A Contribution to Functional Data Analysis

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von
Heiko Wagner
aus Euskirchen

Bonn
2016

Dekan: Prof. Dr. Daniel Zimmer, LL.M
Erstreferent: Prof. Dr. Alois Kneip
Zweitreferent: Prof. Dr. Michael Vogt

Tag der mündlichen Prüfung: 03.11.2016

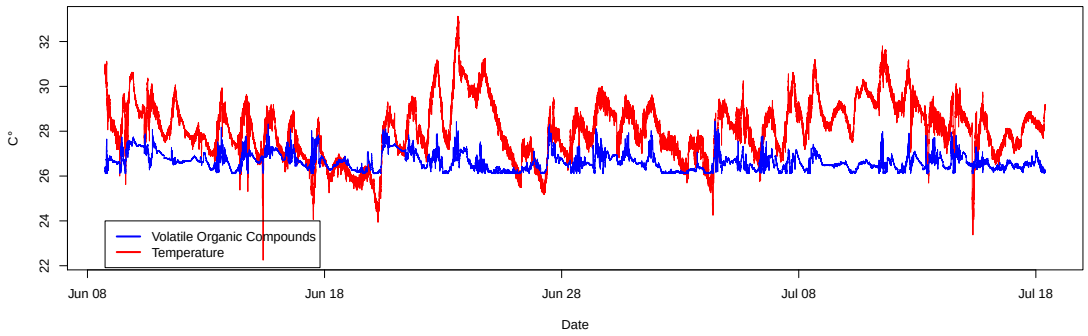
Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn (http://hss.ulb.uni-bonn.de/diss_online) elektronisch publiziert.

Acknowledgments

First, and foremost, I would like to thank my supervisor Alois Kneip. He was a great guidance to me and helps me to canalize my sometimes unstructured ideas and taught me how to write clearly structured papers. I also want to thank James O. Ramsay and his family for my wonderful stay in Ottawa. I am also very grateful for his guidance how to write good papers and how to look at statistical problems from a different angle. I would also like to thank Michael Vogt who accepted on short term to act as my second supervisor, Wolfgang Härdle and Maria Grith to give me the opportunity to travel to Berlin several times to work on the paper from *Chapter 2*, John Aston to invite me to Cambridge and having interesting discussions about the registration problem. Not to forget all my colleagues from the statistic department in Bonn and in particular my officemates Dominik Poß and Daniel Becker who helped <sarcasm>to keep the office temperature high and the air quality low (see Figure 0-1)</sarcasm> and always give great support to scientific questions.

The financial support by the Deutsche Forschungsgemeinschaft (DFG) through DFG Sachbeihilfe KN 567/3-1 and BE 5550/1-1 is greatly acknowledged.

Figure 0-1: Office temperature and air quality measured between June and July 2016 using a Raspberry Pi.



Contents

List of Figures

List of Tables

Chapter 1

Introduction

Subject of this dissertation is the low dimensional representation of random functions. There is a close connection to traditional statistics where usually the properties of a random variable or a multivariate random variable is studied. During the last three decades Functional data analysis (FDA) becomes popular to carry out an statistical analysis of functions as objects of interest, see for example Ramsay and Silverman (2005) or Ferraty and Vieu (2006). In this context let $(\Omega, \mathcal{F}, P)$ be a probability space and $L^2(\Omega)$ be the space of all random variables $\mathbf{X} : \Omega \to L^2(\mathcal{I})$, where $L^2(\mathcal{I})$ is the space of square integrable functions on a compact interval $\mathcal{I}$. Accordingly then for $X \in L^2(\Omega)$ the functional analogy of mean and covariance is given by $E(X(u)) = \int_\Omega X(w,u)dP(w),\ u \in \mathcal{I}$ and $E(X(u)X(s)) = \int_\Omega X(w,u)X(w,s)dP(w),\ u,s \in \mathcal{I}$ the covariance function is defined by $K : \mathcal{I} \times \mathcal{I} \to \mathbb{R},\ K(u,v) = cov(X(u),X(v)) = E(X(u)X(s)) - E(X(u))E(X(s))$. Sloppy speaking functions can thus be considered as "highly multivariate objects". However, considering random functions rises new questions which have no analogues in the analysis of multivariate random variables because a major difference between the two approaches is that for functions observation has to be considered as ordered in $\mathcal{I}$.

One major aspect of statistical data analysis is to reduce the dimensionality of the data to make interpretation easier. Oftentimes this provides an opportunity for further analysis, for example: clustering or prediction. A popular tool for multivariate data is Principal Component Analysis (PCA) invented at the beginning of the last century by Pearson (1901) and independently by Hotelling (1933). PCA is a method to structure multivariate data sets, to do so the n-dimensional data space is projected into a q-dimensional subspace but loosing as little information as possible. Among others Dauxois et al. (1982) developed a functional counterpart to PCA (FPCA) based on an eigendecomposition of the covariance operator $(\Gamma\gamma)(u) = \int_{\mathcal{I}} K(u,v)\gamma(v)dv$.

Usually $L^2(\Omega)$ is not directly observable but only a discrete i.i.d sample $i = 1,\dots,n, j = 1,\dots,T$ often contaminated with noise ϵ_{ij} such that $Y_i(t_j) = X_i(t_j) + \epsilon_{ij}$ is observed. Functional Data of such type often occurs in reality for example considering Weather data, Stock prices or even statistical objects like density functions. In a functional context dimension reduction means finding a low dimensional functional subspace. First attempts here by using FPCA to derive estimates if a sample of discrete noisy data is observed where for example made by Besse and Ramsay (1986). The important theoretical framework was then carried out by Hall and Hosseini-Nasab (2006). Using FPCA as low dimensional representation has the advantages that orthogonal components are obtained and these components span the best low dimensional subspace in terms of the L^2 error function. In *Chapter 2* a detailed introduction to FPCA and suitable estimators are given for the case where $\mathcal{I} = [0,1]^g, g \in \mathbb{N}$.

The ordering in $\mathcal{I}$ rises new chances and problems which have no analogues in the analysis of multivariate random variables. For example smoothing techniques can be applied to handle the ϵ_{ij} term. In *Chapter 2* a local polynomial is used, prominent alternatives are for example using splines as done by Rice and

Silverman (1991) or kernel smoothing as done by Benko et al. (2009). Further the so called registration problem where the observed functions are additionally warped by some strictly increasing function h_i such that $Y_i(t_j) = X_i(h_i(t_j)) + \epsilon_{ij}$ has no comparable counterpart in the analysis of multivariate data. In *Chapter 3* we illustrate that due to warping structurally very different curves can share the same covariance and thus the corresponding eigenfunctions are identically. But in presence of warping using traditional covariance based methods like FPCA or even easier representations like the mean are not meaningful anymore. In general it is hard to model h_i using linear basis expansions because by construction such expansions mostly explain amplitude variation.

FPCA can be considered as an "all-rounder" suitable to analyze most data sets because it always gives the best low dimensional subspace representation. In various specific cases where additional knowledge of the structure of the data set is present or specific questions are posed it is often better to stick to another decomposition which not necessary fulfills the best basis property but is tailored for the corresponding data set. For example if a sample of density functions $f_1, \ldots, f_n$ is considered, then in general the low dimensional approximation of f_i using FPCA as done by Kneip and Utikal (2001) is not a density function anymore. Petersen and Müller (2016) presents a different low dimensional representation not based on a linear basis decomposition of f_i but of $\Psi \circ f_i$ were Ψ is suitably chosen. The choice of Ψ then ensures that the low dimensional approximation of $\Psi \circ f_i$ "reversed" by Ψ^{-1} is still a density. Another decomposition different from FPCA is given by the Principal Differential Analysis (PDA) by Ramsay (1996) which is based on the linear differential operator $L = D^m - \sum_{j=0}^{m-1} w_j D^j$. Here the task is to estimate $w_j \in L^2(\mathcal{I})$ minimizing $n^{-1} \sum_{i=1}^{n} (LY_i)^2(t) = n^{-1} \sum_{i=1}^{n} \{\sum_{j=0}^{m} w_j(t) D^j Y_i(t)\}^2$ instead of estimating eigenfunctions of the covariance operator. A possible advantage of this decomposition is of course given if the curves are of simple differential equation

nature but also a general decomposition can benefit since this approach makes explicit use of smoothness properties due to the derivatives.

In this context two alternative ways to present functional data not using traditional FPCA are presented. In *Chapter 2* for observed discrete noisy $Y_1, \ldots, Y_n$ the aim is to estimate a low dimensional decomposition for derivatives $X_i^{(d)},\ d \in \mathbb{N}_+^g$ of high dimensional spacial curves $\mathcal{I} = [0,1]^g$. The reason not to stick to traditional FPCA in this case is that the usual estimators which rely on some kind of smoothing suffer from the curse of dimensionality. Thus a different estimator closely related to classical FPCA is used where the curse of dimensionality has an lesser impact. Therefore better rates of convergence are obtained and the presented method usually gives better estimates. *Chapter 2* is joined work with Maria Grith, Wolfgang K. Härdle and Alois Kneip and is planned to be submitted to "Statistica Sinica". In *Chapter 3* the registration problem is discussed, in particular registration deals with separating amplitude and phase variation. While traditional registration procedures usually register the observed curves to a single template, oftentimes some kind of structural mean, a method to register the curves to a finite dimensional linear function space is presented. The curves are then decomposed in this finite dimensional space. It turns out that a sample of curves can always be registered to a finite dimensional space if the curves have a special structure such that the number of extrema per curve is finite. We use the term "curves of bounded shape variation" to classify these curves. Assuming a sample of "curves of bounded shape variation" seems to be a very natural condition in many applications in biomedicine, technics, chemometrics, etc. *Chapter 3* is joined work with Alois Kneip and has been submitted to "Journal of the Royal Statistical Society: Series B".

The presented methods were applied to various real data sets from different scientific fields. In *Chapter 2* an empirical study is carried out where the state

price density (SPD) surfaces from call option prices is estimated. Three main components were identified, which can be interpreted as volatility, skewness and kurtosis factors. Also effects introduced by the term structure variation could be identified. *Chapter 3* provides applications to human growth curves, genetic data and the Aneurisk65 data-set. Using registration we were able to get a better data classification and may discover patterns unobservable before. The juggling data-set discussed in *Chapter 4* as well as the Aneurisk65 data-set are 3D-curves $\mathbb{R}^3 \to \mathbb{R}$. A method to register and to analyze these kind of data is presented. *Chapter 4* is joint work with Dominik Poss and was an implication of the CTW: "Statistics of Time Warpings and Phase Variations" at the Ohio State University and has been published in the "Electronic Journal of Statistics" (Poss and Wagner (2014)).

Chapter 2

Functional Principal Component Analysis for Derivatives of High-Dimensional Spatial Curves

Abstract

We present two approaches based on the functional principal component analysis (FPCA) to estimate smooth derivatives of noisy and discretely observed high-dimensional spatial curves. One method is based on the eigenvalue decomposition of the covariance operator of the derivatives and the other assumes the operator of the curves. To handle observed data, both approaches rely on local polynomial regressions. We analyze the requirements under which the methods are asymptotically equivalent, and establish that the first approach requires very strong smoothness assumptions to achieve similar convergence rates to the second one. If the curves are contained in a finite-dimensional function space, we show that using both our methods provides better rates of convergence than estimating the curves individually. We illustrate the methodology in a simulation and empirical study, in which we estimate state price density (SPD) surfaces from call option prices. We identify three main components, which can be interpreted as volatility, skewness and convexity factors. We also find effects introduced by the term structure variation.

2.1 Introduction

Over the last two decades, functional data analysis became a popular tool to handle data entities that are random functions. Usually, discrete and noisy versions of them are observed. Oftentimes, these entities are high-dimensional spatial objects. Examples include brain activity recordings generated during fMRI or EEG experiments, e.g., Majer et al. (2015). In a variety of applications though the object of interest is not directly observable but it is a function of the observed data. Typical examples in the financial applications include functionals that can be retrieved from the observed prices by means of derivatives, such as implied state price density, e.g., Grith et al. (2012), pricing kernel, e.g., Grith et al. (2013) or the market price of risk, e.g., Härdle and Lopez-Cabrera (2012). Motivated by such data analysis situations, we address the problem of estimating high-dimensional spatial curves that are not empirically observable but can be recovered from the existing discrete and noisy data by means of derivatives.

Functions, which are objects of an infinite-dimensional vector space, require specific methods that allow a good approximation of their variability with a small number of components. FPCA is a convenient tool to address this task because it allows us to explain complicated data structures with only a few orthogonal principal components that fulfill the optimal basis property in terms of its L^2 accuracy. These components are given by the Karhunen-Loève theorem, see for instance Bosq (2000). In addition, the corresponding principal loadings to this basis system can be used to study the variability of the observed phenomena. An important contribution in the treatment of the finite dimensional PCA was done by Dauxois et al. (1982), followed by subsequent studies that fostered the applicability of the method to samples of observed noisy curves. Besse and Ramsay (1986), among others, derived theoretical results for observations that are affected by additive errors. Some of the most important contributions for the extension of the PCA

to functional data belong to Cardot et al. (1999), Cardot et al. (2007), Ferraty and Vieu (2006), Mas (2002) and Mas (2008). To date, simple, one-dimensional spatial curves are well understood from both numerical and theoretical perspective. In one-dimensional case FPCA is also easy to implement. High-dimensional objects, with more complicated spatial and temporal correlation structures, or not-directly observable functions of these objects, such as derivatives, lack a sound theoretical framework. Furthermore, the computational issues are not negligible in high-dimensions.

To our best knowledge, FPCA for derivatives has been tackled by Hall et al. (2009) and Liu and Müller (2009). The first study handles one-dimensional directional derivatives and gradients. The second paper analyses a particular setup in one-dimension where the observations are sparse. This method can be applied to non-sparse data but may be computationally inefficient when dealing with large amounts of observations per curve. There are no studies of derivatives using FPCA in more than one spatial dimension. For the study of observed functions, there are a series of applied papers for the two-dimensional case, see Cont and da Fonseca (2002) for an application close to our empirical study. Other complicated attempts to implement FPCA when the object of interest are the observed functions, rather than their derivatives, have been done in more than two dimensions, in particular in the area of brain imaging. For example, Zipunnikov et al. (2011) split the recorded data into smaller parts to make it manageable. This method, called multilevel FPCA, developed through previous studies, see Staicu and Carroll (2010), Di et al. (2009), is well suited to analyze different groups of individuals. However, a thorough derivation of the statistical properties of the estimators is missing in these papers.

In this paper, we aim to fill in the existent gaps in the literature on FPCA for the study of derivatives of functions in high-dimensional space. We present

two alternative approaches to obtain the derivatives. The paper is organized as follows: the theoretical framework, estimation procedure and statistical properties are derived through Section 2.2. Our empirical study in Section 2.3 is guided by the estimation and the dynamics analysis of the option implied state price densities. It includes a simulation study and a real data example.

2.2 Methodology

2.2.1 Two approaches to the derivatives of high-dimensional functions using FPCA

The representation of derivatives of high-dimensional spatial curves requires a careful choice of notation. In this section, we review the FPCA from a technical point of view and make the reader familiar with our notations.

Let X be a centered smooth random function in $L^2([0,1]^g)$, where g denotes the spatial dimension, with finite second moment $\int_{[0,1]^g} \mathsf{E}\left[X(t)^2\right] dt < \infty$ for $t = (t_1, \ldots, t_g)^\top$. The underlying dependence structure can be characterized by the covariance function $\sigma(t,v) \stackrel{\text{def}}{=} \mathsf{E}\left[X(t)X(v)\right]$ and the corresponding covariance operator Γ

$$(\Gamma\vartheta)(t) = \int_{[0,1]^g} \sigma(t,v)\vartheta(v)dv.$$

Mercer's lemma guarantees the existence of a set of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$ and a corresponding system of orthonormal eigenfunctions $\gamma_1, \gamma_2, \ldots$ called functional principal components s.t.

$$\sigma(t,v) = \sum_{r=1}^{\infty} \lambda_r \gamma_r(t)\gamma_r(v), \tag{2.1}$$

where the eigenvalues and eigenfunctions satisfy $(\Gamma\gamma_r)(t) = \lambda_r\gamma_r(t)$. Moreover,

$\sum_{r=1}^{\infty} \lambda_r = \int_{[0,1]^g} \sigma(t,t)dt$. The Karhunen-Loève decomposition for the random function X gives

$$X(t) = \sum_{r=1}^{\infty} \delta_r \gamma_r(t), \tag{2.2}$$

where the loadings δ_r are random variables defined as $\delta_r = \int_{[0,1]^g} X(t)\gamma_r(t)dt$ that satisfy $\mathsf{E}[\delta_r^2] = \lambda_r$, as well as $\mathsf{E}[\delta_r \delta_s] = 0$ for $r \neq s$. Throughout the paper the following notation for the derivatives of a function X will be used

$$X^{(d)}(t) \stackrel{\text{def}}{=} \frac{\partial^d}{\partial t^d} X(t) = \frac{\partial^{d_1}}{\partial t_1^{d_1}} \cdots \frac{\partial^{d_g}}{\partial t_g^{d_g}} X(t_1, \ldots, t_g), \tag{2.3}$$

for $d = (d_1, ..., d_g)^\top$ and $d_j \in \mathbb{N}$ the partial derivative in the spatial direction $j = 1, \ldots, g$. We denote $|d| = \sum_{j=1}^{g} |d_j|$ and require that X is at least $|d|+1$ times continuously differentiable.

