right]\\
&\times\left[\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)^2\right.\\
&\left.\qquad - \frac{h^2\sigma_t^2\left(\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)\right)^2}{\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\right]^{-1}\\
= &\frac{\left[\mu_t(t,\tau) + O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right] - \frac{[0+O(h^2)]\left[\mu(t,\tau)+O_P\left(\frac{1}{\sqrt{I}h^2}\right)\right]}{[1+O(h^2)]}}{[1+O(h^2)] - \frac{h^2\sigma_t^2[0+O(h^2)]^2}{[1+O(h^2)]}}
\end{aligned}$$

$$= \mu_t(t,\tau) + O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

uniformly in t, τ, i.e.

$$\sup_{\substack{t \in T \\ \tau \in \mathcal{T}}} |D_1 - \mu_t(t,\tau)| = O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

and analogously

$$\sup_{\substack{t \in T \\ \tau \in \mathcal{T}}} |D_2 - \mu_\tau(t,\tau)| = O_P\left(\frac{1}{\sqrt{I}h^3}\right).$$

For E_1 we obtain:

$$E_1 = \Bigg[h^4\sigma_\tau^2 \left(\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)\right)\left(\frac{1}{Ih^4\sigma_\tau^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{m}{M}-\tau)\right)$$

$$-h^2\frac{\sigma_{t\tau}^2}{\sigma_t^2}\frac{1}{Ih^6\sigma_{t\tau}^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)(\frac{m}{M}-\tau)\Bigg]$$

$$* \Bigg[\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)^2$$

$$-\frac{h^2\sigma_t^2\left(\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)\right)^2}{\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\Bigg]^{-1}$$

$$= \frac{h^4\sigma_\tau^2\left[0+O(h^2)\right]\left[0+O(h^2)\right] - h^2\frac{\sigma_{t\tau}^2}{\sigma_t^2}\left[0+O(h^2)\right]}{\left[1+O(h^2)\right] - \frac{h^2\sigma_t^2[0+O(h^2)]^2}{[1+O(h^2)]}} = O(h^4)$$

uniformly in t, τ, i.e.

$$\sup_{\substack{t \in T \\ \tau \in \mathcal{T}}} |E_1| = O(h^4) \quad \text{and analogously} \quad \sup_{\substack{t \in T \\ \tau \in \mathcal{T}}} |E_2| = O(h^4).$$

With this information, $\hat{\beta}_{11}(t,\tau)$ and $\hat{\beta}_{12}(t,\tau)$ can be evaluated as follows (uniformly in t, τ):

$$\hat{\beta}_{11}(t,\tau) = \frac{\left[\mu_t(t,\tau) + O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right] + \left[0+O(h^2)\right]\left[\mu_\tau(t,\tau) + O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right]}{1 - \left[0+O(h^2)\right]\left[0+O(h^2)\right]}$$

$$= \mu_t(t,\tau) + O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

$$\hat{\beta}_{12}(t,\tau) = \mu_\tau(t,\tau) + O_P\left(\frac{1}{\sqrt{I}h^3}\right)$$

For $\hat{\beta}_0(t,\tau)$ we now obtain:

$$\begin{aligned}\hat{\beta}_0(t,\tau) =& \frac{\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}Y^i_{nm}}{\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}} - \frac{h^2\sigma_t^2\frac{1}{Ih^4\sigma_t^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{n}{N}-t)}{\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\hat{\beta}_{11}(t,\tau)\\ &- \frac{h^2\sigma_\tau^2\frac{1}{Ih^4\sigma_\tau^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}(\frac{m}{M}-\tau)}{\frac{1}{Ih^2}\sum_i \frac{1}{NM}\sum_n\sum_m w_{nm}}\hat{\beta}_{12}(t,\tau)\\ =& \frac{\left[\mu(t,\tau)+O_P\left(\frac{1}{\sqrt{I}h^2}\right)\right]}{[1+O(h^2)]} - \frac{h^2\left[0+O(h^2)\right]}{[1+O(h^2)]}\left[\mu_t(t,\tau)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right]\\ &- \frac{h^2\left[0+O(h^2)\right]}{[1+O(h^2)]}\left[\mu_\tau(t,\tau)+O_P\left(\frac{1}{\sqrt{I}h^3}\right)\right]\\ =& \mu(t,\tau)+O_P\left(\frac{1}{\sqrt{I}h^2}\right)+h^4O_P\left(\frac{1}{\sqrt{I}h^3}\right)\end{aligned}$$

uniformly in t,τ, such that we obtain the final uniform convergence rate

$$\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}}|\hat{\mu}(t,\tau)-\mu(t,\tau)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right).$$

□

Theorem 16 (Theorem for covariance estimation)**.** *Assume that (a),(b) and (d) of Lemma 14 are valid here as well and that*

(c) $h\to 0, Ih^6\to\infty, Ih^{12}=O(1), \frac{1}{N}=O(h^9), M\geq N$ *for* $I\to\infty$.

Further $\mathcal{K}$ *is assumed to be a four-dimensional product kernel of a kernel* $\mathcal{K}^*$ *of order* $(0,2)$*, e.g.* $\mathcal{K}(\vec{t},\vec{\tau})=\mathcal{K}^*(t_1)\mathcal{K}^*(t_2)\mathcal{K}^*(\tau_1)\mathcal{K}^*(\tau_2)$ *for* $\vec{t}\in T^2$, $\vec{\tau}\in\mathcal{T}^2$. $\mathcal{K}$ *is therefore of order ((0,0,0,0),2) (see calculation* (2.7)*).*
The estimator for the covariance function as defined in Section 2.2 satisfies the following uniform convergence rate:

$$\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}}|\hat{G}(\vec{t},\vec{\tau})-G(\vec{t},\vec{\tau})| = O_P\left(\frac{1}{\sqrt{I}h^4}\right) \tag{6.22}$$

Proof. As the estimation of the covariance function is based on the raw covariances with estimated mean function $G_i(\frac{\vec{n}}{N},\frac{\vec{m}}{M}) = (Y^i_{n_1,m_1}-\hat{\mu}(\frac{n_1}{N},\frac{m_1}{M}))(Y^i_{n_2,m_2}-\hat{\mu}(\frac{n_2}{N},\frac{m_2}{M}))$ instead of $\tilde{G}_i(\frac{\vec{n}}{N},\frac{\vec{m}}{M}) = (Y^i_{n_1,m_1}-\mu(\frac{n_1}{N},\frac{m_1}{M}))(Y^i_{n_2,m_2}-\mu(\frac{n_2}{N},\frac{m_2}{M}))$, we will first show that $G_i(\frac{\vec{n}}{N},\frac{\vec{m}}{M})$ is asymptotically equivalent to $\tilde{G}_i(\frac{\vec{n}}{N},\frac{\vec{m}}{M})$ and work with $\tilde{G}_i(\frac{\vec{n}}{N},\frac{\vec{m}}{M})$ in the further proof.
This can be seen observing that (for $\vec{t}\in T^2$ and $\vec{\tau}\in\mathcal{T}^2$)

$$\begin{aligned}G_i\left(\vec{t},\vec{\tau}\right) = \tilde{G}_i\left(\vec{t},\vec{\tau}\right) &+ \left(Y^i_{t_1,\tau_1}-\mu\left(t_1,\tau_1\right)\right)\left(\mu\left(t_2,\tau_2\right)-\hat{\mu}\left(t_2,\tau_2\right)\right)\\ &+ \left(Y^i_{t_2,\tau_2}-\mu\left(t_2,\tau_2\right)\right)\left(\mu\left(t_1,\tau_1\right)-\hat{\mu}\left(t_1,\tau_1\right)\right)\\ &+ \left(\mu\left(t_1,\tau_1\right)-\hat{\mu}\left(t_1,\tau_1\right)\right)\left(\mu\left(t_2,\tau_2\right)-\hat{\mu}\left(t_2,\tau_2\right)\right)\end{aligned}$$

As $\mathrm{Var}(Y_{t,\tau})$ is finite for all $t \in T$, $\tau \in \mathcal{T}$, $\left(Y^i_{t_1,\tau_1} - \mu\left(t_1, \tau_1\right)\right)$ and $\left(Y^i_{t_2,\tau_2} - \mu\left(t_2, \tau_2\right)\right)$ are stochastically bounded. Further $\sup_{\substack{t\in T\\ \tau\in\mathcal{T}}} |\mu(t,\tau) - \hat{\mu}(t,\tau)| = O_P\left(\frac{1}{\sqrt{I}h^2}\right)$ such that

$$\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}} \left|G_i\left(\vec{t},\vec{\tau}\right) - \tilde{G}_i\left(\vec{t},\vec{\tau}\right)\right| = O_P\left(\frac{1}{\sqrt{I}h^2}\right)$$

which shows the asymptotic equivalence of G_i and $\tilde{G}_i$ with respect to (6.22).

In the following we use the abbreviations

$$R := IN(N-1)M(M-1)h^4$$

$$\sum := \sum_i \sum_n \sum_m$$

$$W := \mathcal{K}\left(\frac{t_1 - \frac{n_1}{N}}{h}, \frac{t_2 - \frac{n_2}{N}}{h}, \frac{\tau_1 - \frac{m_1}{M}}{h}, \frac{\tau_2 - \frac{m_2}{M}}{h}\right)$$

$$G_i := G_i\left(\frac{n_1}{N}, \frac{n_2}{N}, \frac{m_1}{M}, \frac{m_2}{M}\right), G := G(\vec{t},\vec{\tau})$$

To obtain the covariance estimation, the following term is to be minimized:

$$F = \sum W\left[G_i - \beta_0 - \beta_{11}(t_1 - \frac{n_1}{N}) - \beta_{12}(t_2 - \frac{n_2}{N}) - \beta_{21}(\tau_1 - \frac{m_1}{M}) - \beta_{22}(\tau_2 - \frac{m_2}{M})\right]^2$$

Hence we evaluate the gradient

$$\nabla F =$$

$$\begin{pmatrix} 2\sum W\left(G_i - \beta_0 - \beta_{11}(t_1 - \frac{n_1}{N}) - \beta_{12}(t_2 - \frac{n_2}{N}) - \beta_{21}(\tau_1 - \frac{m_1}{M}) - \beta_{22}(\tau_2 - \frac{m_2}{M})\right) \\ 2\sum W\left(G_i - \beta_0 - \beta_{11}(t_1 - \frac{n_1}{N}) - \beta_{12}(t_2 - \frac{n_2}{N}) - \beta_{21}(\tau_1 - \frac{m_1}{M}) - \beta_{22}(\tau_2 - \frac{m_2}{M})\right)(t_1 - \frac{n_1}{N}) \\ 2\sum W\left(G_i - \beta_0 - \beta_{11}(t_1 - \frac{n_1}{N}) - \beta_{12}(t_2 - \frac{n_2}{N}) - \beta_{21}(\tau_1 - \frac{m_1}{M}) - \beta_{22}(\tau_2 - \frac{m_2}{M})\right)(t_2 - \frac{n_2}{N}) \\ 2\sum W\left(G_i - \beta_0 - \beta_{11}(t_1 - \frac{n_1}{N}) - \beta_{12}(t_2 - \frac{n_2}{N}) - \beta_{21}(\tau_1 - \frac{m_1}{M}) - \beta_{22}(\tau_2 - \frac{m_2}{M})\right)(\tau_1 - \frac{m_1}{M}) \\ 2\sum W\left(G_i - \beta_0 - \beta_{11}(t_1 - \frac{n_1}{N}) - \beta_{12}(t_2 - \frac{n_2}{N}) - \beta_{21}(\tau_1 - \frac{m_1}{M}) - \beta_{22}(\tau_2 - \frac{m_2}{M})\right)(\tau_2 - \frac{m_2}{M}) \end{pmatrix}$$

with respect to β and solve the equation $\nabla F = 0$:

$$\nabla F = 0$$

$$\Leftrightarrow \quad \underbrace{\frac{1}{R}\begin{pmatrix} \sum WG_i \\ \sum WG_i(t_1 - \frac{n_1}{N}) \\ \sum WG_i(t_2 - \frac{n_2}{N}) \\ \sum WG_i(\tau_1 - \frac{m_1}{M}) \\ \sum WG_i(\tau_2 - \frac{m_2}{M}) \end{pmatrix}}_{\text{stochastic terms}} \tag{6.23}$$

$$= \frac{1}{R} \underbrace{\begin{pmatrix} \sum W & \sum W(t_1 - \frac{n_1}{N}) & \dots & \sum W(\tau_2 - \frac{m_2}{M}) \\ \sum W(t_1 - \frac{n_1}{N}) & \sum W(t_1 - \frac{n_1}{N})^2 & \dots & \sum W(t_1 - \frac{n_1}{N})(\tau_2 - \frac{m_2}{M}) \\ \sum W(t_2 - \frac{n_2}{N}) & \sum W(t_1 - \frac{n_1}{N})(t_2 - \frac{n_2}{N}) & \dots & \sum W(t_2 - \frac{n_2}{N})(\tau_2 - \frac{m_2}{M}) \\ \sum W(\tau_1 - \frac{m_1}{M}) & \sum W(t_1 - \frac{n_1}{N})(\tau_1 - \frac{m_1}{M}) & \dots & \sum W(\tau_1 - \frac{m_1}{M})(\tau_2 - \frac{m_2}{M}) \\ \sum W(\tau_2 - \frac{m_2}{M}) & \sum W(t_1 - \frac{n_1}{N})(\tau_2 - \frac{m_2}{M}) & \dots & \sum W(\tau_2 - \frac{m_2}{M})^2 \end{pmatrix}}_{\text{deterministic terms}} \begin{pmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{pmatrix}$$

The single terms can be evaluated using Lemma 14. As before, we want to show one example how Lemma 14 is used. Let $\theta_1(t_1, t_2, y_1, y_2) := 1$, then using the kernel $\mathcal{K}$ of order $((0,0,0,0),2)$,

$$\theta_1(t) = \int\int 1\, g_2(y_1, y_2; \vec{t}, \vec{\tau})\, dy_1\, dy_2 = 1 \text{ and } \Theta_{1I}(t) = \frac{1}{R}\sum 1W.$$