Building on equations (2.1) and (2.2), we consider two approaches to model a decomposition for derivatives $X^{(d)}$. The first one is stated in terms of the Karhunen-Loève decomposition applied to their covariance function. We define $\sigma^{(d)}(t,v) \stackrel{\text{def}}{=} \mathsf{E}\left[X^{(d)}(t)X^{(d)}(v)\right]$ and $\lambda_1^{(d)} \geq \lambda_2^{(d)} \geq \ldots$ be the corresponding eigenvalues. The principal components $\varphi_r^{(d)}$ are solutions to

$$\int_{[0,1]^g} \sigma^{(d)}(t,v)\varphi_r^{(d)}(v)dv = \lambda_r^{(d)} \varphi_r^{(d)}(t). \tag{2.4}$$

For nonderivatives, i.e., $|d| = 0$, we introduce the following notation $\varphi_r^{(0)}(t) \equiv \gamma_r(t)$. Similarly to (2.2), the decomposition of $X^{(d)}$ with principal components $\varphi_r^{(d)}(t)$ is

$$X^{(d)}(t) = \sum_{r=1}^{\infty} \delta_r^{(d)} \varphi_r^{(d)}(t), \tag{2.5}$$

for $\delta_r^{(d)} = \int_{[0,1]^g} X^{(d)}(t)\varphi_r^{(d)}(t)dt$.

A different way to think about a decomposition for derivatives, is to take the

derivatives of the functional principal components in (2.2)

$$X^{(d)}(t) = \sum_{r=1}^{\infty} \delta_r \gamma_r^{(d)}(t), \tag{2.6}$$

where the d-th derivative of the r-th eigenfunction is the solution to

$$\int_{[0,1]^g} \frac{\partial^d}{\partial v^d} \left(\sigma(t,v)\gamma_r(v)\right) dv = \lambda_r \gamma_r^{(d)}(t). \tag{2.7}$$

In general, for $|d| > 0$ it holds that $\varphi_r^{(d)}(t) \neq \gamma_r^{(d)}(t)$, but both basis systems span the same function space. In particular, there always exists a projection with $a_{rp} = \left\langle \gamma_p^{(d)}, \varphi_r^{(d)} \right\rangle = \int_{[0,1]^g} \gamma_p^{(d)}(t)\varphi_r^{(d)}(t)dt$ such that $\sum_{r=1}^{\infty} a_{rp}\varphi_r^{(d)}(t) = \gamma_p^{(d)}(t)$. However, if we consider a truncation of (2.2) after a finite number of components this is no longer true in general. An advantage of using $\varphi_r^{(d)}$ instead of $\gamma_r^{(d)}$ is that the decomposition of covariance function of the derivatives gives orthonormal basis that fulfill the best basis property, such that for any fixed $L \in \mathbb{N}$ and every other orthonormal basis system φ'

$$E||X^{(d)} - \sum_{r=1}^{L} \left\langle X^{(d)}, \varphi_r^{(d)} \right\rangle \varphi_r^{(d)}|| \leq E||X^{(d)} - \sum_{r=1}^{L} \left\langle X^{(d)}, \varphi_r' \right\rangle \varphi_r'||. \tag{2.8}$$

This guarantees that by using $\varphi_r^{(d)},\ r = 1, \ldots, L$ we always achieve the best L dimensional subset selection in terms of the L^2 error function. In the next section we show that estimating the basis functions with such desirable features, for nonzero derivatives, comes at the cost of inferior rate of convergence. However, if the true underlying structure lies in a L-dimensional function space, which is equivalent to a factor model setup, the advantage of deriving the best L-orthogonal basis vanishes, because it is possible to derive a basis system with the same features using span$(\gamma^{(d)})$. This is achieved by performing a spectral decomposition of the

finite-dimensional function space of $\gamma_r^{(d)},\ r = 1, \ldots, L$ to get an orthonormal basis system fulfilling (2.8).

2.2.2 Sample inference

Let $X_1, \ldots, X_N \in L^2([0,1]^g)$ be a sample of i.i.d. realizations of the smooth random function X. The empirical approximation for the covariance function based on the N curves is given by the sample counterpart

$$\hat{\sigma}^{(d)}(t,v) = \frac{1}{N}\sum_{i=1}^{N} X_i^{(d)}(t) X_i^{(d)}(v) \tag{2.9}$$

and for the covariance operator by

$$\hat{\Gamma}_N^{(d)} \hat{\varphi}_r^{(d)}(t) = \int_{[0,1]^g} \hat{\sigma}^{(d)}(t,v)\hat{\varphi}_r^{(d)}(v)dv, \tag{2.10}$$

where the eigenfunction $\hat{\varphi}_r^{(d)}$ corresponds to the r-th eigenvalue of $\hat{\Gamma}_N^{(d)}$. For inference, it holds that $||\varphi_r^{(\nu)} - \hat{\varphi}_r^{(\nu)}|| = \mathcal{O}_p(N^{-1/2})$ and $|\lambda_r^{(\nu)} - \hat{\lambda}_r^{(\nu)}| = \mathcal{O}_p(N^{-1/2})$, see for instance Dauxois et al. (1982) or Hall and Hosseini-Nasab (2006). The loadings corresponding to each realization X_i can be estimated via the empirical eigenfunctions as $\hat{\delta}_{ri}^{(d)} = \int_{[0,1]^g} X_i^{(d)}(t)\hat{\varphi}_r^{(d)}(t)dt$.

2.2.3 The model

In most applications, the curves are only observed at discrete points and data is corrupted by additive noise. To model these aspects, we assume that each curve in the sample is observed at independent randomly-distributed points $t_i = (t_{i1}, \ldots, t_{iT_i})^\top,\ t_{ik} \in [0,1]^g,\ k = 1, \ldots, T_i, i = 1, \ldots, N$ from a continuous distribu-

tion with density f such that $\inf_{t\in[0,1]^g} f(t) > 0$. Our model is then given by

$$Y_i(t_{ik}) = X_i(t_{ik}) + \varepsilon_{ik} = \sum_{r=1}^{\infty} \delta_{ri}\gamma_r(t_{ik}) + \varepsilon_{ik}, \tag{2.11}$$

where ε_{ik} are i.i.d. random variables with $\mathsf{E}[\varepsilon_{ik}] = 0$, $\mathrm{Var}(\varepsilon_{ik}) = \sigma^2_{i\varepsilon}$ and ε_{ik} is independent of X_i.

2.2.4 Estimation procedure

Dual method

An alternative to the Karhunen-Loève decomposition relies on the duality relation between the row and column space. The method was first used in a functional context by Kneip and Utikal (2001) to estimate density functions and later adapted by Benko et al. (2009) for general functions. Let $\nu = (\nu_1, \ldots, \nu_g)^\top$, $\nu_j \in \mathbb{N}$, $|\nu| < \rho \leq m$ and $M^{(\nu)}$ be the dual matrix of $\hat{\sigma}^{(\nu)}(t, v)$ from (2.9) consisting of entries

$$M^{(\nu)}_{ij} = \int_{[0,1]^g} X^{(\nu)}_i(t) X^{(\nu)}_j(t) dt. \tag{2.12}$$

Let $l^{(\nu)}_r$ be the eigenvalues of matrix $M^{(\nu)}$ and $p^{(\nu)}_r = (p^{(\nu)}_{1r}, \ldots, p^{(\nu)}_{Nr})$ be the corresponding eigenvectors. For $\nu = d$, the estimators for the quantities in the right-hand side of equations (2.4) and (2.5) are given by

$$\hat{\varphi}^{(d)}_r(t) = \frac{1}{\sqrt{l^{(d)}_r}} \sum_{i=1}^{N} p^{(d)}_{ir} X^{(d)}_i(t) \ , \ \hat{\lambda}^{(d)}_r = \frac{l^{(d)}_r}{N} \text{ and } \hat{\delta}^{(d)}_{ri} = \sqrt{l^{(d)}_r} p^{(d)}_{ir}. \tag{2.13}$$

Important for the representation given in equation (2.6) are the eigenvalues and eigenvectors of $M^{(0)}$ denoted by $l_r \overset{\text{def}}{=} l^{(0)}_r, p_r \overset{\text{def}}{=} p^{(0)}_r$ and the corresponding or-

thonormal basis $\hat{\gamma}_r \stackrel{\text{def}}{=} \hat{\varphi}_r^{(0)}$ and loadings $\hat{\delta}_{ri} \stackrel{\text{def}}{=} \hat{\delta}_{ri}^{(0)}$. It is straightforward to derive

$$\hat{\gamma}_r^{(d)}(t) = \frac{1}{\sqrt{l_r}} \sum_{i=1}^{N} p_{ir} X_i^{(d)}(t). \tag{2.14}$$

Quadratic integrated regression

Before deriving estimators of $M^{(0)}$ and $M^{(d)}$ using the model from Section 2.2.3, we outline some results needed to construct these estimators. For any vectors $a, b \in \mathbb{R}^g$ and $c \in \mathbb{N}^g$, we define $|a| \stackrel{\text{def}}{=} \sum_{j=1}^{g} |a_j|$, $a^{-1} \stackrel{\text{def}}{=} (a_1^{-1}, \ldots, a_g^{-1})^\top$, $a^b \stackrel{\text{def}}{=} a_1^{b_1} \times \cdots \times a_g^{b_g}$, $a \circ b \stackrel{\text{def}}{=} (a_1 b_1, \ldots, a_g b_g)^\top$ and $c! \stackrel{\text{def}}{=} c_1! \times \cdots \times c_g!$.

Consider a curve Y observed at points t_l, $l = 1, \ldots, T$ generated as in equation (2.11). Let $k = (k_1, \ldots, k_g)^\top$, $k_l \in \mathbb{N}$ and consider a multivariate local polynomial estimator $\hat{\beta}(t) \in \mathbb{R}^\rho$ that solves

$$\min_{\beta(t)} \sum_{l=1}^{T} \left[Y(t_l) - \sum_{0 \le |k| \le \rho} \beta_k(t)(t_l - t)^k \right]^2 K_B(t_l - t). \tag{2.15}$$

K_B is any non-negative, symmetric and bounded multivariate kernel function and B a $g \times g$ bandwidth matrix. For simplicity, we assume that B has main diagonal entries $b = (b_1, \ldots, b_g)^\top$ and zero elsewhere.

As noted by Fan et al. (1997) the solution of the minimization problem (2.15) can also be represented using a weight function W_ν^T, see Appendix 2.5.2, such that

$$\hat{X}_b^{(\nu)}(t) = \nu! \hat{\beta}_\nu(t) = \nu! \sum_{l=1}^{T} W_\nu^T \left((t_l - t) \circ b^{-1} \right) Y(t_l). \tag{2.16}$$

Local polynomial regression estimators are better suited to estimate integrals like (2.12) than other kernel estimators, e.g., Nadaraya-Watson or Gasser-Müller estimator, since the bias and variance are of the same order of magnitude near the

boundary as well as in the interior, see for instance Fan and Gijbels (1992). We propose the following estimator for the squared integrated functions $\int_{[0,1]^g} X^{(\nu)}(t)^2 dt$

$$\theta_{\nu,\rho} = \int_{[0,1]^g} \nu!^2 \sum_{k=1}^{T}\sum_{l=1}^{T} W_\nu^T\left((t_k - t)\circ b^{-1}\right) W_\nu^T\left((t_l - t)\circ b^{-1}\right) Y(t_l)Y(t_k)dt \\ -\nu!^2\hat{\sigma}_\varepsilon^2 \int_{[0,1]^g} \sum_{k=1}^{T} W_\nu^T\left((t_k - t)\circ b^{-1}\right)^2 dt. \tag{2.17}$$

where $\hat{\sigma}_\varepsilon^2$ is a consistent estimator of σ_ε^2. The second term is introduced to cancel the bias in $\mathsf{E}\left[Y^2(t_k)\right] = X(t_k)^2 + \sigma_\varepsilon^2$.

Lemma 2.2.1 *Under Assumptions 1- 4, X is $m \geq 2|\nu|$ times continuously differentiable, the local polynomial regression is of order ρ with $|\nu| \leq \rho < m$ and $|\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2| = \mathcal{O}_P(T^{-1/2})$. Then as $T \to \infty$ and $\max(b)^{\rho+1}b^{-\nu} \to 0$, $\frac{\log(T)}{Tb_1\times\cdots\times b_g} \to 0$ as $Tb_1 \times \cdots \times b_g b^{4\nu} \to \infty$, then conditioned on $t_1, \ldots t_T$*

$$\begin{aligned} &\mathsf{E}\left[\theta_{\nu,\rho}\right] - \int_{[0,1]^g} X^{(\nu)}(t)^2 dt = \mathcal{O}_p\left(\max(b)^{\rho+1}b^{-\nu} + \frac{1}{T^{3/2}(b^{2\nu}b_1 \times \cdots \times b_g)}\right) \\ &\mathrm{Var}(\theta_{\nu,\rho}) = \mathcal{O}_p\left(\frac{1}{T^2 b_1 \times \cdots \times b_g b^{4\nu}} + \frac{1}{T}\right), \end{aligned} \tag{2.18}$$

where the expectation and variance denote the conditional operators with respect to the observations Y. The proof of Lemma 2.2.1 is given in Appendix 2.5.2.

Estimation of $M^{(0)}$ and $M^{(d)}$

The curves Y_i in equation (2.11) are assumed to be observed at different random points. For uniformly sampled points $t_1, \ldots, t_T \in [0,1]^g$ with $T = \min_{i\in 1,\ldots,N} T_i$, we

replace the integrals in (2.17) with the Riemann sums, such that

$$\hat{M}_{ij}^{(\nu)} = \begin{cases} \nu!^2 \sum_{k=1}^{T_i} \sum_{l=1}^{T_j} w_\nu^T(t_{ik}, t_{jl}, b) Y_j(t_{jl}) Y_i(t_{ik}) & \text{if } i \neq j \\ \nu!^2 \left(\sum_{k=1}^{T_i} \sum_{l=1}^{T_i} w_\nu^T(t_{ik}, t_{il}, b) Y_i(t_{il}) Y_i(t_{ik}) - \hat{\sigma}_{i\varepsilon}^2 \sum_{k=1}^{T_i} w_\nu^T(t_{ik}, t_{ik}, b) \right) & \text{if } i = j. \end{cases}$$

where $w_\nu^T(t_{ik}, t_{jl}, b) := T^{-1} \sum_{m=1}^{T} W_\nu^T\left((t_{ik} - t_m) \circ b^{-1}\right) W_\nu^T\left((t_{jl} - t_m) \circ b^{-1}\right)$. The estimator for $M^{(0)}$ is given by setting $\nu = (0, \ldots, 0)^\top$ and the estimator for $M^{(d)}$ by $\nu = d$.

There are two possible sources of error in the construction of the estimator $\hat{M}^{(\nu)}$. One is coming from smoothing noisy curves at a common grid, and has been analyzed in Lemma (2.2.1). The other one is from approximating the integral in (2.17) by a sum, see equation above. In Appendix (2.5.3) we show that the error of the integral approximation is of order $T^{-1/2}$.

Proposition 2.2.2 *Under the requirements of Lemma 2.2.1*

$$|M_{ij}^{(\nu)} - \hat{M}_{ij}^{(\nu)}| = \mathcal{O}_P\left(\max(b)^{\rho+1} b^{-d} + \left(\frac{1}{T^2 b_1 \times \cdots \times b_g b^{4d}} + \frac{1}{T} \right)^{1/2} \right).$$

By Proposition 2.2.2 estimating $M^{(d)}$ gives an asymptotic higher bias and also a higher variance than estimating $M^{(0)}$. This effect becomes more pronounced in high-dimensions. However, by using local polynomial regression with large ρ one can still get parametric rates within each method.

Remark 2.2.3 *Under the assumptions of Lemma 2.2.1 and using Proposition 2.2.2 we can derive estimators for $M^{(\nu)}$, which attain parametric rates. If $m > \rho \geq \frac{g}{2} - 1 + 3\sum_{l=1}^{g} \nu_l$, $b = T^{-\alpha}$ with $\frac{1}{2(\rho+1-\sum_{l=1}^{g}\nu_l)} \leq \alpha \leq \frac{1}{g+4\sum_{l=1}^{g}\nu_l}$ then $|M_{ij}^{(\nu)} - \hat{M}_{ij}^{(\nu)}| = \mathcal{O}_P(1/\sqrt{T})$.*

We can see that the orders of polynomial expansion and the bandwidths for estimating $M^{(\nu)}$ will differ for $\nu = (0, \ldots, 0)^\top$ and $\nu = d$. In particular, the estimator

of $M^{(d)}$ requires higher smoothness assumptions - via $m > \rho$ - and a higher bandwidth to achieve the same parametric convergence rate as the estimator for $M^{(0)}$.