According to the remark of Lemma 14 we therefore obtain: ´

$$\sup_{t_1,t_2\in T} |\frac{1}{R}\sum W - 1| = O(h^4)$$

Only terms in t_1 and mixed terms of t_1 and t_2 are presented, as the other occurring terms of the same scheme also have the same convergence rates:

$$\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}} \left|\frac{1}{R}\sum W - 1\right| = O(h^2) \text{ (using kernel } \mathcal{K})$$

$$\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}} \left|\frac{1}{Rh^2\sigma_{t_1}^2}\sum W(t_1 - \frac{n_1}{N}) - 0\right| = O(h^2) \text{ (using kernel } \mathcal{K}_{t_1})$$

$$\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}} \left|\frac{1}{Rh^2\sigma_{t_1}^2}\sum W(t_1 - \frac{n_1}{N})^2 - 1\right| = O(h^2) \text{ (using kernel } \mathcal{K}_{t_1^2})$$

$$\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}} \left|\frac{1}{Rh^4\sigma_{t_1t_2}^2}\sum W(t_1 - \frac{n_1}{N})(t_2 - \frac{n_2}{N}) - 0\right| = O(h^2) \text{ (using kernel } \mathcal{K}_{t_1t_2})$$

$$\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}} \left|\frac{1}{R}\sum WG_i - G\right| = O_P\left(\frac{1}{\sqrt{I}h^4}\right) \text{ (using kernel } \mathcal{K})$$

$$\sup_{\substack{\vec{t}\in T^2\\ \vec{\tau}\in\mathcal{T}^2}} \left|\frac{1}{Rh^2\sigma_{t_1}^2}\sum WG_i(t_1 - \frac{n_1}{N}) - G_{t_1}\right| = O_P\left(\frac{1}{\sqrt{I}h^5}\right) \text{ (using kernel } \mathcal{K}_{t_1})$$

Now equation (6.23) can be evaluated using the uniform convergence rates of the single terms:

$$\begin{pmatrix} G+O_P\left(\frac{1}{\sqrt{Ih^4}}\right) \\ h^2\left(G_{t_1}+O_P\left(\frac{1}{\sqrt{Ih^5}}\right)\right) \\ h^2\left(G_{t_2}+O_P\left(\frac{1}{\sqrt{Ih^5}}\right)\right) \\ h^2\left(G_{\tau_1}+O_P\left(\frac{1}{\sqrt{Ih^5}}\right)\right) \\ h^2\left(G_{\tau_2}+O_P\left(\frac{1}{\sqrt{Ih^5}}\right)\right) \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} 1+O(h^2) & h^2(0+O(h^2)) & h^2(0+O(h^2)) & h^2(0+O(h^2)) & h^2(0+O(h^2)) \\ h^2(0+O(h^2)) & 1+O(h^2) & h^4(0+O(h^2)) & h^4(0+O(h^2)) & h^4(0+O(h^2)) \\ h^2(0+O(h^2)) & h^4(0+O(h^2)) & 1+O(h^2) & h^4(0+O(h^2)) & h^4(0+O(h^2)) \\ h^2(0+O(h^2)) & h^4(0+O(h^2)) & h^4(0+O(h^2)) & 1+O(h^2) & h^4(0+O(h^2)) \\ h^2(0+O(h^2)) & h^4(0+O(h^2)) & h^4(0+O(h^2)) & h^4(0+O(h^2)) & 1+O(h^2) \end{pmatrix}}_{=:(\text{first column}|A')=:A} \begin{pmatrix} \hat\beta_0 \\ \hat\beta_{11} \\ \hat\beta_{12} \\ \hat\beta_{21} \\ \hat\beta_{22} \end{pmatrix}$$

According to Cramer's rule (e.g. Fischer [2002]) we now obtain the following convergence rate for the interesting estimator $\hat\beta_0$:

$$\hat\beta_0 = \frac{\det\begin{pmatrix} G+O_P\left(\frac{1}{\sqrt{Ih^4}}\right) & | & \\ h^2\left(G_{t_1}+O_P\left(\frac{1}{\sqrt{Ih^5}}\right)\right) & | & \\ h^2\left(G_{t_2}+O_P\left(\frac{1}{\sqrt{Ih^5}}\right)\right) & | & A' \\ h^2\left(G_{\tau_1}+O_P\left(\frac{1}{\sqrt{Ih^5}}\right)\right) & | & \\ h^2\left(G_{\tau_2}+O_P\left(\frac{1}{\sqrt{Ih^5}}\right)\right) & | & \end{pmatrix}}{\det(A)}$$

$$= \frac{\left(G+O_P\left(\frac{1}{\sqrt{Ih^4}}\right)\right)\left(1+O(h^2)\right)^4 + \left(G+O_P\left(\frac{1}{\sqrt{Ih^4}}\right)\right)O(h^{10}) + h^2 O_P\left(\frac{1}{\sqrt{Ih^5}}\right)O(h^4)}{\left(1+O(h^2)\right)^9 + \left(1+O(h^2)\right)O(h^{10}) + O(h^{26})}$$

$$= G + O_P\left(\frac{1}{\sqrt{Ih^4}}\right)$$

uniformly in $(\vec{t},\vec{\tau})$ and therewith

$$\sup_{\substack{\vec{t}\in\mathcal{T}^2 \\ \vec{\tau}\in\mathcal{T}^2}} \left|G(\vec{t},\vec{\tau}) - \hat{G}(\vec{t},\vec{\tau})\right| = O_P\left(\frac{1}{\sqrt{Ih^4}}\right)$$

□

Finally convergence rates for eigenvalues and eigenfunctions can be derived. Weyl's eigenvalue inequality of Lemma 11 can be applied in the two-dimensional case as well.

Theorem 17 (Theorem for eigenanalysis estimation)**.** *Under the assumptions of Theorems 15* *and 16, we obtain the following consistency rate for an eigenvalue k for any fixed $k \in \mathbb{N}$:*

$$|\hat{\lambda}_k - \lambda_k| = O_P\left(\frac{1}{\sqrt{I h^4}}\right). \tag{6.24}$$

If λ_k has multiplicity 1, the appropriately standardized estimator for the eigenfunction $\hat{\rho}_k$ has the following convergence rate in the L_2 sense

$$||\hat{\rho}_k - \rho_k||_{L_2} = O_P\left(\frac{1}{\sqrt{I h^4}}\right) \tag{6.25}$$

and fulfills furthermore the uniform consistency rate

$$\sup_{\substack{t \in T \\ \tau \in \mathcal{T}}} |\hat{\rho}_k(t,\tau) - \rho_k(t,\tau)| = O_P\left(\frac{1}{\sqrt{I h^4}}\right). \tag{6.26}$$

Proof. The framework is the same as in the one-dimensional case. We consider the Hilbert space $H = \{f : T \times \mathcal{T} \to \mathbb{R} | \int |f(t)|^2 \, dt < \infty\}$, this time with the inner product $\langle f, g \rangle = \int \int |f(t,\tau) g(t,\tau)| \, dt \, d\tau$ and norm $||f||_{L_2} = \sqrt{\int \int |f(t,\tau)|^2 \, dt \, d\tau}$.

An operator $T : H \to H$ is called Hilbert-Schmidt operator on H, if for an orthonormal system $\{e_i\}$ of H, the Hilbert-Schmidt norm is finite, e.g. $\sum_i ||T e_i||^2 < \infty$. All Hilbert-Schmidt operators on H form the space $\sigma_2(H)$.

The inner product on $F := \sigma_2(H)$ for $T_1, T_2 \in \sigma_2(H)$ is given by $\langle T_1, T_2 \rangle_F = \sum_j \langle T_1 e_j, T_2 e_j \rangle_H$. One can show that the real and estimated covariance operators

$$G(f) = \int \int G(\vec{t}, \vec{\tau}) f(t_1, \tau_1) \, dt_1 \, d\tau_1$$

$$\hat{G}(f) = \int \int \hat{G}(\vec{t}, \vec{\tau}) f(t_1, \tau_1) \, dt_1 \, d\tau_1$$

have a finite Hilbert-Schmidt norm and therefore are Hilbert-Schmidt operators (see Heuser [2006, Chapter 12]).

Like in the proof of Theorem 12, we obtain the following relationship between the supremum norm (for which we know the order of convergence from Theorem 16) and the Hilbert-Schmidt norm:

$$\begin{aligned} ||\hat{G} - G||_F = ||\hat{G} - G||_{L_2} &= \sqrt{\int \int \left|\hat{G}(\vec{t}, \vec{\tau}) - G(\vec{t}, \vec{\tau})\right|^2 \, d\vec{t} d\vec{\tau}} \\ &\leq \sup_{\substack{\vec{t} \in T^2 \\ \vec{\tau} \in \mathcal{T}^2}} \left|\hat{G}(\vec{t}, \vec{\tau}) - G(\vec{t}, \vec{\tau})\right| \underbrace{\sqrt{\int \int 1 \, d\vec{t} d\vec{\tau}}}_{|T||\mathcal{T}|} = O_P\left(\frac{1}{\sqrt{I h^4}}\right) \end{aligned}$$

Applying Weyl's inequality as in Theorem 12, the rate for the eigenvalues (6.24) follows:

$$|\hat{\lambda}_k - \lambda_k| = O_P\left(\frac{1}{\sqrt{I h^4}}\right)$$

Further we can deduce exactly as in the one-dimensional case, that in the case of uni- and multi-dimensional eigenspaces, we get for the projections

$$P_k = \sum_{j \in I_k} \rho_j \otimes \rho_j$$

$$\hat{P}_k = \sum_{j \in I_k} \hat{\rho}_j \otimes \hat{\rho}_j$$

with $I_k := \{j : \lambda_j = \lambda_k\}$:

$$||\hat{P}_k - P_k||_F = O_P\left(\frac{1}{\sqrt{I}h^4}\right)$$

and for the case of a one-dimensional eigenspace k:

$$||\hat{\rho}_k - \rho_k||_H^2 \leq ||\hat{P}_k - P_k||_F^2$$

Hence equation (6.25) follows.
In order to prove (6.26) we first observe using the Cauchy-Schwarz inequality, Theorem 16 and (6.25) that:

$$|\hat{\lambda}_k \hat{\rho}_k(t_2, \tau_2) - \lambda_k \rho_k(t_2, \tau_2)| = \left| \int \int \hat{G}(\vec{t}, \vec{\tau}) \hat{\rho}_k(t_1, \tau_1)\, dt_1\, d\tau_1 - \int \int G(\vec{t}, \vec{\tau}) \rho_k(t_1, \tau_1)\, dt_1\, d\tau_1 \right| \tag{6.27}$$

$$\leq \int \int |\hat{G}(\vec{t}, \vec{\tau}) - G(\vec{t}, \vec{\tau})| |\hat{\rho}_k(t_1, \tau_1)|\, dt_1\, d\tau_1$$

$$+ \int \int |G(\vec{t}, \vec{\tau})| |\hat{\rho}_k(t_1, \tau_1) - \rho_k(t_1, \tau_1)|\, dt_1\, d\tau_1$$

$$\leq \sqrt{\int \int (\hat{G}(\vec{t}, \vec{\tau}) - G(\vec{t}, \vec{\tau}))^2\, dt_1\, d\tau_1} + \sqrt{\int \int G^2(\vec{t}, \vec{\tau})\, dt_1\, d\tau_1} ||\hat{\rho}_k - \rho_k||_H$$

$$= O_P\left(\frac{1}{\sqrt{I}h^4}\right).$$

$\sup_{\substack{t \in \mathcal{T} \\ \tau \in \mathcal{T}}} |\frac{\hat{\lambda}_k \hat{\rho}_k(t,\tau)}{\lambda_k} - \rho_k(t, \tau)| = O_P\left(\frac{1}{\sqrt{I}h^4}\right)$ follows in (6.27) and using (6.24), we also obtain $\sup_{\substack{t \in \mathcal{T} \\ \tau \in \mathcal{T}}} |\hat{\rho}_k(t, \tau) - \rho_k(t, \tau)| = O_P\left(\frac{1}{\sqrt{I}h^4}\right)$.

□

7 Implementation in R

In order to execute the one- and two-dimensional nonparametric FPCA we implemented the methods as described in Chapters 2 and 3 in the statistical computing software R. We chose an object-oriented approach based on the so-called *S4 classes* as described in Chambers [2009] in order to make the functions easily applicable. The implementation resulted in a package called FPCA. In the following the main functions and their usage are presented. For a more detailed description of all functions please refer to the manual of the package (Winzenborg [2011]).
Before using the functions, the package has to be installed and loaded with the command `library(FPCA)`.

7.1 One-dimensional implementation

As mentioned above the package is organized object-oriented. This means for the one-dimensional case that we defined a class for functional data objects (called `fpcadat`) that combines all information necessary to understand the data and a second class that includes the results and inputs of the FPCA calculation (the class is called `fpcaobj`). `fpcaobj` further inherits all information from `fpcadat` and is therefore a child class of `fpcadat`.
Hence, at first one has to create an object of the class `fpcadat` which is created in order to define the structure of the data such that the later FPCA analysis knows which variable defines the time, which variable the outcome and which the subject. Furthermore, a title can be specified that is used in some plots as part of the default title.
The class syntax of `fpcadat` is the following:

- `dt`: Object of class `"data.frame"`: data frame containing (at least) the columns xvar, yvar and subvar
- `xvar`: Object of class `"character"`: name of x column (e.g. time)
- `yvar`: Object of class `"character"`: name of y column (outcome)
- `subvar`: Object of class `"character"`: name of subject column
- `title`: Object of class `"character"`: title which is used in plots

The usage of the package is shown with the example of the Wiener process that was analyzed in Section 2.3. The package also includes the functions to simulate the one- and two-dimensional Wiener process. For the one-dimensional process we need to specify the number of processes N and the number of steps s for the simulation. In order to simulate 50 realizations of a Wiener process with a step width of $\frac{1}{20}$, we write:

```
N = 50; s = 20
wienermat = wiener(N,s)
```

The result is a matrix with a row for each observation. As the FPCA method expects a data frame, we transform the matrix into a data frame with columns for time, outcome and subject:

```
wienerdat = data.frame(t = wienermat$t,
  wiener = as.vector(t(wienermat$wienermat)),
  obs = rep(1:N,each = length(wienermat$t)))
```

Afterwards we are able to construct the FPCA data object by calling:

```
wiener_fpca_dat = new("fpcadat", dt = wienerdat, xvar = "t", yvar = "wiener",
  subvar = "obs", title = paste("Wiener, step width:", 1/s))
```

The structure of `wiener_fpca_dat` can be displayed via

```
str(wiener_fpca_dat)
```

and the result is the following:

```
Formal class 'fpcadat' [package "FPCA"] with 5 slots
  ..@ dt    :'data.frame': 1050 obs. of  3 variables:
  .. ..$ t     : num [1:1050] 0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 ...
  .. ..$ wiener: num [1:1050] 0 0.157 0.63 0.588 0.739 ...
  .. ..$ obs   : int [1:1050] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ xvar  : chr "t"
  ..@ yvar  : chr "wiener"
  ..@ subvar: chr "obs"
  ..@ title : chr "Wiener, step width: 0.05"
```

This object contains the basis information which is necessary to call the FPCA analysis. Only the bandwidths have to be specified additionally:

```
wiener_fpca_out = fpca(wiener_fpca_dat, bwmean = 0.2, bwcov = 0.2,
  bwerrvar = 0.2)
```