In Lemma 2.2.1 it is required that $|\sigma_{i\varepsilon}^2 - \hat{\sigma}_{i\varepsilon}^2| = \mathcal{O}_p(T^{-1/2})$, which ensures parametric rates of convergence for $\hat{M}^{(\nu)}$ under the conditions of Remark 2.2.3. By Assumption 2, in the univariate case, a simple class of estimators for $\sigma_{i\varepsilon}^2$, which achieve the desired convergence rate, are given by successive differentiation, see von Neumann et al. (1941) and Rice (1984). However, as pointed out in Munk et al. (2005), difference estimators are no longer consistent for $g \geq 4$ due to the curse of dimensionality. A possible solution is to generalize the kernel based variance estimator proposed by Hall and Marron (1990) to higher dimensions with

$$\hat{\sigma}_{i\varepsilon}^2 = \frac{1}{v_i} \sum_{l=1}^{T_i} \left(Y_i(t_{il}) - \sum_{k=1}^{T_i} w_{ilk} Y(t_{ik}) \right)^2, \tag{2.19}$$

where $w_{ilk} = K_{r,H}(t_{il} - t_{ik}) / \sum_{k=1}^{T_i} K_{r,H}(t_{il} - t_{ik})$ and $v_i = T_i - 2\sum_l w_{ilk} + \sum_{l,k} w_{ilk}^2$ and $K_{r,H}$ is a g-dimensional product kernel of order r with bandwidth matrix H. Munk et al. (2005) show that if $4r > g$ and if the elements of the diagonal matrix H are of order $\mathcal{O}(T^{-2/(4r+g)})$ then the estimator $\hat{\sigma}_{\varepsilon i}$ in equation (2.19) achieves parametric rates of convergence.

Note that if the curves are observed at a common random grid with $T = T_i = T_j,\ i, j = 1, \ldots, N$, a simple estimator for $M^{(0)}$ is constructed by replacing the integrals with Riemann sums in (2.12). This estimator is given by

$$\tilde{M}_{ij}^{(0)} = \begin{cases} \frac{1}{T} \sum_{l=1}^{T} Y_i(t_l) Y_j(t_l) \text{ if } i \neq j \\ \frac{1}{T} \sum_{k=1}^{T} Y_i(t_l)^2 - \hat{\sigma}_{i\varepsilon}^2 \text{ if } i = j \end{cases}. \tag{2.20}$$

In Appendix (2.5.3) we verify that the convergence rate of $\tilde{M}_{ij}^{(0)}$ does not depend on g.

When working with more than one spatial dimension, in practice data is often recorded using an equidistant grid with T points in each direction. For our approach, this strategy will not improve the convergence rate of $\tilde{M}^{(0)}$ due to the curse of dimensionality. If one can influence how data is recorded, we recommend using a common random grid, which keeps computing time and the storage space for data to a minimum and still gives parametric convergence rates for the estimator of $M_{ij}^{(0)}$. If $T \gg N$ equation (2.20), gives a straightforward explanation why the dual matrix is preferable to derive the eigendecomposition of the covariance operator, because taking sums has a computational cost that is linear.

Estimating the basis functions

We keep notations $\nu = d$ to refer to the specification in equation (2.5) and $\nu = (0, \ldots, 0)^\top$ to (2.6). A spectral decomposition of $\hat{M}^{(\nu)}$ is applied to obtain the eigenvalues $\hat{l}_r^{(\nu)}$ and eigenvectors $\hat{p}_r^{(\nu)}$ for $r, j = 1, \ldots, N$. This gives straightforward empirical counterparts $\hat{\lambda}_{r,T}^{(\nu)} = \hat{l}_r^{(\nu)}/N$ and $\hat{\delta}_{rj,T}^{(\nu)} = \sqrt{\hat{l}_r^{(\nu)}}\hat{p}_{rj}^{(\nu)}$.

To estimate $\varphi_r^{(d)}$ and $\gamma_r^{(d)}$, a suitable estimator for $X_i^{(d)}$, $r, j = 1, \ldots, N$ is needed. Given a set of T observations $Y = \{Y(t_1), \ldots, Y(t_T)\}$ of variable X, we use a local polynomial kernel estimator, denoted $\hat{X}_{i,h}^{(d)}$, similarly to (2.16), with a polynomial of order p and bandwidth vector $h = (h_1, \ldots, h_g)$. Here, h is not equal to b, the bandwidth used to smooth the entries of the $\hat{M}^{(0)}$ and $\hat{M}^{(d)}$ matrix. In fact, we show below that the optimal order for the bandwidth vector h differs asymptotically from that of b derived in the previous section. An advantage of using local polynomial estimators, compared for example to spline or wavelet estimators, is that the bias and variance can be derived analytically. For the univariate case these results can be found in Fan and Gijbels (1996) and for the multivariate case in Masry (1996) and Gu et al. (2015). We summarize them in terms of order

of convergence below conditioned on $t_{1j}, \dots t_{T_j j}$

$$\begin{aligned} \mathsf{E}\left[X_j^{(d)}(t) - \hat{X}_{j,h}^{(d)}(t)\right] &= \mathcal{O}_p(\max(h)^{p+1}h^{-d}) \\ \operatorname{Var}\left(\hat{X}_{j,h}^{(d)}(t)\right) &= \mathcal{O}_p\left(\frac{1}{Th_1 \times \cdots \times h_g h^{2d}}\right). \end{aligned} \tag{2.21}$$

Let $\max(h)^{p+1}h^{-d} \to 0$ and $\left(\max(h)^{p+1}Th^{-d}\right)^{-1} \to 0$ as $T \to \infty$. If p is chosen such that $p - |d|$ is odd then

$$\begin{aligned} \mathsf{E}\left[\frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N} p_{ir}^{(\nu)}\left(X_i^{(d)}(t) - \hat{X}_{i,h}^{(d)}(t)\right)\right] =& \frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{j=1}^{N} p_{jr}^{(\nu)} \operatorname{Bias}\left(\hat{X}_{j,h}^{(d)}(t)\right) + o_p\left(\max(h)^{p+1}h^{-d}\right) \\ =& \mathcal{O}_p(\max(h)^{p+1}h^{-d}) \end{aligned}$$

$$\begin{aligned} \operatorname{Var}\left(\frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N} p_{ir}^{(\nu)}\hat{X}_{i,h}^{(d)}(t)\right) =& \frac{1}{l_r^{(\nu)}}\sum_{j=1}^{N}\left(p_{jr}^{(\nu)}\right)^2 \operatorname{Var}\left(\hat{X}_{j,h}^{(d)}(t)\right) + o_p\left(\frac{1}{NTh_1 \times \cdots \times h_g h^{2d}}\right) \\ =& \mathcal{O}_p\left(\frac{1}{NTh_1 \times \cdots \times h_g h^{2d}}\right). \end{aligned}$$

We show that under certain assumptions the asymptotic mean squared error of $\hat{\varphi}_{r,T}^{(d)}$ and $\hat{\gamma}_{r,T}^{(d)}$ is dominated by the two terms.

Proposition 2.2.4 *Under the requirements of Lemma 2.2.1, Assumptions 6 and 7, Remark 2.2.3, and for* $\inf_{s \neq r}|\lambda_r - \lambda_s| > 0$, $r, s = 1, \dots, N$ *and* $\max(h)^{p+1}h^{-d} \to 0$ *with* $NTh_1 \dots h_g h^{2d} \to \infty$ *as* $T, N \to \infty$ *we obtain*

a) $|\gamma_r^{(d)}(t) - \hat{\gamma}_{r,T}^{(d)}(t)| = \mathcal{O}_p\left(\max(h)^{p+1}h^{-d}\right) + \mathcal{O}_p\left((NTh_1 \times \cdots \times h_g h^{2d})^{-1/2}\right)$

b) $|\hat{\varphi}_r^{(d)}(t) - \hat{\varphi}_{r,T}^{(d)}(t)| = \mathcal{O}_p\left(\max(h)^{p+1}h^{-d}\right) + \mathcal{O}_p\left((NTh_1 \times \cdots \times h_g h^{2d})^{-1/2}\right)$

A proof of Proposition 2.2.4 is provided in Appendix 2.5.4. As a consequence, the resulting global optimal bandwidth is given by $h_{r,opt} = \mathcal{O}_p\left((NT)^{-1/(g+2p+2)}\right)$

for both basis and all $r = 1, \ldots, N$. Even if the optimal bandwidth for both approaches is of the same order of magnitude, the values of the actual bandwidths may differ. A simple rule of thumb for the choice of bandwidths in practice is given in Section 2.3.1.

2.2.5 Properties under a factor model structure

Often, the variability of the functional curves can be expressed with only a few basis functions modeled by a truncation of (2.2) after L basis functions. If a true factor model is assumed, the basis representation to reconstruct $X^{(d)}$ is arbitrary, in sense that

$$X^{(d)}(t) = \sum_{r=1}^{L} \delta_r \gamma_r^{(d)}(t) = \sum_{r=1}^{L_d} \delta_r^{(d)} \varphi_r^{(d)}(t). \tag{2.22}$$

Here L is always an upper bound for L_d. The reason for this is that by taking derivatives it is possible that $\gamma_r^{(d)}(t) = 0$ or that there exits some $a_r \in \mathbb{R}^{L-1}$ such that $\gamma_r^{(d)}(t) = \sum_{s \neq r} a_{sr} \gamma_s^{(d)}(t)$.

Based on the methodology described in Section 2.2.4, the two estimators for derivatives are given by

$$\hat{X}^{(d)}_{i,FPCA_1}(t) \stackrel{\text{def}}{=} \sum_{r=1}^{L} \hat{\delta}_{ir,T} \hat{\gamma}^{(d)}_{r,T}(t) \;\approx\; \hat{X}^{(d)}_{i,FPCA_2}(t) \stackrel{\text{def}}{=} \sum_{r=1}^{L_d} \hat{\delta}^{(d)}_{ir,T} \hat{\varphi}^{(d)}_{r,T}(t). \tag{2.23}$$

Proposition 2.2.5 *Assume that a factor model with L factors holds for X. For $NT^{-1} \to 0$, together with the requirements of Proposition 2.2.4, the true curves can be reconstructed*

a) $|X_i^{(d)}(t) - \hat{X}^{(d)}_{i,FPCA_1}(t)| = \mathcal{O}_p\left(T^{-1/2} + \max(h)^{p+1} h^{-d} + (NTh_1 \times \cdots \times h_g h^{2d})^{-1/2}\right)$

b) $|X_i^{(d)}(t) - \hat{X}^{(d)}_{i,FPCA_2}(t)| = \mathcal{O}_p\left(T^{-1/2} + \max(h)^{p+1} h^{-d} + (NTh_1 \times \cdots \times h_g h^{2d})^{-1/2}\right)$.

A proof of Proposition (2.2.5) is given in Appendix (2.5.5). Compared with the

convergence rates of the individual curves estimators, see (2.21), the variance of our estimators reduces not only in T but also in N. Equations (2.13) and (2.14) can be interpreted as an average over N curves for only a finite number of L components. The intuition behind it is that only those components are truncated that are related to the error term and thus a more accurate fit is possible. If N increases at a certain rate, it is possible to get close to parametric rates. Such rates are not possible when smoothing the curves individually.

For the estimation of $\hat{X}^{(d)}_{i,FPCA_2}$, as illustrated in Remark 2.2.3, additional assumptions on the smoothness of the curves are needed to achieve the same rates of convergence for the estimators $\hat{M}^{(d)}$ and $\hat{M}^{(0)}$. With raising g and $d_j,\ j = 1, \ldots, g$ it is required that the true curves become much smoother which makes the applicability of estimating $\hat{X}^{(d)}_{i,FPCA_2}$ limited for certain applications. In contrast, the estimation of $M^{(0)}$ still gives almost parametric rates if less smooth curves are assumed. In addition, if the sample size is small, using a high degree polynomial needed to estimate $M^{(d)}$ might lead to unreliable results. To learn more about these issues, we check the performance of both approaches in a simulation study in Section 2.3.2 using different sample sizes.

2.3 Application to state price densities implied from option prices

In this section we analyse the state price densities (SPDs) implied by the stock index option prices. As state dependent contingent claims, options contain information about the risk factors driving the underlying asset price process and give information about expectations and risk patterns on the market. Mathematically, SPDs are equivalent martingale measures for the stock index and their existence is guaranteed in the absence of arbitrage plus some technical condi-

tions. In mathematical-finance terminology they are known as risk neutral densities (RNDs). A very restrictive model, with log-normal marginals for the asset price, is the Black-Scholes model. This model results in log-normal SPDs that correspond to a constant implied volatility surface across strikes and maturity. This feature is inconsistent with the empirically documented volatility smile or skew and the term structure. Therefore, richer specifications for the option dynamics have to be used. Most of earlier works adopt a static viewpoint; they estimate curves separately at different moments in time, see the methodology reviews by Bahra (1997), Jackwerth (1999) and Bliss and Panigirtzoglou (2002). In order to exploit the information content from all data available, it is reasonable to consider them as collection of curves.

The relation between the SPDs and the European call prices has been demonstrated by Breeden and Litzenberger (1987) and Banz (1978) for a continuum of strike prices spanning the possible range of future realizations of the underlying asset. For a fixed maturity, the SPD is proportional to the second derivative of the European call options with respect to the strike price. In this case, SPDs are one-dimensional functions. A two-dimensional point of view can be adopted if maturities are taken as an additional argument and the SPDs are viewed as a family of curves.

Let $C : \mathbb{R}^2_{\geq 0} \to \mathbb{R}$ denote the price function of a European call option with strike price k and maturity τ such that

$$C(k,\tau) = \exp\left(-r_\tau \tau\right) \int_0^\infty (s_\tau - k)^+ q(s_\tau, \tau)\, ds_\tau, \tag{2.24}$$

where r_τ is the annualized risk free interest rate for maturity τ, s_τ the unknown price of the underlying asset at maturity, k the strike price and q the state price

density of s_τ. One can show that

$$q(s_\tau, \tau) = \exp(r_\tau \tau) \left. \frac{\partial^2 C(k,\tau)}{\partial k^2} \right|_{k=s_\tau}. \tag{2.25}$$

Let s_0 be the asset price at the moment of pricing and assume it to be fixed. Then by the no-arbitrage condition, the forward price for maturity τ is

$$F_\tau = \int_0^\infty s_\tau q(s_\tau, \tau) ds_\tau = s_0 \exp(r_\tau \tau). \tag{2.26}$$

Suppose that the call price is homogeneous of degree one in the strike price. Then

$$C(k, \tau) = F_\tau C(k/F_\tau, \tau). \tag{2.27}$$

If we denote $m = k/F_\tau$ the moneyness, it is easy to verify that

$$\frac{\partial^2 C(k,\tau)}{\partial k^2} = \frac{1}{F_\tau} \frac{\partial^2 C(m,\tau)}{\partial m^2}. \tag{2.28}$$

Then one can show that for $d = (2, 0)$, $C^{(d)}(m, \tau)|_{m=s_\tau/F_\tau} = q(s_\tau/s_0, \tau) = s_0 q(s_\tau, \tau)$. In practice, it is preferable to work with densities of returns instead of prices when analyzing them jointly because prices are not stationary. Also, notice that call price curves are not centered. This imply that equations (2.4) and (2.6) will include an additional additive term, which refers to the population mean. We show in the next section how to handle this in practice.

In the application, X will refer to the rescaled call price $C(m, \tau)$. Therein, we also assume that the index $i = 1, \ldots, N$ refers to ordered time-points.

2.3.1 Implementation

Centering the observed curves

Throughout the paper it is assumed that the curves are centered. To insure this assumption, we subtract the empirical mean $\bar{X}^{(\nu)}(t_k) = \frac{1}{N}\sum_{i=1}^{N} \hat{X}_{i,b}^{(\nu)}(t_k)$ from the the observed call prices to obtained centered curves. A centered version $\overline{M}^{(\nu)}$, $\nu = (0,d)$ is given by

$$\overline{M}_{ij}^{(\nu)} = \hat{M}_{ij}^{(\nu)} - \frac{1}{T}\sum_{k=1}^{T}\left(\bar{X}^{(\nu)}(t_k)\hat{X}_{i,b}^{(\nu)}(t_k) + \bar{X}^{(\nu)}(t_k)\hat{X}_{j,b}^{(\nu)}(t_k) - \bar{X}^{(\nu)}(t_k)^2\right). \quad (2.29)$$

There is still space for improvement when centering of the curves. One possibility is to use a different bandwidth to compute the mean because averaging will necessarily lower the variance. Secondly, by the arguments of Section 2.2.4, the $\frac{1}{T}\sum_{k=1}^{T}\bar{X}^{(\nu)}(t_k)^2$ term can be improved accordingly to Lemma 2.2.1 by subtracting $\hat{\sigma}_\varepsilon$ weighted by suitable parameters. We decide to omit these fine tunings in our application because it would involve a huge amount of additional computational effort in contrast to only minor improvements in the results.

Bandwidth selection

To get parametric rates of convergence for $\hat{M}^{(d)}$ related to Remark 2.2.3 we choose $\rho = 7$ and b has to lie between $\mathcal{O}(T^{-1/10})$ and $\mathcal{O}(T^{-1/12})$. The choice of b to estimate $\hat{M}^{(0)}$ is similar, with the difference that $\rho > 0$, we choose $\rho = 1$ and b has to lie between $\mathcal{O}(T^{-1/3})$ and $\mathcal{O}(T^{-1/5})$. We use a very easy criteria to choose the bandwidth because by Proposition 2.2.4 the dominating error depends mainly on the choice of h. Let $t_{ik} = (t_{ik1}, \dots t_{ikg})$, then the bandwidth for direction j is determined by $b_j = ((\max_k(t_{ikj}) - \min_k(t_{ikj}))T_i)^\alpha$. When estimating state price densities $t_{ik} = (\tau_{ik}, m_{ik})$ and T_i is replaced by the cardinality of $\tau_i = \{\tau_{i1}, \dots \tau_{iT_i}\}$

and m_i respectively. In the estimation of $\hat{M}^{(d)}$ we set $\alpha = -1/10$ and $\alpha = -1/3$ for $\hat{M}^{(0)}$.