The calculation takes a few seconds.
In this case the eigenfunctions are calculated for the time points where the original observations were measured (or in this case simulated). In order to evaluate the eigenfunctions at other time points, we can specify the evaluation points `xeval` accordingly:

```
wiener_fpca_out2 = fpca(wiener_fpca_dat, bwmean = 0.2, bwcov = 0.2,
  bwerrvar = 0.2, xeval = seq(0,1,1/40))
```

Naturally the calculation takes longer for larger `xeval` vectors.
The function `fpca` automatically creates an object of class `fpcaobj` which summarizes the data and all information of the FPCA analysis and has the following structure:

- bwmean: Object of class "numeric": bandwidth for mean smoothing
- bwcov: Object of class "numeric": bandwidth for covariance smoothing
- bwerrvar: Object of class "numeric": bandwidth for error variance smoothing
- npc: Object of class "integer": number of principal components to be calculated
- xeval: Object of class "numeric": time points where discretized functions are evaluated
- xuni: Object of class "numeric": all time points that occur in the observations
- meanfct: Object of class "numeric": smoothed and discretized mean function
- covfct: Object of class "matrix": smoothed and discretized covariance function
- rawcov: Object of class "matrix": data table with raw covariance values for each subject
- errvar: Object of class "numeric": estimated error variance
- eigvalues: Object of class "numeric": eigenvalues in the same order as the eigenfunctions
- eigfcts: Object of class "matrix": discretized eigenfunctions, each column defines one eigenfunction
- eigscores: Object of class "matrix": each row contains the scores of one subject ordered as in dt
- eigscores_shrinked: Object of class "matrix": each row contains the shrinked scores of one subject ordered as in dt
- eigpercvar: Object of class "numeric": vector with percentages of variability explained by each eigenfunction
- dt: Object of class "data.frame": data frame containing (at least) columns xvar, yvar and subvar
- xvar: Object of class "character": name of x column (e.g. time)
- yvar: Object of class "character": name of y column (outcome)
- subvar: Object of class "character": name of subject column
- title: Object of class "character": title which is used in plots

Several functions are available to plot the results. The standard plot command calls plotpcs if used on an fpcaobj object. Therefore

```
plot(wiener_fpca_out)
plot(wiener_fpca_out2)
```

lead to plots where all calculated eigenfunctions are plotted in one graph. In order to obtain plots as in Figures 2.4, the additional option `trellis = T` has to be specified. This is helpful for black-and-white figures as in this case. Four is the default number of eigenfunctions that is calculated and plotted, but other numbers of eigenfunctions can be specified for calculation and for plotting.

Furthermore, mean and covariance functions as well as the scores of the observations can be plotted by calling:

```
plotmean(list(wiener_fpca_out))
plotcov(wiener_fpca_out)
plotscores(wiener_fpca_out)
```

The `plotmean` function expects a list of `fpcaobj` objects. This has the advantage that one is able to plot several mean functions in one graph as was for example done in Figure 4.9(a).

Standard options are used for all plot functions if no further specifications are made. If no specifications of titles and/or axes labels are made, the package derives reasonable default values. Alternatively it is possible to specify these and other options using the standard R plot syntax. For details please consider the package help (Winzenborg [2011]).

Additionally, the package gives the possibility to cluster the resulting scores. We did not perform clustering procedures on Wiener scores as this is of limited use, but for a demonstration how to use the package, we show the function calls.

```
kmeans_out = kmeans_scores(wiener_fpca_out, numclust = 3, nscores = 3)
```

`kmeans_out` is a list with two entries - the first entry is the input `fpca_obj` object with an additional column `clusters` in the original data set `@dt` which reflects the cluster membership of each subject. The second entry of the return list is simply an integer vector with the cluster indices for each subject. If we now call again the `plotscores` function with the additional parameter `index = "clusters"`, the resulting graph is colored according to cluster memberships:

```
plotscores(kmeans_out[[1]], index = "clusters")
```

By the way, `index` may be any other variable name of the data set by which the scores shall be colored.

Another interesting graph is provided by

```
plotclust(kmeans_out[[1]])
```

and was used for example to produce Figure 4.4. It displays the original observation curves according to their cluster memberships.

Additionally to K-means clustering we provide hierarchical clustering via the function

```
hierarch_scores(wiener_fpca_out)
```

with the same in- and outputs. It is designed to use standard inputs for the hierarchical clustering procedures `hclust` and `cutree`. Alternative clustering procedures can easily be

applied and the plotting functions can also be used if the additional column to the object corresponding to `wiener_fpca_out` is likewise added.
If one decides that it is not reasonable to form clusters as we did regarding Figure 4.11, one can use another plot function in order to still divide the subjects for plotting. The function is called via:

```
plotquad(wiener_fpca_out)
```

Again, further specifications are possible, for example the coloring according to a specified variable as was used in Figure 4.11.
These were essentially the functions available for one-dimensional FPCA with their main calls. For a more detailed treatment of all functions with all options available, please consider the package reference.

7.2 Two-dimensional implementation

The implementation and call of functions is essentially the same as in the one-dimensional case, but one should consider that the calculations are computationally more intense than before, mainly due to the four-dimensional covariance function estimation.
First of all, we again defined a class, this time called `fpcadat2d`, containing the functional data and a child class `fpcaobj2d` which additionally includes the FPCA results.
The class `fpcadat2d` has the same slots as its one-dimensional analogue apart from the `xvar` slot. The former `xvar` slot is this time separated into two slots for both space directions. Hence the structure looks as follows:

- `dt`: Object of class `"data.frame"`: data frame containing (at least) columns x1var, x2var, yvar and subvar
- `x1var`: Object of class `"character"`: name of x1 column (e.g. spatial dimension 1)
- `x2var`: Object of class `"character"`: name of x2 column (e.g. spatial dimension 2)
- `yvar`: Object of class `"character"`: name of y column (outcome)
- `subvar`: Object of class `"character"`: name of subject column
- `title`: Object of class `"character"`: title which is used in plots

The procedure is again illustrated by using a Wiener process to construct some test data and show the function calls. Realizations of Wiener processes are simulated via

```
N = 50; s = 20
wienerdat = wiener2d(N, s)
```

for 50 realizations with a step width of $\frac{1}{20}$ in both dimensions. `wienerdat` contains the information about the evaluation points in both dimensions as well as the realizations in a three-dimensional array. We transform the data into a data frame and create the `fpcadat2d` object afterwards:

```
dat = data.frame(subj = rep(1:N, each = length(wienerdat$t)^2),
  x1 = rep(wienerdat$t, each = length(wienerdat$t)),
  x2 = wienerdat$t, y = as.vector(wienerdat$wienermat))

wiener2d_fpca_dat = new("fpcadat2d", dt = dat, x1var = "x1", x2var = "x2",
  yvar = "y", subvar = "subj", title = "Wiener2D")
```

In difference to the one-dimensional case, we provide more options in the estimation of the functional principal components in the two-dimensional case. This is because we allow performing the estimation with different degrees of smoothing.
The call for the full smoothing procedure is analogous to the one-dimensional call and calculates a default number of four principal components:

```
wiener2d_fpca_out_fullsmo =
  fpca2d(wiener2d_fpca_dat, bwmean = 0.2, bwcov = 0.2, smoothest = T)
```

`smoothest` is not necessary to specify in this case, as `smoothest = T` is the standard parameter setting. The output is an object of class `fpcaobj2d`, which is similar to `fpcaobj` with some adjustments to the two-dimensional case:

- `bwmean`: Object of class `"numeric"`: bandwidth for mean smoothing
- `bwcov`: Object of class `"numeric"`: bandwidth for covariance smoothing
- `bwerrvar`: Object of class `"numeric"`: bandwidth for error variance smoothing
- `npc`: Object of class `"integer"`: number of principal components to be calculated
- `xeval`: Object of class `"matrix"`: spatial points where discretized functions are evaluated
- `xuni`: Object of class `"matrix"`: all spatial points that occur in the observations
- `meanfct`: Object of class `"numeric"`: smoothed and discretized mean function
- `covfct`: Object of class `"data.frame"`: smoothed and discretized covariance function
- `errvar`: Object of class `"numeric"`: estimated error variance (not calculated)
- `eigvalues`: Object of class `"numeric"`: eigenvalues in the same order as the eigenfunctions
- `eigfcts`: Object of class `"array"`: discretized eigenfunctions, each matrix indexed by first dimension defines one eigenfunction
- `eigscores`: Object of class `"matrix"`: each row contains the scores of one subject ordered as in dt
- `eigpercvar`: Object of class `"numeric"`: vector with percentages of variability explained by each eigenfunction
- `dt`: Object of class `"data.frame"`: data frame containing (at least) columns x1var, x2var, yvar and subvar

- `x1var`: Object of class `"character"`: name of x1 column (e.g. spatial dimension 1)
- `x2var`: Object of class `"character"`: name of x2 column (e.g. spatial dimension 2)
- `yvar`: Object of class `"character"`: name of y column (outcome)
- `subvar`: Object of class `"character"`: name of subject column
- `title`: Object of class `"character"`: title which is used in plots

The first alternative to the full smoothing solution, which already allows a lot faster calculation and which was used in the calculations of Section 3.3, is to smooth observations and eigenfunctions, but else perform the FPCA calculation on the discretized data like in the multivariate case:

```
wiener2d_fpca_out_simpsmo = fpca2d(wiener2d_fpca_dat, bwmean = 0.2, bwcov = 0.2,
  smoothest = F, smoothobs = T, smootheigfcts = T)
```

As can be seen regarding the specified options, it is also possible to either smooth the observations or the eigenfunctions instead of both. The fastest computation is reached if no smoothing is performed as in the following call:

```
wiener2d_fpca_out_nosmo = fpca2d(wiener2d_fpca_dat, bwmean = 0, bwcov = 0,
  smoothest = F)
```

The bandwidths are ignored in this call, but as the code needs values in order to create the `fpcaobj2d` object, we specify them here as zero. This last way can only be used if the observations are available on a regular grid without any missing data, because here no chance is given to compensate for missing values.

After performing the FPCA calculation, similar plotting methods like in the one-dimensional case are available. In order to provide the spatial plots, we use the graphic functions `wireframe`, `levelplot` and `contourplot` of the `lattice` package and allow specifying which type should be used. Additionally, the opportunity is given to choose between a colored and a black-and-white color scheme.

Plotting methods exist for mean function, principal components and the FPC scores. The mean function is plotted calling

```
plotmean2d(wiener2d_fpca_out_simpsmo, type = "wireframe")
```

The default is to produce a colored `wireframe` plot. Other plot types can be specified by using for example `plottype = contourplot`.

In order to plot the principal components, we need again only to call the `plot` function with an appropriate object:

```
plot(wiener2d_fpca_out_simpsmo)
```

If `plot` is called without specifying the number of principal components to plot, all calculated eigenfunctions are plotted in trellis graphs. Else the eigenfunction(s) to be plotted can be specified as a vector (option `npc`).

The scores are plotted via

```
plotscores2d(wiener2d_fpca_out_simpsmo)
```

and this plot can be modified using the same options as in the one-dimensional variant. Specific cluster functions for the two-dimensional case do not exist, but the functions presented for the one-dimensional case work here as well.
For details and further options we again refer to Winzenborg [2011].

8 Summary and Discussion

This thesis deals with the extension of nonparametric functional principal component analysis to spatial data. We have seen that the theoretic FPCA can be derived in the one- as well as the spatial case from the general Hilbert space spectral theory.

For estimating the principal components we used a nonparametric approach, which means in our case to estimate at first the mean and covariance function using local smoothing methods and afterwards calculate the eigenfunctions, eigenvalues and scores through discrete approximations. As this method is computationally very intense, above all in the spatial case if many observations were made or the evaluation time points are relatively dense, we considered alternatively the method of smoothing the single observations and afterwards calculating the mean and covariance function without smoothing. This method is usable if the observations are sufficiently regular and dense.

The aforementioned computational complexity is certainly a disadvantage of this kind of nonparametric approach. In comparison with the parametric approach in Chapter 4.3, where the observations are represented in a system of basis function and the PCA is performed multivariately on the coefficients of the basis representation, it was much slower, but we decided to use this approach anyway as it has the advantage of not being dependent on the choice of the basis. The approach of Ramsay and Silverman could easily be adapted to the two-dimensional case as it is only necessary to represent the observations in a two-dimensional functional basis. The rest of the calculation would be identical to the one-dimensional case.

We showed in various applications in the one- as well as the two-dimensional case that FPCA is very useful to understand and describe the variability structure of a data set. Further one obtains the possibility to compare observations based on few scores instead of the whole curves. This is useful for example in order to find groups of similar curves using cluster analysis or to detect outliers. In application 5.2 the spatial observations had further a time dimension and we showed how the scores can be used to evaluate structure changes over time. In this situation one could derive a more sophisticated framework to include the time dimension directly into the FPC analysis, but for this data application the way we performed the spatial FPCA and analyzed the scores described the data sufficiently.

One could ask whether it is even reasonable to estimate covariance functions in an infinite-dimensional setting as we know from multivariate questions, that one has to consider the curse of dimensionality, which denotes the sparseness of data in high-dimensional spaces. This means that one in general needs a high number of observations if the number of variables grows in order to evaluate a data set (see for example Hastie et al. [2001, Section 2.5] for a demonstrative description of this matter). In functional data analysis we have the contrary situation: Many (theoretically infinite many) variables and often a small number of observations in comparison. But the crucial difference to multivariate analysis is the smoothness and therefore high correlation of neighboring points in time or space. Hence it is even favorable to have the measurements as dense as possible (see also Ferraty and Vieu [2006] or Hall and Hosseini-Nasab [2006] for a discussion of this matter).

The nonparametric approach we performed is also secured through consistency calculations. For the one-dimensional case Yao et al. already made consistency calculations which we carried out in greater detail. Further we could extend the calculations to the two-dimensional case, if the parameters I (number of observations), h_1, h_2 (smoothing bandwidths for mean and covariance) and N, M (number of measurement points of each observation in both dimensions) have certain relationships. We assumed for the calculations that the measurements occur on a fixed regular grid as is the case in our analysis, but we see no conceptual problem in calculating the results also in the case of irregularly distributed measurements if N and M are the expected numbers of measurements per observation as Yao et al. did in the one-dimensional case (Yao et al. [2005]).

Further one-dimensional consistency results for somewhat different data or estimation situations are for example derived by Hall and Hosseini-Nasab [2006] and Kneip and Utikal [2001]. In Hall et al. [2006], the authors made simulations in order to obtain confidence bands for the eigenfunctions.

Rates can only be calculated for eigenfunctions with one-dimensional eigenspaces, because in the case of multi-dimensional eigenspaces the eigenfunctions cannot be unambiguously defined. This makes it impossible to compare directly compare theoretic and estimated eigenfunctions in multi-dimensional eigenspaces.