The choice of bandwidths h is a crucial parameter for the quality of the estimators. To derive an estimator for the bandwidths first note that in the univariate case ($g = 1$) the theoretical optimal univariate asymptotic bandwidth for the r-th basis is given by

$$h_{r,opt}^{d,\nu} = C_{d,p}(K) \left[T^{-1} \frac{\int_0^1 \sum_{i=1}^N (p_{ir}^{(\nu)})^2 \sigma_{\varepsilon i}^2(t) f_i(t)^{-1} dt}{\int_0^1 \left\{ \sum_{i=1}^N p_{ir}^{(\nu)} X_i^{(p+1)}(t) \right\}^2 dt} \right]^{1/(2p+3)}, \tag{2.30}$$

$$C_{d,p}(K) = \left[\frac{(p+1)!^2 (2d+1) \int K_{p,d_j}^{*2}(t) dt}{2(p+1-d)\{\int u^{p+1} K_{d,p}^*(t) dt\}^2} \right]^{1/(2p+3)}.$$

Like in the conventional local polynomial smoothing case $C_{d,p}(K)$ does not depend on the curves and is an easily computable constant. It only depends on the chosen kernel, the order of the derivative and the order of the polynomial, see for instance Fan and Gijbels (1996).

For our bandwidth estimator we treat every dimension separately, as if we have to choose an optimal bandwidth for derivatives in the univariate case, and correct for the asymptotic order, see Section 2.2.4. In practice, we can not use equation (2.30) to determine the optimal bandwidth because some variables are unknown and only discrete points are observed. As a rule-of-thumb, we replace these unknown variables using approximations. Estimates of $p_{ir}^{(0)}$ from $\hat{M}^{(0)}$ and $p_{ir}^{(d)}$ from $\hat{M}^{(d)}$ are further used. With these approximations, a feasible rule for computing the optimal bandwidth in direction j for the r-th basis function h_{jr} is

given by

$$h_{jr,rot}^{d,\nu} = \left(T^{-1} \frac{C_{d,p}^{2p+3} \hat{\sigma}_\varepsilon^2}{f_j \int_0^1 \left\{ \sum_{i=1}^N \hat{p}_{ir}^{(\nu)} \tilde{X}_i^{(p+1)}(t_j) \right\}^2 dt_j} \right)^{1/(g+2p+2)} . \tag{2.31}$$

In our application as well as our simulation we have $g = 2$, $d = (0, 2)$ and do a third order local polynomial regression. The integrals are approximated by Riemann sums.

- The density of the observed points is approximated by a uniform distribution with $f_1 = \max_{i,j}(\tau_{ij}) - \min_{i,j}(\tau_{ij})$ and $f_2 = \max_{i,j}(m_{ij}) - \min_{i,j}(m_{ij})$.

- To get a rough estimator for $X_i^{(p+1)}$ based on X_i, we use a polynomial regression. For our application, we take $p = 3$ and are thus interested in estimates for $X_i^{(4)}(m)$ and $X_i^{(4)}(\tau)$. We expect the curves to be more complex in the moneyness direction than in the maturity direction and we adjust the degree of the polynomials to reflect this issue. The estimates are then given by

$$\begin{aligned} a_i^* &= \arg\min_{a_i} \left(X_i(m,\tau) - a_{i0} + \sum_{l=1}^{5} a_{il} m^l + \sum_{l=6}^{9} a_{il} \tau^{(l-5)} \right) \\ \tilde{X}_i^{(4)}(m) &= 24 a_{i4}^* + 120 a_{i5}^* m \\ \tilde{X}_i^{(4)}(\tau) &= 24 a_{i9}^*. \end{aligned} \tag{2.32}$$

- To estimate the variance for each curve we use the kernel approach given in (2.19) using a Epanechnikov kernel with a bandwidth of $T^{-2/(4+g)}$ for each spatial direction. These estimates are used as well to correct for the diagonal bias when $\hat{M}^{(0)}$ and $\hat{M}^{(d)}$ are estimated. In (2.31) the average over all $\hat{\sigma}_{i\epsilon}$ is used.

For technical reasons, we use the product of Gaussian kernel functions to con-

struct local polynomial estimators. The reason for is that, as we can verify from Proposition 2.2.4, the optimal bandwidth h will decrease in N. By using a global bandwidth and a compact kernel the matrix given in equation (2.43) may become singular when N is large and T is small.

In our simulation and application we use the mean optimal $h_{i,rot}^{d,\nu} = L^{-1}\sum_{r=1}^{L} h_{ir,rot}^{d,\nu}$ for each $\hat{\gamma}_r^{(d)}, \hat{\varphi}_r^{(d)}$ to save computation time. Since we had to demean the sample in (2.29), finally we add $N^{-1}\sum_{i=1}^{N} \hat{X}^{(d)}_{i,h_{i,rot}^{d,\nu}}$ to the resulting truncated decomposition to derive the final estimate.

Estimation of the number of components

In Section 2.2.5 we assumed that the number of components is given. In general, the number of basis functions needed is unknown a priori. For the case $|d| = 0$ there exists a wide range of criteria that can be adapted to our case to determine the upper bound L. The easiest way to determine the number of components is by choosing the model accuracy by an amount of variance explained by the eigenvalues. In (2.69) we show that under the conditions from Proposition 2.2.4 $\hat{\lambda}_r^{(d)} - \hat{\lambda}_{r,T}^{(d)} = \mathcal{O}_p(N^{-1/2}T^{-1/2} + T^{-1})$ and $\lambda_r^{(d)} - \hat{\lambda}_r^{(d)} = \mathcal{O}_p(N^{-1/2})$. The assumptions in Corollary 1 from Bai and Ng (2002) can be adapted to our case and give several criteria for finding L or L_d by generalizing Mallows (1973) C_p criteria for panel data settings. These criteria imply minimizing the sum of squared residuals when k factors are estimated and penalizing the overfitting. One such formulation suggests choosing the number of factors using the criteria

$$PC^{(\nu)}(k^*) = \min_{k\in\mathbb{N}, k\leq L_{\max}} \left[\left(\sum_{r=k+1}^{N} \hat{\lambda}_r^{(\nu)} \right) + k \left(\sum_{r=L^{\max}}^{N} \hat{\lambda}_r^{(\nu)} \right) \left(\frac{\log(C_{NT}^2)}{C_{NT}^2} \right) \right], \quad (2.33)$$

for the constant $C_{NT} = \min(\sqrt{N}, \sqrt{T})$ and a prespecified $L^{\max} < \min(N, T)$. Bai and Ng (2002) propose an information criteria that do not depend on the choice

of $L_{\max}$. We consider the above modified version

$$IC^{(\nu)}(k^*) = \min_{k\in\mathbb{N}, k\leq L} \left[\log\left(\frac{1}{N} \sum_{r=k+1}^{N} \hat{\lambda}_r^{(\nu)} \right) + k \left(\frac{\log(C_{NT}^2)}{C_{NT}^2} \right) \right]. \tag{2.34}$$

Here using $\nu = (0, \ldots, 0)^\top$ will give L while using $\nu = d$ will give the factors L_d.

Another possibility for the choice of number of components is to compute the variance explained by each nonorthogonal basis by

$$\mathrm{Var}(\hat{\delta}_{r,T}^{(d)} \hat{\gamma}_{r,T}^{(d)}) = \langle \hat{\gamma}_{r,T}^{(d)}, \hat{\gamma}_{r,T}^{(d)} \rangle \hat{\lambda}_r, \tag{2.35}$$

sort them in decreasing order and use equations (2.33) or (2.34) to select the number of components.

2.3.2 Simulation Study

We investigate the finite sample behavior of our estimators in a simulation study, which is guided by the real data application in Section 2.3.3. Simulated SPDs are modeled as mixtures of G components, $q(m,\tau) = \sum_{l=1}^{G} w_l q^l(m,\tau)$, where q^l are fixed basis functions and w_l are random weights. For fixed τ we consider $q^l(\cdot,\tau)$ to be a log-normal density functions, with mean $\left(\mu_l - \frac{1}{2}\sigma_l^2\right)\tau$ and variance $\sigma_l^2\tau$, and simulate weights w_{il} with $\sum_{l=1}^{G} w_{il} = 1$, where $i = 1, \ldots, N$ is the index for the day, then

$$q_i(m,\tau) = \sum_{l=1}^{G} w_{il} \frac{1}{m\sqrt{2\pi\sigma_l^2\tau}} \exp\left[-\frac{1}{2} \left\{ \frac{\log(m) - \left(\mu_l - \frac{1}{2}\sigma_l^2\right)\tau}{\sigma_l\sqrt{\tau}} \right\}^2 \right]. \tag{2.36}$$

Following Brigo and Mercurio (2002) the prices of call options for these SPDs are

$$C_i(m,\tau) = \exp\left(-r_{i\tau}\tau\right) \sum_{l=1}^{G} w_{il} \left\{\exp(\mu_l \tau)\Phi(y_1) - m\Phi(y_2)\right\} \tag{2.37}$$

where $y_1 = \frac{\log\left(m^{-1}\right)+\left(\mu_l+\frac{1}{2}\sigma_l^2\right)\tau}{\sigma_l\sqrt{\tau}}$, $y_2 = \frac{\log\left(m^{-1}\right)+\left(\mu_l-\frac{1}{2}\sigma_l^2\right)\tau}{\sigma_l\sqrt{\tau}}$ and Φ is the standard normal cdf. This representation corresponds to a factor model in which the mixture components are densities associated with a particular state of nature and the loadings are equivalent with probabilities of states.

We illustrate the finite sample behavior for $G = 3$ with $\mu_1 = 0.4$, $\mu_2 = 0.7$, $\mu_3 = 0.1$, and $\sigma_1 = 0.5$, $\sigma_2 = 0.3$, $\sigma_3 = 0.3$. The loadings are simulated from the positive half-standard normal distribution, then standardized to sum up to one. One can verify that the correlation matrix for the loadings is

$$R = \begin{bmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{bmatrix},$$

which is singular with $\text{rank}(R) = 2$. As a result, the covariance operator of the SPD curves has $L = G-1$ nonzero eingenvalues. This implies that in this example, using a mixture of 3 factors only 2 principal components are necessary to explain the variance in the true curves.

Without loss of generality, we set $r_{i\tau} = 0$, for each day $i = i, \ldots, N$. We construct a random grid for each observed curve X_i by simulating points $t_{ik} = (m_{ik}, \tau_{ik})$, $k = 1, \ldots, T$ from a uniform distribution with continuous support $[0.5, 1.8] \times [0.2, 0.7]$. Finally, we record noisy discrete observations of the call functions with the additive error term i.i.d. $\varepsilon_{ik} \sim \text{N}(0, 0.1^2)$.

The true SPDs given by equation (2.36) are used to verify the performance of $\hat{X}^{(d)}_{FPCA_1}$, $\hat{X}^{(d)}_{FPCA_2}$ and of the individually estimated curves $\hat{X}^{(d)}_{Indiv.}$, in terms of

	T	50				250			
N	$\hat{X}_\bullet^{(d)}$	Mean	Var	Med	IQR	Mean	Var	Med	IQR
10	$FPCA_1$	0.1876	0.0367	0.1300	0.1325	0.0780	0.0025	0.0643	0.0546
	$FPCA_2$	0.2238	0.1212	0.1295	0.1466	0.0762	0.0026	0.0630	0.0518
	Indiv.	0.2709	0.0900	0.1928	0.1838	0.1105	0.0054	0.0916	0.0708
25	$FPCA_1$	0.0917	0.0066	0.0680	0.0580	0.0404	0.0006	0.0336	0.0223
	$FPCA_2$	0.1553	0.0966	0.0878	0.0887	0.0586	0.0016	0.0489	0.0406
	Indiv.	0.2691	0.0995	0.1889	0.1848	0.1111	0.0052	0.0916	0.0719

Table 2.1: Results of the simulation described in Section 2.3.2 with different values for T and N. $FPCA_1$ and $FPCA_2$ are superior in sense of MSE over the individual estimation of the derivatives in each setting. $FPCA_1$ is better than $FPCA_2$ except for $N = 10$, $T = 250$. For $FPCA_1$ and $FPCA_2$ the estimation improves with raising N and T. These results support our asymptotic results given by Proposition 2.2.2 and 2.2.5.

mean integrated squared error (MSE), i.e., $T^{-1}\sum_{k=1}^{T}\left\{X^{(d)}(t_{ik}) - \hat{X}_\bullet^{(d)}(t_{ik})\right\}^2$, for $d = (2, 0)$. For evaluation we generate a common grid of 256 points from a uniform distribution. To derive the optimal bandwidth in each case we stick to the rule-of-thumb approach presented in Section 2.3.1. The bandwidth for the individually smoothed curve i is derived by replacing $\hat{p}_{ir}^{(\nu)}$ in (2.31) by one and zero otherwise. The performance is recorded for sample sizes N of 10 and 25 with T observations per day of size 50 and 250. This procedure is repeated 500 times to get reliable results, mean, variance and the inter quartile distance based at the MSE of the repetitions are given in Table 2.1.

Both FPCA based approaches give better estimates for the derivative of the call functions than an individually applied local polynomial estimator of the individual curves. Both the mean and the median of the MSE are smaller which is a result of the additional average over N for the basis functions as given by Proposition 2.2.5. However, the $FPCA_1$ method performs decisively better for small T than the other two both in terms of mean and standard deviation of the mean squared error. In addition $FPCA_1$ benefits more from increasing N than $FPCA_2$. With small T for $FPCA_2$ and individual smoothing the variability of MSE is much bigger than for $FPCA_1$ while the median of $FPCA_1$ and $FPCA_2$ are comparable. This means

individual smoothing and $FPCA_2$ must behave much worse than $FPCA_1$ in some instances while $FPCA_1$ was able to stabilize the estimates. To get the same effect using $FPCA_2$ a much bigger T is needed. A possible explanation for this behavior is given by Proposition 2.2.2. The rates of convergence for the estimators of the dual matrix entries rely on T. Thus in finite sample, when T is small, the estimated loadings might be biased.

2.3.3 Real Data Example

Data description

We use daily settlement European call option prices written on the underlying DAX 30 stock index. The sample spans the period between January 2, 2002 and December 3, 2011 and includes a total of 2557 days. The option prices are computed at the end of the trading day by EUREX based on the recorded intraday transaction prices. The expiration dates for the options are set on every third Friday of a month. Therefore, only option prices with a few maturities are available on a particular day, see Figure 2-1. The distance between two consecutive maturities is increasing with the maturity, while the distance between two consecutive strikes for the settlement option prices is relatively constant. This data structure, with only a few available maturities daily, still allow the use of local polynomial method for smoothing in our application because the estimates in the maturity direction can be interpreted as weighted averages of the neighboring estimates for fixed observed maturities. This is similar to interpolation that is often used in practice for option prices. We include call options with maturity between one day and one year. Our sample contains prices of options with an average of six maturities and sixty-five strikes per day.

We assume 'sticky' coordinates for the daily observations, see equation (2.27),

and standardize both the strike and the call prices within one day by the forward stock index value to ensure that the observation points are in the same range. We then apply the estimation methodology to the rescaled call prices observed at discrete moneyness and maturity points and report the decomposition results for their second derivative with respect to moneyness. Our proxy for the risk-free interest rates are the EURIBOR rates, which are listed daily for several maturities. We perform a linear interpolation to calculate the rate values at desired maturities.

In our subsequent analysis we use the VDAX index computed by the Deutsche Börse AG from the prices of call and put options. It reflects the market expectation for the 30 day ahead square root of implied variance for the DAX log-returns under the risk neutral measure, which is then annualized.

r, $L_{\max}$	1	2	3	4	5	6	7	8	9	10
$\hat{\lambda}_{r,T} \times 10^6$	133.29	18.90	2.69	1.62	0.49	0.34	0.26	0.09	0.08	0.05
$\hat{\lambda}_{r,T}/\hat{\lambda}_{r+1,T}$	7.05	7.01	1.66	3.28	1.44	1.31	2.83	1.18	1.70	1.35
$k^*(PC^{(0)})$	-	-	-	-	-	-	7	8	9	9
$k^*(IC^{(0)})$	-	-	-	-	-	-	7	-	-	-

Table 2.2: Estimated eigenvalues and eigenvalue ratios. Number of factors by $PC^{(0)}$ criterion

Estimation results

The first eigenvalue of the dual covariance matrix $\hat{M}^0$ for the call option surfaces has a dominant explanatory power. The order of magnitude of the following eigenvalues decreases by a factor of ten for every few additional components. To detect the relative contribution of consecutive components, we construct the ratio of two adjacent estimated eigenvalues in descending order, see Ahn and Horenstein (2013). The first two terms are dominating the sequence and there are spikes at the fourth and seventh component ratio. $PC^{(0)}$ criterion suggests at least seven

components, see values of k^* for $L_{max} \geq 7$ in Table 2.2. $IC^{(0)}$ criterion, which does not depend on the truncation parameter L_{max}, suggests seven components.