Furthermore, as in the estimation process we always obtain pairs of eigenvalues and eigenfunctions, it is likely that we have multi-dimensional eigenspaces in the theoretic process, but not in the estimated process.

This problem occurs in the case of the two-dimensional Wiener process in Section 3.3 and its cause is the symmetry in both dimensions. We showed in this section, how one can alternatively compare projections on eigenspaces. In general, if we deal with a process that is symmetric in both dimensions, all non-symmetric eigenfunctions must belong to an at least two-dimensional eigenspace, because the structure of the process is the same in both dimensions. Hence the condition of one-dimensional eigenspaces is more critical in the two- as in the one-dimensional case.

Otherwise, in practice one seldom deals with processes that are assumed to be symmetric in both dimensions. In our applications in Section 5 for example the axes of both dimensions are already different such that the underlying process cannot be symmetric. In non-symmetric settings however the risk of having multi-dimensional eigenspaces should not be much higher than in one-dimensional problems.

The FPCA framework and estimation technique is theoretically expendable to more than two dimensions, but this expansion would mainly be of theoretic interest, because application areas are missing. Furthermore, the estimation of the covariance function, which also is complicated in the two-dimensional case, would become even more critical.

9 German Introduction

In vielen Bereichen (z.B. Medizin, Biologie, Ökologie und Ökonometrie) werden Daten gemessen, die von Natur aus einen funktionalen Zusammenhang haben. Ein Beispiel für Messungen über die Zeit im medizinischen Bereich stellen regelmäßige Arztbesuche von Patienten dar, bei denen Indikatoren für den medizinischen Zustand, sogenannte Biomarker, gemessen werden. Räumliche Beispiele sind ebenso häufig anzutreffen, zum Beispiel im Bereich der Bildanalyse und der Ökologie. Der funktionale Zusammenhang der Daten soll adressiert werden, indem Methoden der funktionalen Datenanalyse verwendet werden. Doch was bedeutet funktionale Datenanalyse?

Um dies mittels Wahrscheinlichkeitstheorie auszudrücken, sei T eine Indexmenge und $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Die Funktion $Y : T \times \Omega \to \mathbb{R}$ ist eine funktionale Variable, falls $Y(t, \cdot) : \Omega \to \mathbb{R}$ eine univariate Zufallsvariable für jedes $t \in T$ und T unendlich ist. Falls T endlich wäre, würde es sich wiederum um den multivariaten Fall handeln. Daher besteht ein funktionaler Datensatz aus Beobachtungen von I funktionalen, gleichverteilten Variablen $Y_1, \ldots, Y_I$. Diese Arbeit behandelt den Fall, in dem T entweder ein reelles Intervall (z.B. in der Zeit) oder ein Rechteck in $\mathbb{R}^2$, also wir räumliche Messungen vorliegen haben, und in dem die Pfade $Y(\cdot, \omega)$ stetig für alle $\omega \in \Omega$ sind.

In praktischen Anwendungen beobachtet man nicht die Funktionen an sich, sondern stattdessen werden Beobachtungen an diskreten Messpunkten gemacht. Weiterhin können Messungen fehlerbehaftet sein und an unterschiedlichen Messpunkten pro Individuum auftreten (z.B. im Fall von Arztbesuchen werden diese häufig nicht exakt an den gleichen Tagen von allen Patienten vorgenommen). Daher müssen geeignete Methoden angewandt werden, um glatte Beobachtungen oder z.B. Momentenschätzer zu erhalten.

Es könnte an dieser Stelle die Frage auftreten, warum diese Art von Daten funktional und nicht multivariat behandelt wird. Der Grund dafür ist, neben der Behandlung von den soeben angesprochenen Irregularitäten in den Daten, das Einbeziehen von Informationen aus der Umgebung, was gemacht werden kann, sobald ein gewisser Grad von Stetigkeit angenommen wird.

Es existieren im Wesentlichen zwei hauptsächliche Richtungen im Bereich der funktionalen Datenanalyse. Im ersten Ansatz wird ein Satz von Basisfunktionen über T definiert und die Messungen werden durch diese Basisfunktionen repräsentiert. Auswertungen werden dann basierend auf den Koeffizienten dieser Darstellung vorgenommen. Ramsay and Silverman [2006] ist ein anwendungsorientiertes Referenzbuch für diese Richtung mit weiteren Anwendungen in Ramsay and Silverman [2002]. Für eine gute Zusammenfassung siehe Levitin et al. [2007].

Im Gegensatz dazu existiert ein Ansatz, der ohne diese Art der Parametrisierung auskommt. Anstatt dessen werden Glättungsverfahren (hauptsächlich nichtparametrische) angewendet um glatte Beobachtungen zu erhalten. Eine Einführung für diesen Bereich wird in Ferraty and Vieu [2003] gegeben sowie eine ausführlichere Darstellung in Ferraty and Vieu [2006].

Da funktionale Prozesse (zumindest theoretisch) eine unendliche Anzahl von Dimensionen besitzen, ist es äußerst wichtig sich auf die bedeutenden Informationen zu konzentrieren um einen Überblick der Struktur des Prozesses zu erhalten. Eine Methode um dies zu erreichen ist die Hauptkomponentenanalyse, die es erlaubt, die hauptsächlichen Variationsrichtungen zu extrahieren und somit den unendlich-dimensionalen Prozess mit großer Genauigkeit durch eine kleine, endliche Basis auszudrücken. Hauptkomponentenanalyse ist ein effizienter Weg, um die Daten durch ein orthonormales System auszudrücken, denn das Hauptkomponentensystem ist optimal unter allen möglichen orthonormalen Systemen in dem Sinne, dass es den Großteil der Variabilität des originalen Prozesses erhält.

Hauptkomponentenanalyse ist eine sehr populäre Methode in der multivariaten Datenanalyse, da sie die Anzahl der Dimensionen eines hoch-dimensionalen Datensatzes auf ein paar wenige relevante reduziert. Die ersten Hauptkomponenten, die lineare Kombinationen der ursprünglichen Variablen sind, sind optimal im Sinne, dass sie die meiste Variation des Datensatzes unter allen möglichen orthogonalen Linearkombinationen erklären. Einen ausgiebigen Überblick zur multivariaten Datenanalyse liefert Jolliffe [2004]. Die Hauptkomponentenanalyse ist konzeptionell einfach auf den funktionalen Fall zu übertragen und ist dort sogar von noch größerem Wert, da multivariate Hauptkomponentenanalyse häufig das Problem hat, dass Variablen in einem multivariaten Datensatz verschiedenste Eigenschaften mit sehr unterschiedlichen Wertebereichen und Bedeutungen enthalten können. Durch die stetige Indexmenge im funktionalen Fall sind die Hauptkomponenten wie die Beobachtungen ebenfalls Kurven und können direkt als die hauptsächlichen Varianzrichtungen interpretiert werden.

Das Hauptthema dieser Arbeit ist die Erweiterung der funktionalen Hauptkomponentenanalyse (FPCA) über die Zeit auf räumliche Daten. The Möglichkeit dieser Erweiterung wurde von einigen Autoren erwähnt, z.B. von Jolliffe [2004, Section 12.3], Ramsay and Silverman [2006, Section 8.5.3] und Yao et al. [2005]. Eine frühe Arbeit (Preisendorfer and Mobley [1988, Section 2d]) führt zwei räumliche FPCA-Ansätze aus: Im ersten Ansatz wird die Fragestellung durch ein diskretes duales Problem ausgedrückt, das einfacher zu lösen ist. Im zweiten Ansatz werden räumliche Basisfunktionen verwendet. Dieser Ansatz ist im Wesentlichen eine räumliche Variante des Ansatzes von Ramsay and Silverman [2006]. Braud et al. [1993] wenden die Methode von Bouhaddou [1987] an, die ebenfalls einen dualen Ansatz benutzt. Erweiterungen des nichtparametrischen Ansatzes inklusive Glätten, Konvergenzresultaten und Implementierungen fehlen unseres Wissens nach.