We assess the quality and importance of estimated components by looking first at equation (2.26), which expresses the pricing rule under the risk neutral measure Q. After rearranging we obtain that $\mathsf{E}_i^Q\left(s_{i+\tau}/F_i\right) = \int_0^\infty m q_i(m,\tau)dm = 1$ and $\mathrm{Var}_i^Q\left(s_{i+\tau}/F_i\right) = \int_0^\infty m^2 q_i(m,\tau)dm$, where $F_{i\tau} = s_i \exp(r_{i\tau}\tau)$. By equation (2.6)

$$1 = \int_0^\infty m\bar{q}(m,\tau)dm + \sum_{r=1}^{\infty} \delta_{ir} \int_0^\infty m\gamma_r^{(d)}(m,\tau)dm, \tag{2.38}$$

where $\bar{q}$ is the population mean. Similarly

$$\frac{\mathrm{Var}_i^Q\left(s_{i+\tau}/s_i\right)}{\left\{\exp(r_{i\tau}\tau)\right\}^2} - 1 = \int_0^\infty m^2\bar{q}(m,\tau)dm + \sum_{r=1}^{\infty} \delta_{ir} \int_0^\infty m^2\gamma_r^{(d)}(m,\tau)dm. \tag{2.39}$$

Equations (2.38) and (2.39) can be used to select those components used to minimize the difference between the left-hand side and a linear combination of their loadings, over the entire sample. We fit such linear regressions for different combinations of components and assess their goodness in a MSE sense for $\tau = 1$ month. For the left hand side of equation (2.39) we use the square of VDAX index in the numerator. While this index refers to the standard deviation of the log-returns under the risk neutral measure, it can still be used in the regression because the transformation $q(\log m, \tau) = mq(m,\tau)$ maintains the linear-relationship between the dependent variable and the loadings. The first eight loadings explain over 99% of the variance of the interest rate and respectively risk neutral variance. A higher adjusted R^2 is obtained by regressing VDAX on the loadings instead of VDAX^2. Among the regressors, $\hat{\delta}_{1,T}$, $\hat{\delta}_{2,T}$, $\hat{\delta}_{3,T}$, $\hat{\delta}_{7,T}$ explain most of the variation in the system of equations. In particular, $\hat{\delta}_{1,T}$ is highly correlated with VDAX (-94.90%) and VDAX^2 (-91.15%), and $\hat{\delta}_{3,T}$ with the one-month EURIBOR rate (-65.54%).

A closer look at the dynamics of the loadings $\hat{\delta}_{2,T}$ shows an unusual behavior

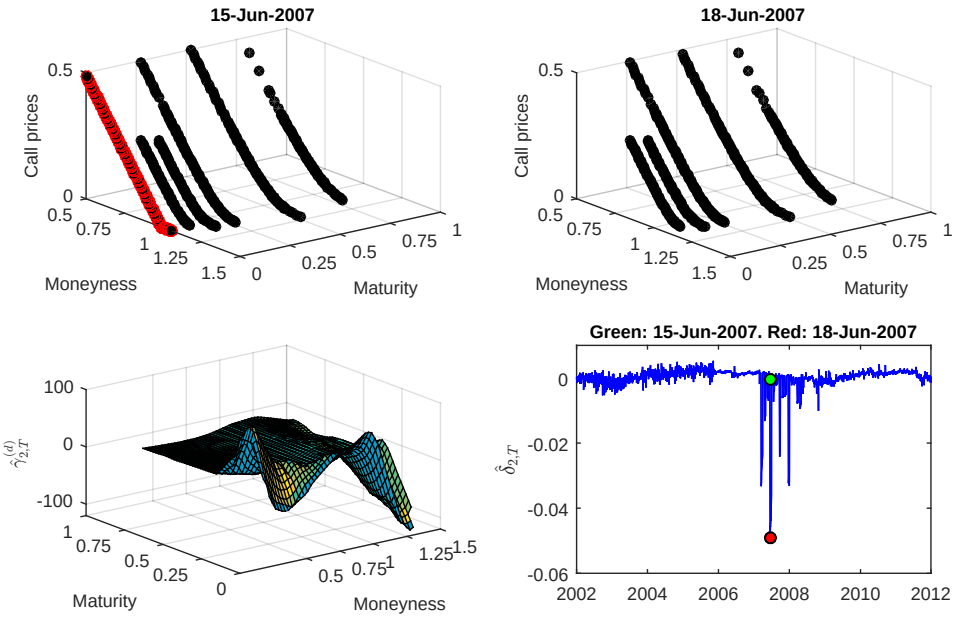


Figure 2-1: The effect of the expiration date on $\hat{\delta}_{2,T}$

between mid-February 2007 through mid-June 2008. This interval spans the period before the beginning of the financial crisis and extends to the end of the recession in the Euro Area - according to the Center for Economic and Policy Research (CEPR) recession indicator. The loadings are extremely volatile and display a certain time regularity of jumps. We identify these jumps with the Mondays following an expiration date (options expire at a monthly frequency, always on a Friday). Figure 2-1 highlights the dynamics of $\hat{\delta}_{2,T}$ on and following an expiration day. After roughly two weeks, they revert to a 'normal' level.

During this period, there are few observations available for the call prices with strikes larger than the current stock index price for small maturities. Together with the absence of a call string with close enough maturity on the following trading Monday, this introduces bias in the smooth estimated call surface, for grid values outside the range of observation points. However, we cannot rule out the

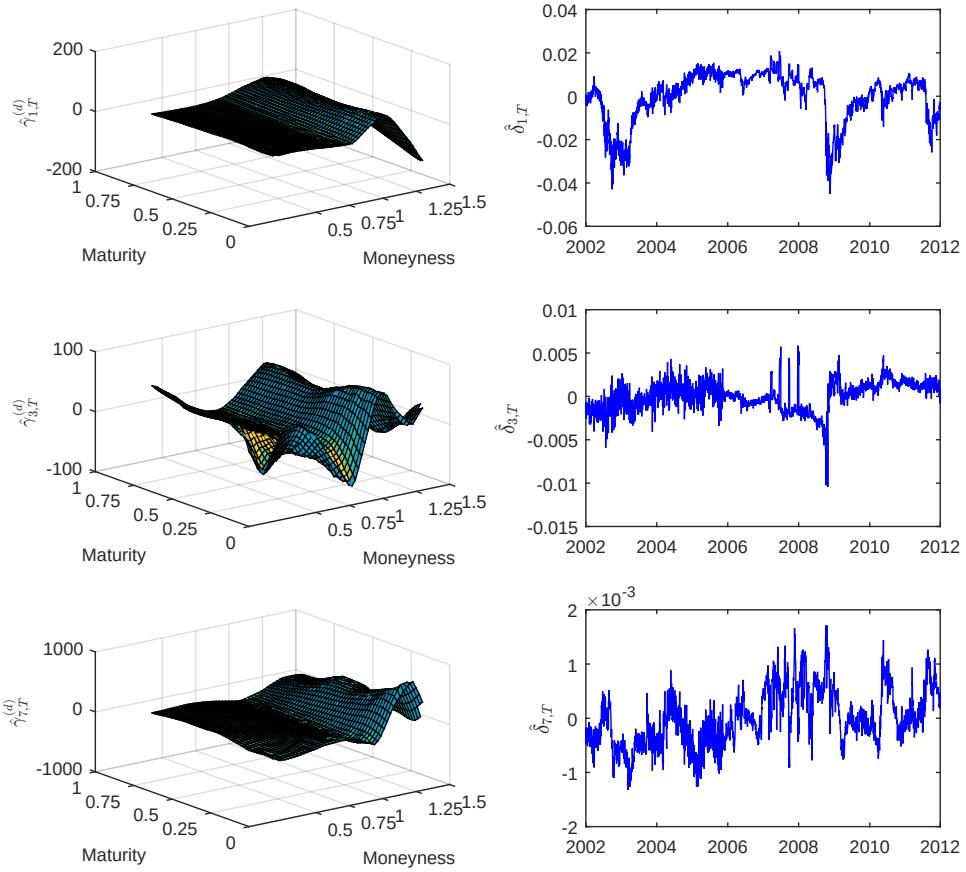


Figure 2-2: Estimated components $\hat{\gamma}_{1,T}^{(d)}$, $\hat{\gamma}_{3,T}^{(d)}$ and $\hat{\gamma}_{7,T}^{(d)}$ and their loadings obtained by the decomposition of the dual covariance matrix $\hat{M}^{(0)}$

possibility that the importance of the second component is not due to an error in pre-smoothing of the call options used for the estimation of $M^{(0)}$ because even if we recalculate the explained variance for all the components after excluding the estimated loadings from this time interval, this factor still remains the second most important. The shape of the second estimated component $\hat{\gamma}_{2,T}^{(d)}$, displayed in Figure 2-1, suggests that it is related to the short end of the SPD term structure effect.

The estimated components $\hat{\gamma}_{1,T}^{(d)}$, $\hat{\gamma}_{3,T}^{(d)}$ and $\hat{\gamma}_{7,T}^{(d)}$ together with their loadings are displayed in Figure 2-2, in order of their explained variance, see equation (2.35). These three components describe three types of asymmetry present in the dynamics of the SPDs. The first component, has a long tail on the left side of the peak, similarly to the mean SPD curve. It takes positive values around the peak and negative around the tails and is closely related to the volatility of the SPD dynamics. An increase in the loadings of this component decreases the volatility of SPD. $\hat{\gamma}_{3,T}^{(d)}$ has a relatively symmetric 'valley-hill' pattern, which shifts mass around the central region of the density. It also influences the density far left tail. A positive shock in the direction of this components increases the negative skewness, while a large enough negative shock will render the SPD positively skew. This component is interpreted as the sign of skewness factor. $\hat{\gamma}_{7,T}^{(d)}$ has a lean 'valley' at the left of the sample mean, which takes negative values, and a more pronounced 'hill' at the right, which feature positive values. This component emphasized the dynamics of negative skewness and induces changes in the kurtosis of the density as well. We interpret it as the negative skewness factor.

A negative skewness of the SPD reflects the market expectation that the future stock index will be above its forward value. Usually, the negative skew increases together with the implied volatility. While negative skewness risk can bear excess returns, during periods of economic downturn, the investors prefer positively skewed distributions. This can be seen when looking at the large negative values

of $\hat{\delta}_{3,T}$ which, in effect, shift the SPD mass from the positive to the negative side of the distribution, in conformity with an increase in the risk aversion of investors.

The functional principal components for the reduced model $\sum_{r\in\{1,3,7\}} \hat{\delta}_{r,T}\hat{\gamma}_{r,T}^{(d)}$ resemble closely the three components from Figure 2-2. Further analysis shows that if we add any of the term structure components, whose loading feature a behavior similar to $\hat{\delta}_{2,T}$, with their inherent jump before, the shape of the components changes slightly. In addition, the loadings of all orthogonalized components are 'contaminated' with jumps. In fact, all the loadings of the estimated components (not displayed here) by decomposing $\hat{M}^{(d)}$, for $d = (2, 0)$ display the jump-behavior we described before between mid-February 2007 and mid-September 2008. In that sense, previous approach seems to provide more accurate estimates that allow for a better interpretation of the results. The other components $\hat{\gamma}_{4,T}^{(d)}$, $\hat{\gamma}_{5,T}^{(d)}$, $\hat{\gamma}_{6,T}^{(d)}$ and $\hat{\gamma}_{8,T}^{(d)}$ have similar shape features to the four components discussed so far: $\hat{\gamma}_{1,T}^{(d)}$, $\hat{\gamma}_{2,T}^{(d)}$, $\hat{\gamma}_{3,T}^{(d)}$ and $\hat{\gamma}_{7,T}$. Their loadings have "jumps" alike $\hat{\delta}_{2,T}$. We contend that they are related to the asymmetric behavior of the option prices along the maturity direction, i.e., the term structure effect of the SPDs.

Dynamic analysis of the loadings

In this section, we investigate the dynamics of the loadings in the reduced model. The loadings times series have serial autocorrelations that decay slowly similarly to the integrated processes that feature a stochastic trend. Unit root and stationarity test results (not reported here) are mixed. Whenever the null hypothesis assumes the existence of a unit root (augmented Dickey-Fuller unit-root test, Phillips-Perron test, variance-ratio test for random walk) the tests reject the null, while stationary tests that have the unit root hypothesis as an alternative (KPSS test, Leybourne-McCabe stationarity test) favor the alternative. Based on these results, we further investigate if the loadings feature fractional integration between zero

and one, typical to long-memory processes. This means that a current shock impacts their future levels over a long period and eventually dissipates. To detect the long-range dependence in each series, we employ Lo (1991) modified R/S statistic with $[N^{1/4}] = 9$ and obtain $N^{1/2}R_9^1 = 5.1582$, $N^{1/2}R_9^3 = 4.5248$ and $N^{1/2}R_9^7 = 4.9893$, with 95% confidence interval $(0, 809, 1, 862)$. The tests reject the hypothesis that loadings have short-memory. We also apply Geweke and Porter-Hudak (1983) log-periodogram regression model to estimate the Hurst exponent. The estimates are $H_{GPH}^1 = 1.3736$, $H_{GPH}^3 = 1.1761$ and $H_{GPH}^7 = 1.1433$ for the cutoff $[N^{1/2}] = 50$ and the 95% confidence interval for the the GPH estimator (0.2981, 0.7019) is calculated by Weron (2002) using the bootstrapping procedure. This implies an estimated order of integration $\hat{d}^r = H_{GPH}^r - 0.5$. The simplest long-memory formulation is an autoregressive fractionally integrated moving-average model ARFIMA$(0, \hat{d}^r, 0)$. Additional AR and MA terms can be estimated but the estimation is nontrivial. For the purpose of the current work we do not pursue this endeavor.

Another way of analyzing the loading dynamics is to use a moving average window. Examining closer the dynamic relation for the loadings first difference, represented in Figure 2-3 through the 100-days moving window correlation coefficient, we see that for most of the times the volatility and negative skewness factors move together. Oftentimes the correlation of their difference $corr = (\Delta\hat{\delta}_{i1}, \Delta\hat{\delta}_{i7})$ is close to -1 and its strength weakens and is sometimes reversed, in a strong connection to the movements of the volatility of implied volatility index (Vol_{IV}), computed as a 100-days moving window standard deviation of the daily implied volatility index. The reversion in the correlation sign following the financial crises means that the OTM put options become more expensive as volatility increases. This phenomenon is explained in the empirical financial literature through the net buying pressure of index options (Bollen and Whaley (2004), Gârleanu et al.

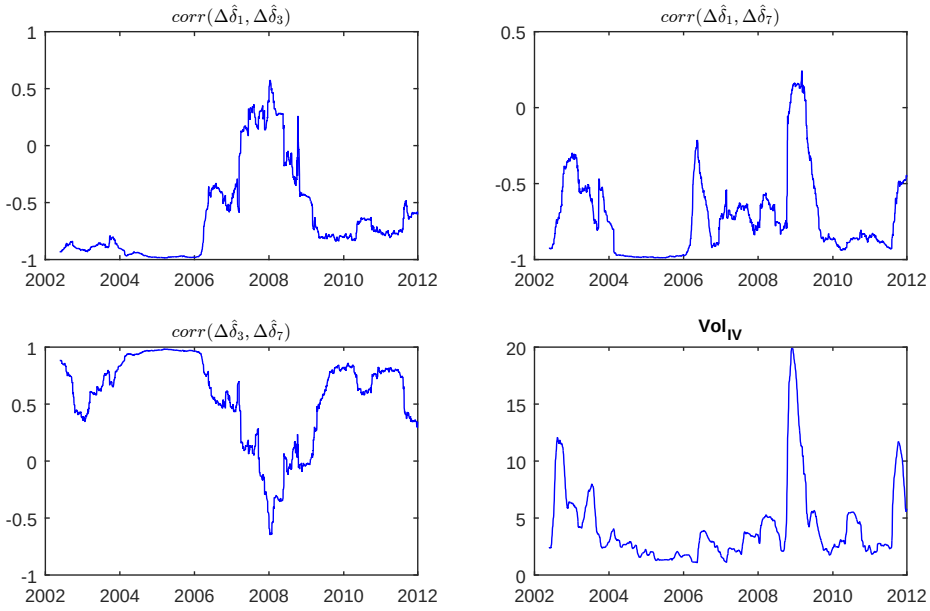


Figure 2-3: 100-days moving window correlation coefficient for the first-difference of the loadings and volatility of implied volatility

(2009)). Overall, $\hat{\delta}_{7,T}$ is linked to sudden and short-term changes in volatility.

After sustained periods of increases in the implied volatility, particularly between 2006 to the end of 2008, $\hat{\delta}_{3,T}$ decreases substantially, giving rise to more expensive OTM puts and relatively cheaper deep OTM options. The overall flattening of the left tail together with volatility increases is a manifestation of the implied volatility skew puzzle, as documented by Constantinides and Lian (2015). The authors explain it through the reduction in supply of put options from credit-constrained market makers when the demand for puts increases. Our findings according to which the difference between the prices of OTM and ATM put options decreases during the financial crisis, is consistent with their observation that the implied volatility skew declines.

Comparison with the existing literature on DAX implied volatility surfaces

The analysis of the call options traditionally takes place within the implied volatility framework. There exists a direct mapping - based on the Black-Scholes formula - between the call prices and the implied volatility. A large body of literature is concerned with the dynamics of the implied volatility surfaces. The focus is on a stylized asymmetric U-shape feature that varies across different maturities and strike prices. This pattern is called the 'smile' or 'smirk' effect. Application of PCA or FPCA to the implied volatility curves or surfaces of index options reveal usually three driving sources for its variability: a shift or level effect, a Z-shaped slope twist that impacts the skewness of the implied density, a curvature or butterfly mode that changes the convexity in the IV surface e.g. Cont and da Fonseca (2002), Fengler et al. (2003). When looking at the term structure of implied volatility, usually for fixed moneyness at the money, one factor explains most of the variability for the maturities between one month and one year, e.g. Mixon

(2002) for a study of S&P 500 index implied volatility. Fengler et al. (2002) find that the dynamics of term structure in implied volatility as measured by VDAX subindices can be represented as a two-factor model.

The decomposition of SPD variation is important because it gives the counterpart of the implies volatility surface variation, which is already fairly well understood in the financial literature. The level changes in the implied volatility surfaces are well represented for the case of SPDs by the first component. In our model, changes in skewness and kurtosis occur simultaneously and manifest through two distinct mechanisms: one affects the degree of negative skewness and the other one influences the sign of the skewness. We do not identify in our model a separate residual kurtosis factor. This is because either changes in skewness and kurtosis are manifestations of the same phenomenon or (and) usually, the amount of variance explained by the kurtosis factor is quite small.

2.4 Conclusions

We present two methods for estimating the derivatives of high-dimensional curves using FPCA techniques. In the first approach, FPCA is applied to the dual covariance matrix of the curve derivative. The second approach considers in the decomposition of the dual covariance for the original curves, whereas derivatives are applied to their functional principal components. Thus, the second approach explains the dynamics of derivatives in terms of orthogonal loadings but the components are no longer orthonormal. When an underlying factor model is assumed, we show that when estimating the curves from the observed discrete and noisy data, the second method performs better both asymptotically and in finite sample. In the real data example we find that three components can explain most of the variability in the data. Additional factors describe the variation of the term

structure of the SPD. The empirical analysis provides new insights into the economics behind the option pricing, which suggest the need to reconsider the last generation of arbitrage-free models for option pricing, see representative Bates (2006).

2.5 Appendix

2.5.1 Assumptions summary

Assumption 1 *The curves Y_i, $i = 1, \ldots, N$ are observed at a random grid $t_{i1}, \ldots, t_{iT_i}$, $t_{ij} \in [0,1]^g$ having a common bounded and continuously differentiable density f with support $\operatorname{supp}(f) = [0,1]^g$ and the integrand $u \in \operatorname{supp}(f)$ and $\inf_u f(u) > 0$.*

Assumption 2 *$\mathsf{E}(\varepsilon_{ik}) = 0$, $\operatorname{Var}(\varepsilon_{ik}) = \sigma^2_{i\varepsilon} > 0$ and ε_{ik} are independent of X_i, and $\mathsf{E}[\varepsilon^4_{ik}] < \infty, \forall i, k$.*

Assumption 3 *Let $K_B(u) = \frac{1}{b_1 \times \cdots \times b_g} K(u \circ b)$. K is a product kernel based on symmetric univariate kernels. B is a diagonal matrix with $b = (b_1, \ldots, b_g)^\top$ at the diagonal. The kernel K is bounded and has compact support on $[-1,1]^g$ such that for $u \in \mathbb{R}^g$ $\int uu^T K(u)du = \mu(K)I$ where $\mu(K) \neq 0$ is a scalar and I is the $g \times g$ identity matrix. Conditions 2 and 3 from Masry (1996) are fulfilled.*

Assumption 4 *$\rho - \sum_{l=1}^g d_l$ and $p - \sum_{l=1}^g d_l$ are odd.*

Assumption 5 $|\hat{\sigma}^2_{i\varepsilon} - \sigma^2_{i\varepsilon}| = \mathcal{O}_P(T^{-1/2})$

Assumption 6 *We require that it holds*

$$\sup_{r \in \mathbb{N}} \sup_{t \in [0,1]^g} |\varphi_r^{(d)}(t)| < \infty \ , \ \sup_{r \in \mathbb{N}} \sup_{t \in [0,1]^g} |\gamma_r^{(d)}(t)| < \infty \tag{2.40}$$

$$\sum_{r=1}^{\infty} \sum_{s=1}^{\infty} \mathsf{E}\left[\left(\delta_{ri}^{(\nu)}\right)^2 \left(\delta_{si}^{(\nu)}\right)^2 \right] < \infty \ , \ \sum_{q=1}^{\infty} \sum_{s=1}^{\infty} \mathsf{E}\left[\left(\delta_{ri}^{(\nu)}\right)^2 \delta_{si}^{(\nu)} \delta_{qi}^{(\nu)} \right] < \infty, \quad \nu = (0,d) \tag{2.41}$$

for all $r \in \mathbb{N}$.

Assumption 7 *We require that the eigenvalues are distinguishable such that for any T and N and fixed $r \in 1, \ldots, L$ there exists $0 < C_{1,r} < \infty$, $0 < C_{2,r} \leq C_{3,r} < \infty$ such that*

$$\begin{aligned} NC_{2,r} \leq l_r^{(\nu)} \leq NC_{3,r} \\ \min_{s=1,\ldots,N; s \neq r} |l_r^{(\nu)} - l_s^{(\nu)}| \geq NC_{1,r}. \end{aligned} \tag{2.42}$$

2.5.2 Proof of Lemma 2.2.1

Univariate case (g=1)

In the following proof we use d instead of ν. As noted by Ruppert and Wand (1994) equation (2.16) can be stated up to a vanishing constant using equivalent kernels. Equivalent kernels can be understand as an asymptotic version of W_d^T. Let e_l be a vector of length ρ with 1 at the $l+1$ position and zero else. Then $W_d^T(t)$ evaluates the function at point u and is defined as $(Tb^{d+1})^{-1} e_d^T S_T(u)^{-1}(1, t, \ldots, t^\rho)^T K(t)$. $S_T(u)$ is a $\rho \times \rho$ matrix with entries $S_{T,k}(u) = (Tb)^{-1} \sum_{l=1}^T K\left(\frac{t_l - u}{b}\right) (\frac{t_l - u}{b})^k$ such that

$$S_T(u) = \begin{pmatrix} S_{T,0}(u) & S_{T,1}(u) & \ldots & S_{T,\rho}(u) \\ S_{T,1}(u) & S_{T,2}(u) & \ldots & S_{T,\rho+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{T,\rho}(u) & S_{T,\rho+1}(u) & \ldots & S_{T,2\rho}(u) \end{pmatrix}. \tag{2.43}$$

Accordingly

$$\begin{aligned} \mathsf{E}(S_{T,k}(u)) =& (Tb)^{-1} \int_0^1 \sum_{l=1}^T K\left(\frac{x-u}{b}\right) \left(\frac{x-u}{b}\right)^k f(x) dx \\ =& b^{-1} \int_u^{1+u} K\left(\frac{x}{b}\right) \left(\frac{x}{b}\right)^k f(x) dx = \int_{ub^{-1}}^{(1+u)b^{-1}} K(t)\, t^k f(tb) dt. \end{aligned} \tag{2.44}$$

Since $K(t)$ has compact support and is bounded, for a point at the left boundary

with $c \geq 0$ u is of the form $u = cb$ and at the right boundary $u = 1 - cb$ respectively. We define $S_{k,c} = \int_{-c}^{\infty} t^k K(t)dt$ and $S_{k,c} = \int_{-\infty}^{c} t^k K(t)dt$ respectively and for interior points $S_k = \int_{-\infty}^{\infty} t^k K(t)dt$. Further we construct the $p \times p$ Matrix corresponding to (2.43) with

$$S(u) = \begin{cases} S_c = (S_{j+l,c})_{0 \leq j,l \leq \rho} & , u \text{ is a boundary point} \\ S = (S_{j+l})_{0 \leq j,l \leq \rho} & , u \text{ is an interior point} \end{cases}. \tag{2.45}$$

The equivalent kernel is then defined as $K_{d,\rho}^{u*}(t) = e_d^T S(u)^{-1}(1, t, \ldots, t^\rho)^T K(t)$ and the estimator can be rewritten as

$$\hat{X}_b^{(d)}(u) = d!\beta_d(u) = \frac{d!}{Tf(u)b^{d+1}} \sum_{l=1}^{T} K_{d,\rho}^{u*}\left(\frac{t_l - u}{b}\right) Y(t_l)\{1 + o_P(1)\}. \tag{2.46}$$

The only difference between W_d^T and $K_{d,\rho}^{u*}$ is that $S_T(u)$ is been replaced by $f(u)S(u)$. Regarding Masry (1996) we can further state that with a bandwidth fulfilling $\frac{log(T)}{Tb} \to 0$ we have uniformly in $u \in [0, 1]$ that $S_T(u)^{-1} \to \frac{S(u)^{-1}}{f(u)}$ almost surely as $T \to \infty$. We will drop the $u*$ index concerning the equivalent kernel from now on.

By construction, the equivalent kernel fulfills that using the Kronecker-Delta δ

$$\int u^k K_{d,\rho}^*(u)\, du = \delta_{d,k} \;\; 0 \leq d, k \leq \rho. \tag{2.47}$$

As mentioned by Fan et al. (1997), the design of the kernel automatically adapts to the boundary which gives the same order of convergence for the interior and boundary points, see Ruppert and Wand (1994). The estimator can be rewritten

as

$$\begin{aligned}
&\int d!^2 \sum_{j=1}^{T}\sum_{l=1}^{T} W_d^T\left(\frac{t_j-u}{b}\right) W_d^T\left(\frac{t_l-u}{b}\right) Y(t_l)Y(t_j)du \\
&= \int \frac{d!^2}{T^2 f(u)^2 b^{2d+2}} \sum_{l=1}^{T}\sum_{j=1}^{T} K_{d,\rho}^*\left(\frac{t_j-u}{b}\right) K_{d,\rho}^*\left(\frac{t_l-u}{b}\right) Y(t_l)Y(t_j)\{1+o_P(1)\}du.
\end{aligned} \tag{2.48}$$

For the expectation we get

$$\begin{aligned}
&\mathsf{E}\left(\theta_{d,\rho}|t_1,\ldots,t_T\right) \\
&= \int_0^1 d!^2 \sum_{j=1}^{T}\sum_{l=1}^{T} W_d^T\left(\frac{t_j-u}{b}\right) W_d^T\left(\frac{t_l-u}{b}\right) X(t_l)X(t_j)du \\
&\quad + d!^2\left(\sigma_\varepsilon^2-\hat{\sigma}_\varepsilon^2\right)\int_0^1 \sum_{j=1}^{T} W_d^T\left(\frac{t_j-u}{b}\right)^2 du \\
&= \left\{ d!^2 \int_0^1\int_0^1\int_0^1 \frac{f(x)f(y)}{b^{2(d+1)}f(z)^2} K_{d,\rho}^*\left(\frac{x-z}{b}\right) K_{d,\rho}^*\left(\frac{y-z}{b}\right) X(x)X(y)dxdydz \right. \\
&\quad \left. + \mathcal{O}_P\left(\frac{1}{T^{3/2}b^{2d+1}}\right)\right\}\{1+o_P(1)\} \\
&= \left\{ \int_0^1 X^{(d)}(z)X^{(d)}(z)dz \right. \\
&\quad + 2\frac{d!}{(\rho+1)!}\int_0^1 \frac{b^{\rho+1}}{b^d}\left(\int_0^1 u^{\rho+1}K_{d,\rho}^*(u)\,du\right) X^{(\rho+1)}(z)X^{(d)}(z)dz \\
&\quad + \frac{d!^2}{(\rho+1)!^2}\int_0^1 \frac{b^{2\rho+2}}{b^{2d}}\left(\int_0^1 u^{\rho+1}K_{d,\rho}^*(u)\,du\right)^2 X^{(\rho+1)}(z)X^{(\rho+1)}(z)dz \\
&\quad \left. + \mathcal{O}_P\left(\frac{1}{T^{3/2}b^{2d+1}}\right)\right\}\{1+o_P(1)\}
\end{aligned} \tag{2.49}$$

These results where obtained by substitution with $x = z+ub, y = z+vb$ and using a $\rho+1$ order Taylor expansion of $X(z+ub)$ and $X(z+vb)$ together with (2.47). We get $\int_{[0,1]^g} X(u)^2du - \mathsf{E}(\theta_{d,\rho}|t_1,\ldots,t_T) = \mathcal{O}_p\left(b^{\rho+1-d} + \left(T^{3/2}b^{2d+1}\right)^{-1}\right)$.

First note that using the second mean value integration theorem there exits some $c \in (0,1)$ and we can write

$$\int f(z)^{-2} K^*_{d,\rho}\left(\frac{y-z}{b}\right) K^*_{d,\rho}\left(\frac{x-z}{b}\right) dz = f(c)^{-2} \int K^*_{d,\rho}\left(\frac{y-z}{b}\right) K^*_{d,\rho}\left(\frac{x-z}{b}\right) dz. \tag{2.50}$$

We introduce a kernel convolution with

$$K^C_{d,\rho}(y-x) \stackrel{\text{def}}{=} \int K^*_{d,\rho}(y-z)\, K^*_{d,\rho}(x-z)\, dz \tag{2.51}$$

and thus using $z = \frac{u}{b}$

$$K^C_{d,\rho}\left(\frac{y-x}{b}\right) = \int K^*_{d,\rho}\left(\frac{y}{b} - z\right) K^*_{d,\rho}\left(\frac{x}{b} - z\right) dz = \int b^{-1} K^*_{d,\rho}\left(\frac{y-u}{b}\right) K^*_{d,\rho}\left(\frac{x-u}{b}\right) du. \tag{2.52}$$

Note that the integral over $K^C_{d,\rho}$ is computed over an parallelogram D bounded by the lines $x+y=2, x+y=0, x-y=1, x-y=-1$. Using the substitution $x = \frac{v+u}{2}b,\ y = \frac{u-v}{2}b$

$$\int\int_D K^C_{d,\rho}\left(\frac{y-x}{b}\right) dydx = \frac{b}{2}\int_0^2 \int_{-1}^1 K^C_{d,\rho}\left(\frac{v+u-u+v}{2}\right) dvdu = b\int K^C_{d,\rho}(v)\, dv. \tag{2.53}$$

Note that the variance can be decomposed

$$\begin{aligned}
&\mathrm{Var}\left(\theta_{d,\rho}|t_1,\ldots,t_T\right) && (2.54)\\
=&\frac{d!^4}{T^4(b^{4d+2})f(c)^4}\left\{\sum_{l=1}^{T}K_{d,\rho}^{C}\left(0\right)^2\mathrm{Var}(Y(t_l)^2)\right. && (2.55)\\
&+2\sum_{l=1}^{T}\sum_{k\neq l}^{T}\mathrm{Var}(K_{d,\rho}^{C}\left(\frac{t_l-t_k}{b}\right)Y(t_l)Y(t_k)) && (2.56)\\
&+4\sum_{l=1}^{T}\sum_{k\neq l}^{T}\sum_{k'\neq k}^{T}\mathrm{Cov}(K_{d,\rho}^{C}\left(\frac{t_k-t_l}{b}\right)Y(t_k)Y(t_l),K_{d,\rho}^{C}\left(\frac{t_l-t_{k'}}{b}\right)Y(t_l)Y(t_{k'})) && (2.57)\\
&\left.+\ 24\sum_{l=1}^{T}\sum_{k\neq l}^{T}\sum_{k'\neq k}^{T}\sum_{l'\neq k'}^{T}\mathrm{Cov}(K_{d,\rho}^{C}\left(\frac{t_l-t_k}{b}\right)Y(t_l)Y(t_k),K_{d,\rho}^{C}\left(\frac{t_{l'}-t_{k'}}{b}\right)Y(t_{l'})Y(t_{k'}))\right\} && (2.58)\\
&+\mathcal{O}_P\left(\frac{1}{T}\right). && (2.59)
\end{aligned}$$

Expression (2.58) vanishes and (2.55) given by $\frac{d!^4}{T^3(b^{4d+2})f(c)^4}\int K_{d,\rho}^{C}\left(0\right)^2\mathrm{Var}(Y(y)^2)f(y)dy\{1+\mathcal{O}_P(T^{-1})\}$ is dominated by (2.56) because

$$\begin{aligned}
&\frac{2d!^4}{T^4(b^{4d+2})f(c)^4}\sum_{l=1}^{T}\sum_{k\neq l}^{T}K_{d,\rho}^{C}\left(\frac{t_l-t_k}{b}\right)^2\mathrm{Var}(Y(t_l)Y(t_k))\\
=&\frac{2d!^4}{T^4(b^{4d+2})f(c)^4}\sum_{l=1}^{T}\sum_{k\neq l}^{T}K_{d,\rho}^{C}\left(\frac{t_l-t_k}{b}\right)^2\left\{\mathsf{E}(Y(t_l)^2Y(t_k)^2)-\mathsf{E}(Y(t_l)Y(t_k))^2\right\}\\
=&\frac{2d!^4\int(\sigma_\epsilon^4+2\sigma_\epsilon^2X(x)^2)f(x)^2dx}{T^2b^{4d+1}f(c)^4}\int\left(K_{d,\rho}^{C}(u)\right)^2du+\mathcal{o}_P\left(\frac{1}{T^2b^{4d+1}}\right). && (2.60)
\end{aligned}$$

Before looking at expression (2.57), note that with $m \geq 2d$

$$\begin{aligned}
&\int\int \frac{d!^2}{b^{2d+1}} K_{d,\rho}^C\left(\frac{x-y}{b}\right) X(x)dxdy \\
=&\frac{d!^2}{b^{2d}} \int\int\int K_{d,\rho}^*(m)\, K_{d,\rho}^*(z)\, X\{y+(m-z)b\}\, dzdmdy \qquad (2.61) \\
=&(-1)^d \int_0^1 X^{(2d)}(y)dy + \mathcal{O}_P(1)
\end{aligned}$$

by performing two Taylor expansions with mb first and then $-zb$.