Daher erweitern wir die Schätzmethode von Yao et al. [2003] und Yao et al. [2005] auf räumliche Daten und zeigen Konsistenz der Schätzer unter gewissen Voraussetzungen. Um eine solide Grundlage für diese Erweiterung zu haben, präsentieren wir zunächst den ein-dimensionalen Fall und erweitern anschließend den theoretischen Rahmen sowie die nichtparametrische Schätzung von Mittelwert- und Kovarianzfunktion sowie der Hauptkomponenten. Für die Schätzung der Hauptkomponenten und der zugrunde liegenden Mittelwert- und Kovarianzfunktion verwenden wir lokale Regressionsschätzmethoden, aber geben im zweidimensionalen Fall gleichzeitig eine Alternative, bei der nur Beobachtungen und Hauptkomponenten, nicht aber Mittelwert- und Kovarianzfunktion geglättet werden. Diese Alternative ist ebenso gut anwendbar im Fall von nicht-spärlichen Daten und hat den Vorteil, viel Rechenzeit zu sparen. Für den ein- sowie den zweidimensionalen Fall demonstrieren wir die Methode anhand von Simulationen eines Wiener-Prozesses. Bei der Behandlung des zweidimensionalen Wiener-Prozesses gibt sich dabei auch noch die Gelegenheit, zu zeigen, wie man mit multi-dimensionalen Eigenräumen umgehen kann. Anschließend werden sowohl im ein- als auch im zweidimensionalen Fall Anwendungen in unterschiedlichen Datensituationen gezeigt und schließlich leiten wir Konvergenzraten für die zweidimensionale Schätzung her. Im eindimensionalen Fall sind in leicht anderer Datensituation bereits Konvergenzraten von Yao et al. [2005] vorhanden. Außerdem haben wir die ein- und zweidimensionalen Methoden in einem R-Paket implementiert unter Verwendung von S4-Klassen, dem aktuellen R-Standard objektorientierter Programmierung (siehe Chambers [2009]).

Bibliography

M. Benko, W. Härdle, and A. Kneip. Common functional principal components. *The Annals of Statistics*, 37(1):1–34, 2009.

P. Besse and J. O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2), 1986.

G. Boente and R. Fraiman. Kernel-based functional principal components. *Statistics and Probability Letters*, 48:335–345, 2000.

D. Bosq. *Linear Processes in Function Spaces - Theory and Applications*. Springer, 2000.

O. Bouhaddou. Principal component analysis and interpolation of stochastic processes: methods and simulation. *Journal of Applied Statistics*, 14(3):251–267, 1987.

I. Braud, C. Obled, and A. Phamdinhtuan. Empirical Orthogonal Function (EOF) analysis of spatial random fields: Theory, accuracy of the numerical approximations and sampling effects. *Stochastic Hydrology and Hydraulics*, 7:146–160, 1993.

H. Cardot. Conditional Functional Principal Component Analysis. *Scandinavian Journal of Statistics*, 34:317–335, 2006.

P. E. Castro, W. H. Lawton, and E. A. Sylvestre. Principal modes of variation for processes with continuous sample curves. *Technometrics*, 28:329–337, November 1986. ISSN 0040-1706.

J. M. Chambers. *Software for Data Analysis: Programming with R*. Statistics and Computing. Springer, 2nd edition, 2009.

J.-M. Chiou and P.-L. Li. Functional clustering and identifying substructures of longitudinal data. *Journal of the royal statistical society Series B*, 69:679–699, 2007.

M. Csörgö and P. Révész. *Strong Approximations in Probability and Statistics*. Academic Press, 1981.

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136 – 154, 1982.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Chapman & Hall, 1996.

F. Ferraty and P. Vieu. Functional nonparametric statistics: a double infinite dimensional framework. In Akritas, M. G. and Politis, D. N., editor, *Recent advances and trends in nonparametric statistics*, pages 61–76. Elsevier Science BV, 2003.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice.* Springer, 2006.

G. Fischer. *Lineare Algebra: Eine Einführung für Studienanfänger.* Vieweg, 13th edition, 2002.

P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society Series B*, 68(1):109–126, 2006.

P. Hall, H.-G. Müller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517, 2006.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, 2001.

H. Heuser. *Lehrbuch der Analysis Teil 2.* Teubner, 12th edition, 2002.

H. Heuser. *Funktionalanalysis - Theorie und Anwendung.* Teubner, 4th edition, 2006.

I. M. Johnstone and A. Yu Lu. On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

I.T. Jolliffe. *Principal Component Analysis.* Springer, 2nd edition, 2004.

A. Kneip and K. J. Utikal. Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association*, 96(454), 2001.

F. Leisch. *The flexclust Package; R package description*, 2007.

D. J. Levitin, R. L. Nuzzo, B. W. Vines, and J. O. Ramsay. Introduction to Functional Data Analysis. *Canadian Psychology*, 48(3):135–155, 2007.

H.-G. Müller. *Nonparametric Regression Analysis of Longitudinal Data. Lecture Notes in Statistics 46.* Springer, 1988.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *The Annals of Statistics*, 33:774–805, 2005.

H.-G. Müller, U. Stadtmüller, and F. Yao. Functional Variance Processes. *Journal of the American Statistical Association*, 101:1007–1018, 2006.

A. Obhukov. The statistically orthogonal expansion of empirical functions. *American Geophysical Union*, pages 288–291, 1960.

F. A. Ocana, A. M. Aguilera, and M. Escabias. Computational considerations in functional principal component analysis. *Computational Statistics*, 22(3):449–465, 2007.

P. Pollard. *Convergence of stochastic processes.* Springer, 1984.

R. W. Preisendorfer and C. D. Mobley. *Principal Component Analysis in Meteorology and Oceanography.* Elsevier, 1988.

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

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer, 2nd edition, 2006.

J. O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB.* Springer, 2009.

J. A. Rice and B. W. Silverman. Estimating the Mean and Covariance Structure Nonparametrically when the Data are Curves. *Journal of the Royal Statistical Society Series B*, 53(1): 233–243, 1991.

B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24, 1996.

A. van der Linde. Variational Bayesian functional PCA. *Computational Statistics and Data Analysis*, 53:517–533, 2008.

I. Winzenborg. *The FPCA Package; R package description*, 2011.

I. Winzenborg, A. Geistanger, and U. Stadtmüller. Multivariate and functional clustering applied to measurements of laboratory analysis systems. In Pavese, F. et al., editor, *Advanced Mathematical and Computational Tools in Metrology and Testing VIII*, volume 78 of *Series on Advances in Mathematics for Applied Sciences*, pages 361–368. World scientific publ CO PTE LTD, 2009.

F. Yao. Functional principal component analysis for longitudinal and survival data. *Statistica Sinica*, 17(3):965–983, 2007.

F. Yao and T. C. M. Lee. Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society Series B*, 68:3–25, 2006.

F. Yao, H.-G. Müller, and J.-L. Wang. Shrinkage Estimation for Functional Principal Component Scores with Application to the Population Kinetics of Plasma Folate. *Biometrics*, 59: 676–685, 2003.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.