We can thus derive for expression (2.57) that

$$\begin{aligned}
&H(T)\sum_{l=1}^{T}\sum_{k\neq l}^{T}\sum_{k'\neq k}^{T} \mathrm{Cov}(K_{d,\rho}^C\left(\frac{t_k-t_l}{b}\right) Y(t_k)Y(t_l), K_{d,\rho}^C\left(\frac{t_l-t_{k'}}{b}\right) Y(t_l)Y(t_{k'})) \\
=&H(T)\sum_{l=1}^{T}\sum_{k\neq l}^{T}\sum_{k'\neq k}^{T} K_{d,\rho}^C\left(\frac{t_k-t_l}{b}\right) K_{d,\rho}^C\left(\frac{t_l-t_{k'}}{b}\right) \left\{\mathsf{E}\left(Y(t_k)Y(t_l)^2Y(t_{k'})\right)\right. \\
&\left.-\mathsf{E}\left(Y(t_k)Y(t_l)\right)\mathsf{E}\left(Y(t_l)Y(t_{k'})\right)\right\} \\
=&H(T)\sum_{l=1}^{T}\sum_{k=1}^{T}\sum_{k'=1}^{T} K_{d,\rho}^C\left(\frac{t_k-t_l}{b}\right) K_{d,\rho}^C\left(\frac{t_l-t_{k'}}{b}\right) X(t_k)\sigma_\epsilon^2 X(t_{k'}) \\
&-\frac{2d!^4}{T^4(b^{4d+2})f(c)^4}\sum_{k=1}^{T}\sum_{k'=1}^{T} K_{d,\rho}^C\left(\frac{t_l-t_{k'}}{b}\right)^2 X(t_k)\sigma_\epsilon^2 X(t_{k'}) \\
=&\frac{4\sigma_\epsilon^2}{Tf(c)}\int X^{(2d)}(y)X^{(2d)}(y)dy - \mathcal{O}_P\left(\frac{1}{T^2(b^{4d+1})}\right),
\end{aligned}$$

where $H(T) \stackrel{\text{def}}{=} \frac{4d!^4}{T^4(b^{4d+2})f(c)^4}$. Thus $\mathrm{Var}\left(\theta_{d,\rho}|t_1,\ldots,t_T\right) = \mathcal{O}_P\left(\frac{1}{T^2(b^{4d+1})}\right)$.

Multivariate case ($g > 1$)

The same strategy also works in the multivariate case by using multivariate Taylor series. Using the multi-index notation introduces in section 2.2.4 and $a =$

$(a_1, ..., a_g),\ a_l \in \mathbb{N}^+$ a multivariate taylor series of degree $k < \rho$ is given by

$$X(x - u \circ b) = \sum_{0 \leq |a| \leq k} \frac{X^{(a)}(x)}{a!}(u \circ b)^a + \mathcal{O}_P\left(u^{k+1} \max(b)^{k+1}\right). \tag{2.62}$$

Using the equivalent kernel by Ruppert and Wand (1994) extended to the case and using Masry (1996) we can further state that with a bandwidth fulfilling $\frac{log(T)}{Tb_1 \times \cdots \times b_g} \to 0$ we have uniformly in $u \in [0,1]^g$ that $S_T(u)^{-1} \to \frac{S(u)^{-1}}{f(u)}$ almost surely as $T \to \infty$. Furthermore, the multivariate equivalent kernel has the properties that with $v = (v_1, \ldots, v_g),\ v_l \in \mathbb{N}^+$

$$\int u^v K^*_{d,\rho}(u)\, du = \delta_{d,v},\ |v| \leq \rho,\ 0 \leq d_i\ \forall i = 1, \ldots g. \tag{2.63}$$

Let c be the position of $\max(b)$ in b and $\tilde{\rho}$ be a vector of length g which is $\rho + 1$ at the $c - th$ position and 0 else. Then for the bias

$$\begin{aligned}
&\mathsf{E}\left(\theta_{d,\rho} | t_1, \ldots, t_T\right) \\
&= \Bigg\{ \int_{[0,1]^g} X^{(d)}(z) X^{(d)}(z) dz \\
&\quad + 2\frac{d!}{(\rho+1)!} \int_{[0,1]^g} \frac{\max(b)^{\rho+1}}{b^d} \left(\int u^{\tilde{\rho}} K^*_{d,\rho}(u)\, du \right) X^{(\tilde{\rho})}(z) X^{(d)}(z) dz \\
&\quad + \mathcal{O}_P\left(\frac{\max(b)^{\rho+1}}{b^d} + \frac{1}{T^{3/2}(b^{2d} b_1 \times \cdots \times b_g)} \right) \Bigg\} \{1 + o_P(1)\}
\end{aligned} \tag{2.64}$$

Further note that for the convoluted kernel we get

$$\begin{aligned}
&K^C_{d,\rho}\left((y - x) \circ b^{-1}\right) \\
&= \int (b_1 \times \cdots \times b_g)^{-1} K^*_{d,\rho}\left\{(y - u) \circ b^{-1}\right\} K^*_{d,\rho}\left\{(x - u) \circ b^{-1}\right\} du.
\end{aligned}$$

Accordingly, we get for the multivariate equivalent of expression (2.56) that

$$\begin{aligned}
&\frac{2d!^4}{T^4 f(c)^4 (b_1^2 \times \cdots \times b_g^2 b^{4d})} \sum_{l=1}^{T} \sum_{k \neq l}^{T} K_{d,\rho}^{C} \left((t_l - t_k) \circ b^{-1} \right)^2 Var(Y(t_l) Y(t_k)) \\
=&\frac{2d!^4 \int (\sigma_\epsilon^4 + 2\sigma_\epsilon^2 X(x)^2) f(x)^2 dx}{T^2 f(c)^4 b_1 \times \cdots \times b_g b^{4d}} \int \left(K_{d,\rho}^{C}(u) \right)^2 du \{1 + o_P(1)\}
\end{aligned}$$

and because we assume that $m \geq 2|d|$ we get for the multivariate equivalent of expression (2.57) that

$$\begin{aligned}
&A(T) \sum_{l=1}^{T} \sum_{k \neq l}^{T} \sum_{k' \neq k}^{T} \mathrm{Cov}(K_{d,\rho}^{C} \left((t_k - t_l) \circ b^{-1} \right) Y(t_k) Y(t_l), K_{d,\rho}^{C} \left((t_l - t_{k'}) \circ b^{-1} \right) Y(t_l) Y(t_{k'})) \\
=&A(T) \sum_{l=1}^{T} \sum_{k \neq l}^{T} \sum_{k' \neq k}^{T} K_{d,\rho}^{C} \left((t_k - t_l) \circ b^{-1} \right) K_{d,\rho}^{C} \left((t_l - t_{k'}) \circ b^{-1} \right) X(t_k) \sigma^2 X(t_{k'}) \\
=&\frac{4\sigma_\epsilon^2}{T f(c)} \int X^{(2d)}(y) X^{(2d)}(y) dy + \mathcal{O}_P \left(\frac{1}{T^2 (b^{4d} b_1 \times \cdots \times b_g)} \right)
\end{aligned}$$

where $A(T) \overset{\text{def}}{=} \frac{4d!^4}{T^4 (b^{4d} b_1^2 \times \cdots \times b_g^2) f(c)^4}$.

2.5.3 Proof of Proposition 2.2.2

Asymptotic results

We first have look at the estimator $\tilde{M}^{(0)}$ for the special case when a common random grid is present. The only error here comes from approximating the integral

in equation (2.12) with a sum.

$$
\begin{aligned}
M_{ij}^{(0)} - \tilde{M}_{ij}^{(0)} &= \int_{[0,1]^g} X_i(t)X_j(t)dt - \frac{1}{T}\sum_{l=1}^{T} Y_i(t_{il})Y_j(t_{jl}) + I(i=j)\hat{\sigma}_{i\varepsilon}^2 \\
&= \int_{[0,1]^g} X_i(t)X_j(t)dt - \frac{1}{T}\sum_{l=1}^{T} \left(X_i(t_l) + \varepsilon_{il}\right)\left(X_j(t_l) + \varepsilon_{jl}\right) + I(i=j)\hat{\sigma}_{i\varepsilon}^2 \\
&= \int_{[0,1]^g} X_i(t)X_j(t)dt - \frac{1}{T}\sum_{l=1}^{T} X_i(t_l)X_j(t_l) \\
&\quad - \frac{1}{T}\sum_{l=1}^{T} X_i(t_l)\varepsilon_{jl} - \frac{1}{T}\sum_{l=1}^{T} X_j(t_l)\varepsilon_{il} - \frac{1}{T}\sum_{l=1}^{T_i} \varepsilon_{il}\varepsilon_{jl} + I(i=j)\hat{\sigma}_{i\varepsilon}^2.
\end{aligned}
\tag{2.65}
$$

By construction, it hold that $\mathsf{E}\left[\varepsilon_{il}\varepsilon_{jl}\right] = 0,\ i \neq j$, $\mathsf{E}\left[{\varepsilon_{il}}^2\right] = \sigma_{i\varepsilon}^2$ and $\mathsf{E}\left[Y_i(t_l)\varepsilon_{jl}\right] = 0$. All sums for example $\frac{1}{T}\sum_{l=1}^{T} X_i(t_l)X_j(t_l)$ are the corresponding empirical estimator for the mean, i.e., $\int_{[0,1]^g} X_i(t)X_j(t)dt = \mathsf{E}\left[X_iX_j\right]$. By the law of large numbers, it converges in probability to the theoretical mean as $T \to \infty$. Using the central limit theorem we can further state that $\int_{[0,1]^g} X_i(t)X_j(t)dt - \frac{1}{T}\sum_{l=1}^{T} X_i(t_l)X_j(t_l)$ is approximately normal, which gives an error of order $T^{-1/2}$ regardless of dimension g. By requiring that $\hat{\sigma}_{i\varepsilon}$ is also $T^{-1/2}$ consistent we get $T^{-1/2}$ for all elements.

To understand $\hat{M}^{(0)}$ we investigate two possible sources of error in the construction of the estimator. One coming from interpolation and smoothing at a common grid and the other from approximating the integral with a sum. First note that by the same arguments as for $\tilde{M}^{(0)}$ the error of the integral approximation is of order $T^{-1/2}$. Besides the error for the off diagonal elements is smaller than for the diagonal, thus the leading error source is given by Lemma 2.2.1. The same arguments also work to derive asymptotic results for $\hat{M}^{(d)}$.

2.5.4 Proof of Proposition 2.2.4

Under the assumptions of Proposition 2.2.4 together with the requirements of Lemma 2.2.2 for $\nu = (0, d)$ and the setup of Remark 2.2.3

$$||\hat{M}^{(\nu)} - M^{(\nu)}|| \leq \text{tr}\left\{\left(\hat{M}^{(\nu)} - M^{(\nu)}\right)^{\top}\left(\hat{M}^{(\nu)} - M^{(\nu)}\right)\right\}^{1/2} = \mathcal{O}_p\left(NT^{-1/2}\right). \tag{2.66}$$

Given that $\sum_{l=1}^{T} p_{lr}^{(\nu)} = 0$, $\sum_{l=1}^{T}\left(p_{lr}^{(\nu)}\right)^2 = 1\ \forall r$ and applying Cauchy-Schwarz inequality gives $\sum_{l=1}^{N}|p_{lr}^{(\nu)}| = \mathcal{O}\left(N^{1/2}\right)$. This together with Lemma A from Kneip and Utikal (2001) leads to

$$\mathsf{E}\left[\left(p_r^{(\nu)}\right)^{\top}(\hat{M}^{(\nu)} - M^{(\nu)})p_r^{(\nu)}\right]^2 = \mathcal{O}_p\left(\frac{N}{T}\right) \tag{2.67}$$

We are now ready to make a statement about the basis that span the factor space.

$$\begin{aligned}
&\left|\frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N} p_{ir}^{(\nu)} X_i^{(d)}(t) - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\sum_{i=1}^{N}\hat{p}_{ir}^{(\nu)}\hat{X}_{i,h}^{(d)}(t)\right| \\
\leq &\left|\frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N} p_{ir}^{(\nu)}\left[X_i^{(d)}(t) - \hat{X}_{i,h}^{(d)}(t)\right]\right| + \left|\sum_{i=1}^{N}\left(\frac{1}{\sqrt{l_r^{(\nu)}}}p_{ir}^{(\nu)} - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\hat{p}_{ir}^{(\nu)}\right)\hat{X}_{i,h}^{(d)}(t)\right|.
\end{aligned} \tag{2.68}$$

The first term is discussed in equation (2.2.4). Therefore we take a look at the second term here. As a consequence of Assumption (7), Lemma A (a) from Kneip and Utikal (2001) together with equation (2.67) gives

$$l_r^{(\nu)} - \hat{l}_r^{(\nu)} = (p_r^{(\nu)})^T(\hat{M}^{(\nu)} - M^{(\nu)})p_r^{(\nu)}) + \mathcal{O}_p(NT^{-1}) = \mathcal{O}_p(N^{1/2}T^{-1/2} + NT^{-1}), \tag{2.69}$$

where

$$\frac{1}{\sqrt{\hat{l}_r^{(\nu)}}} - \frac{1}{\sqrt{l_r^{(\nu)}}} = \frac{l_r^{(\nu)} - \hat{l}_r^{(\nu)}}{\sqrt{\hat{l}_r^{(\nu)}}\sqrt{l_r^{(\nu)}}(\sqrt{\hat{l}_r^{(\nu)}} + \sqrt{l_r^{(\nu)}})} = \mathcal{O}_p\left(T^{-1/2}N^{-1} + T^{-1}N^{-1/2}\right). \tag{2.70}$$

Using Lemma A (b) from Kneip and Utikal (2001) we further get

$$|\hat{p}_{ir}^{(\nu)} - p_{ir}^{(\nu)}| = \mathcal{O}_p\left((NT)^{-1/2}\right) \text{ and } ||\hat{p}_r^{(\nu)} - p_r^{(\nu)}|| = \mathcal{O}_p\left(T^{-1/2}\right). \tag{2.71}$$

Putting all results together for the second term gives

$$\begin{aligned}
&\left|\sum_{i=1}^{N}\left(\frac{1}{\sqrt{l_r^{(\nu)}}}p_{ir}^{(\nu)} - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\hat{p}_{ir}^{(\nu)}\right)\hat{X}_{i,h}^{(d)}(t)\right| = \\
=&\left|\sum_{i=1}^{N}\left(\frac{1}{\sqrt{l_r^{(\nu)}}} - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\right)\hat{p}_{ir}^{(\nu)}\hat{X}_{i,h}^{(d)}(t) + \frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N}\left(\hat{p}_{ir}^{(\nu)} - p_{ir}^{(\nu)}\right)\hat{X}_{i,h}^{(d)}(t)\right| \\
\leq&\left|\left(\frac{1}{\sqrt{l_r^{(\nu)}}} - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\right)\right|\sum_{i=1}^{N}|p_{ir}^{(\nu)}|\left|\left(\hat{X}_{i,h}^{(d)}(t)\right)\right| \\
&+\left|\left(\frac{1}{\sqrt{l_r^{(\nu)}}} - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\right)\right|||\hat{p}_r^{(\nu)} - p_r^{(\nu)}||\left|\hat{X}_{i,h}^{(d)}(t)\right| + \frac{1}{\sqrt{l_r^{(\nu)}}}||\hat{p}_r^{(\nu)} - p_r^{(\nu)}||\left|\hat{X}_{i,h}^{(d)}(t)\right| \\
=&\mathcal{O}_p\left((NT)^{-1/2}\right)\left|\hat{X}_{i,h}^{(d)}(t) - X_{i,h}^{(d)}(t) + X_{i,h}^{(d)}(t)\right| \\
\leq&\mathcal{O}_p\left((NT)^{-1/2}\right)(\text{Bias}\left(\hat{X}_{j,h}^{(d)}(t)\right) + \sqrt{\text{Var}\left(\hat{X}_{j,h}^{(d)}(t)\right)} + \left|X_{i,h}^{(d)}(t)\right|).
\end{aligned} \tag{2.72}$$

Using Cauchy-Schwarz and equation (2.70) we see that first term is of order $(NT)^{-1/2}$. For the second term remember that $l_r^{(\nu)}$ is of order N together with (2.71) this also leads to order $(NT)^{-1/2}$. Inserting the right hand side, equation

(2.68) becomes

$$\begin{aligned}
&\mathcal{O}_p\left(\max(h)^{p+1}h^{-d}\right) + \mathcal{O}_p\left((NTh_1\dots h_g h^{2d})^{-1/2}\right) + \mathcal{O}_p\left((NT)^{-1/2}\right)\mathcal{O}_p\left(\max(h)^{p+1}h^{-d}\right)\\
&+ \mathcal{O}_p\left((NT)^{-1/2}\right)\mathcal{O}_p\left((Th_1\dots h_g h^{2d})^{-1/2}\right) + \mathcal{O}_p\left((NT)^{-1/2}\right)\\
&= \mathcal{O}_p\left(\max(h)^{p+1}h^{-d}\right) + \mathcal{O}_p\left((NTh_1\dots h_g h^{2d})^{-1/2}\right).
\end{aligned}$$

2.5.5 Proof of Proposition 2.2.5

Note that

$$\sqrt{l_r^{(v)}} - \sqrt{\hat{l}_r^{(v)}} = (l_r^{(v)} - \hat{l}_r^{(v)})(\sqrt{l_r^{(v)}} + \sqrt{\hat{l}_r^{(v)}})^{-1} = \mathcal{O}_p(T^{-1/2} + N^{1/2}T^{-1}), \tag{2.73}$$

together with (2.71) gives

$$\begin{aligned}
&\hat{\delta}_{ir} - \hat{\delta}_{ir,T} = \sqrt{l_r^{(v)}}p_{ir}^{(v)} - \sqrt{\hat{l}_r^{(v)}}\hat{p}_{ir}^{(v)}\\
&= \left(\sqrt{l_r^{(v)}} - \sqrt{\hat{l}_r^{(v)}}\right)p_{ir}^{(v)} - \sqrt{\hat{l}_r^{(v)}}\left(\hat{p}_{ir}^{(v)} - p_{ir}^{(v)}\right) = \mathcal{O}_p(T^{-1/2} + N^{1/2}T^{-1}).
\end{aligned} \tag{2.74}$$

Using Proposition 2.2.4 it follows that

$$\begin{aligned}
&|Y_i(t) - \hat{Y}_i(t)| = |\sum_{r=1}^{K}\hat{\delta}_{ir}\hat{\gamma}_r^{(v)}(t) - \sum_{r=1}^{K}\hat{\delta}_{ir,T}\hat{\gamma}_{r,T}^{(v)}(t)|\\
=&|\sum_{r=1}^{K}(\hat{\delta}_{ir} - \hat{\delta}_{ir,T})\hat{\gamma}_r + \hat{\delta}_{ir,T}(\hat{\gamma}_r - \hat{\gamma}_{r,T})|\\
=&\mathcal{O}_p\left(T^{-1/2} + N^{1/2}T^{-1} + \max(h)^{p+1}h^{-d} + (NTh_1\times\cdots\times h_g h^{2d})^{-1/2}\right).
\end{aligned} \tag{2.75}$$

Chapter 3

Nonparametric Registration to Low-Dimensional Function Spaces

Abstract

Registration aims to decompose amplitude and phase variation of samples of curves. Phase variation is captured by warping functions which monotonically transform the domains. Resulting registered curves should then only exhibit amplitude variation. Most existing registration method rely on aligning typical shape features like peaks or valleys to be found in each sample function. It is shown that this is not necessarily an optimal strategy for subsequent statistical data exploration and inference. In this context a major goal is to identify low dimensional linear subspaces of functions that are able to provide accurate approximations of the observed functional data. In this paper we present a registration method where warping functions are defined in such a way that the resulting registered curves span a low dimensional linear function space. Problems of identifiability are discussed in detail, and connections to established registration procedures are analyzed. The method is applied to real and simulated data.

3.1 Introduction

The data that we consider are a sample of i.i.d. smooth random functions $x_1, \ldots, x_n$ defined over a closed interval on the real line. Registration literature focuses on

the situation that all functions share a common set of shape features, such as peaks and valleys. The curves displayed in the top panel of Figure 3-1 provide an example. The sizes of the features vary, and we refer to this as *amplitude variation*. The locations of the features also vary from curve to curve, which indicates the existence of *phase variation*. Generally speaking, registration deals with separating amplitude and phase variation in a statistically meaningful way. The aim is to search for a set of smooth strictly monotonic functions h_i, called *warping functions*, which eliminate phase variation such that the *registered* functions $y_i(t)$ of the form $y_i(t) = x_i(h_i(t)) = (x_i \circ h_i)(t)$ represent amplitude variation. Since monotone transformations do not destroy shape features the registered functions will possess the same sequences of peaks and valleys as the original functions x_i.

It is well-known that phase variation is present in many important applications, and it poses severe problems for the application of functional versions of commonly used multivariate data analyses such as computing pointwise means, variances and correlations, principal components analysis and canonical correlation analyses (Ramsay and Silverman (2005); Silverman (1995)).

Traditional literature on the registration problem aims to define warping functions in such a way that registered functions y_i have all shape features aligned. A frequently used method in older studies is *landmark registration*, see e.g. Bookstein (1978, 1997), Kneip and Gasser (1992) and Gasser and Kneip (1995). Many other methods not using landmarks have also been developed, partly in response to situations where shape features used as landmarks are not clearly identifiable in all curves. A common property of the most important methods proposed in this context is to determine warping functions h_i by minimizing a distance $d(x_i \circ h_i, \gamma)$ between registered functions $y_i(t) = x_i(h_i(t))$ and a template $\gamma(t)$. There is a considerable literature proposing algorithms which aim to minimize the distance $d_2(x_i \circ h_i, \gamma) = \|x_i \circ h_i - \gamma\|_2$, where $\|\cdot\|_2$ denotes the L^2-distance, see, for example,

Sakoe and Chiba (1978). Ramsay (1998), Ramsay and Li (1998), or Kneip et al. (2000). Usually additional regularization techniques are applied.

Well-known problems with these techniques have lead to the development of more sophisticated techniques based on alternative distance measures. For example, Ramsay and Silverman (2005), Wang and Gasser (1997, 1998, 1999), or Srivastava et al. (2011)) propose to minimize semi-metrics with the property that $d(x_i \circ h_i, \gamma) = 0$ if $x_i \circ h_i = a_i\gamma$ for some $a_i \in \mathbb{R}$.

All these methods share a common point of view. The success of a registration method is assessed in terms of how well it is able to align visible features. Templates are often determined iteratively from the sample and their construction aims to establish a "structural mean" which possess all common shape features at mean locations and with mean amplitude. Hence, traditionally registration tends to concentrate on establishing a most informative mean curve summarizing the sample functions.

However, more recent work also tends to apply registration procedures in the context of more complex problems of statistical data exploration and inference. In functional data analysis the most frequently applied procedures are based on identifying **low dimensional linear subspaces of functions** that are able to provide accurate approximations of the observed functional data. An essential tool is functional principal component analysis (FPCA), where sample curves are approximated as elements of the linear space generated by a few leading functional principal components. For functions exhibiting a registration problem, Hadjipantelis et al. (2015) use a norm based method for aligning functions, and then apply FPCA separately for registered curves and warping functions. Gervini and Gasser (2005), Claeskens et al. (2010), or Slaets et al. (2012) present multi-resolution approaches to registration. Assuming on discretized observations, they rely on *prespecified* basis expansions for amplitude and phase variation and use algorithms

designed for fitting mixed-effect models.

For clustering functions, Sangalli et al. (2010) propose a procedure which is based on several templates instead of only a single "structural mean". The "k mean" approach assumes, that each observed curve belongs to one of k specific clusters. The method then tries to determine the mean (template) of each cluster iteratively and uses scale and shift to align the curves within the clusters.

In this paper we consider registration from a more general point of view. Registration may be used as a tool for statistical analysis whenever the random functions x_i possess "bounded shape variation", i.e. there exists a fixed value $\mathbf{q} < \infty$ such that with probability 1 the number of shape features to be found within each possible realization does not exceed $\mathbf{q}$. Our approach is based on an observation already made by Kneip and Ramsay (2008) that for random functions with bounded shape variation there exists a finite K and warping functions h_i such that with probability 1

$$x_i(h_i(t)) = \sum_{j=1}^{K} a_{ij}\gamma_j(t). \tag{3.1}$$

for some basis functions $\gamma_1, \ldots, \gamma_K$ and individually different coefficients $a_{i1}, \ldots, a_{iK}$.

We are going beyond Kneip and Ramsay (2008) by studying decomposition (3.1) from a theory-guided, conceptional point of view and by deriving some basic inference results for situations, where the true functions have to be reconstructed from discrete, noisy observations. Appropriate values of K depend on the structure of x_i, and possible non-uniqueness of solutions to (3.1) are resolved by selecting the registration procedure with the least complex warping functions. Furthermore, we present a new algorithm which estimates the components of (3.1) for all possible values of K and seems to work well for many applications.

Assuming that functional shapes are of bounded complexity does not seem to

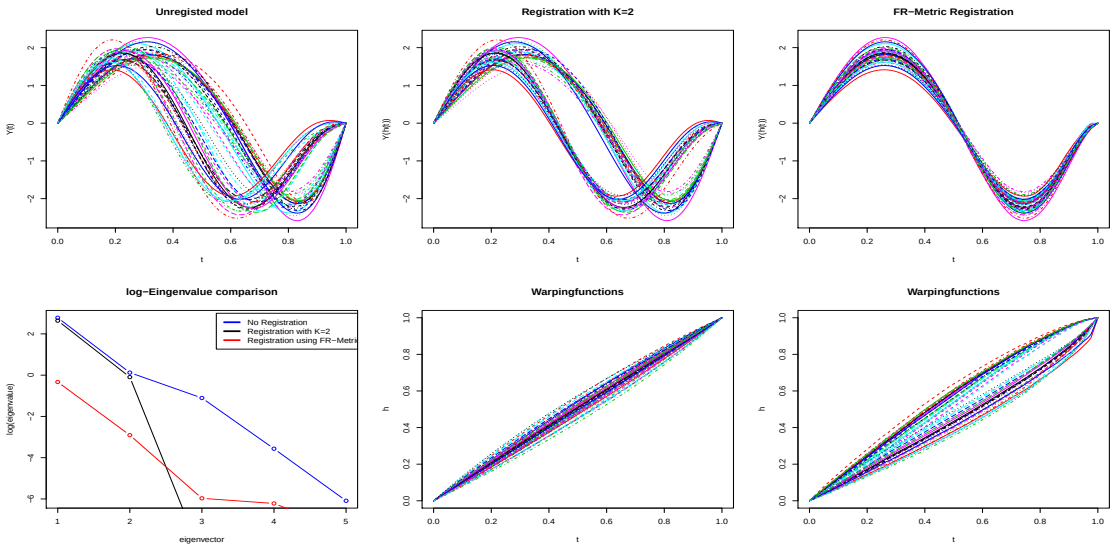


Figure 3-1: Example for curves generated by (3.1) with $K = 2$. The lower left Figure provides the log eigenvalues of an FPCA decomposition for the three types of registration given in the upper figures. The alignment of the peaks increases the model complexity (log-Eigenvalues) and the complexity of the warping functions compared to a registration using $K = 2$.

be restrictive in important applications for instance consider biomedicine, technics, chemometrics, etc., and often the presence of phase variation is already imposed from a substantial point of view (different reactions times, etc.). Our approach then generalizes the rather limited range of applicability of traditional registration techniques. Together with a suitable analysis of warping functions, the method allows to decompose functional data in a way that might be more informative than standard functional principal component analysis (FPCA).

If $K \geq 2$, then an optimal registration based on (3.1) will usually not align shape features, since for the registered curves $y_i(t) = x_i(h_i(t))$ existence and locations of shape features will depend on the interplay between the coefficients $a_{i1}, \ldots, a_{ik}$. Our approach is illustrated by Figures 3-1 and 3-2. They represent simulated data, and a description of the underlying data generating processes is given in the online appendix.

Figure 1 corresponds to the type of data usually considered in a registration

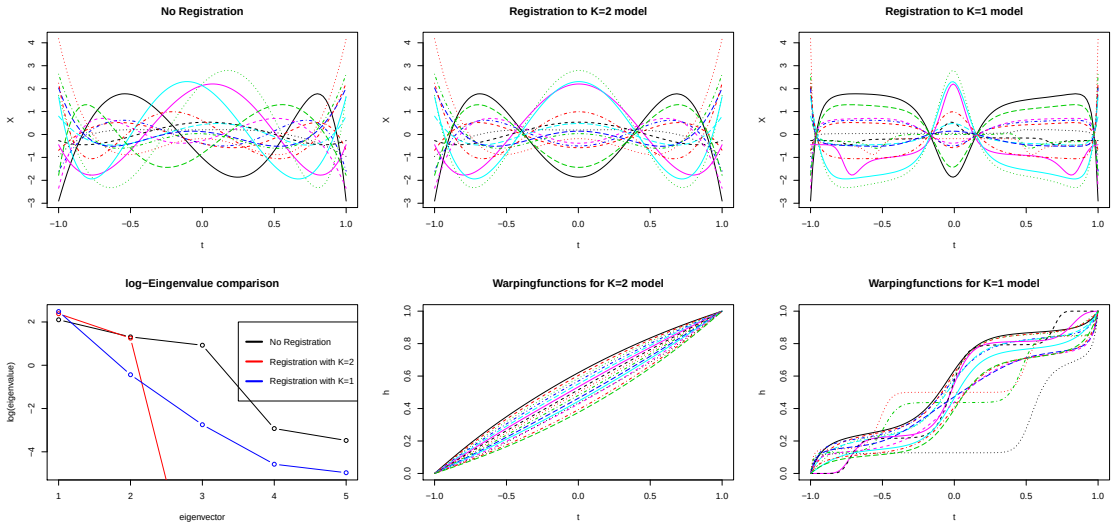


Figure 3-2: Registration of curves generated by Simulation 3.6.2 using our algorithm from Section 3.4. The upper right figure shows a registration using $K = 1$ which results in a visible curve pinching in order to archive some kind of peak alignment. The registration to $K = 2$ shown in the lower right figure does not align peaks but reduces the model complexity as seen due to the log eigenvalues of an FPCA presented in the lower right figure.

context and satisfies (3.1) with $K = 2$. There is a typical sequence of shape features which can be found in all curves. Registration aiming at peak alignment is done with the R-package "fdasrvf" by Tucker (2014). The package implements the method proposed by Srivastava et al. (2011), where the Fisher-Rao metric is used to determining warping functions by minimizing $d(x_i \circ h_i, \gamma)$ with respect to a suitably defined templat functon γ. Since $d(x_i \circ h_i, \gamma) = 0$ if and only if $x_i \circ h_i = a_i\gamma$ for some $a_i \in \mathbb{R}$, this may be interpreted as an algorithm for fitting (3.1) with $K = 1$. Indeed, it is shown in Section 2 that peak alignment in tendency corresponds to adjusting a one dimensional model. At the same time, the figure shows that an optimal selection of warping functions depends on K, and less complex functions are determined by the $K = 2$ dimensional registration. As can be seen from an FPCA decomposition of $y_i = x_i \circ h_i$, for the one dimensional approximation the space of the registered functions is more complex and cannot be described by two components anymore.

The functional data presented in Figure 3-2 are qualitatively different from those of Figure 3-1 in the sense that the functions do not possess a clearly visible "common shape". But a closer look at the unregistered curves also shows that there exist structurally similar curves which quite obviously exhibit some phase variation. Nevertheless, these are not the type of data that may be registered by any conventional method. Indeed, fitting a $K = 1$ dimensional model leads to unreasonable results with extreme warping functions. On the other hand, the number of local extrema of these functions varies between 1 and 3, and these random functions are of bounded shape variation. Indeed the true minimal dimension is $K = 2$, and the figure shows that $K = 2$ dimensional registration rests upon structurally simple warping functions.

The paper is organized as follows. In Section 3.2 we study the qualitative model (3.1) and discuss resulting problems of identifiability. Minimal variability of warping functions is introduced as a criterion to choose between different possible solutions. In Section 3.3 established connections to usual FPCA are studied in detail. Any suitable subspace registration should be based on a sophisticated algorithm which provides an effective solution to the fitting problem introduced by (3.1). In Section 3.4 we describe an algorithm based on nonlinear programming which works well in many applications. Section 3.5 provides applications to human growth curves and genetic data.

Supplementary material is presented in an online appendix. In Section A of this appendix all simulated examples used for illustration purposes are discussed in detail. The section also contains a Monte-Carlo simulation to verify the results of Theorem 1 in a small sample environment as well as a comparison with exiting methods. Proofs of theorems are given in Section 3.7.

3.2 Registering to low dimensional linear spaces

3.2.1 Random functions with bounded shape complexity

We consider observations consisting of a sample of smooth, at least twice continuously differentiable random functions $x_1, \ldots, x_n$ defined on a common interval, that we may take as $[0, 1]$ without losing generality. Let $\mathcal{X}$ denote the function space containing these observations, i.e. $\mathbb{P}(x_i \in \mathcal{X}) = 1$ and additionally assume that the random functions possess bounded variation, $\mathbb{E}\left(\sup_t x_i(t)^2\right) < \infty$.

For an integer m let $\mathcal{W}^m \subset \mathcal{L}^2[0,1]$ denote the Sobolev space of all smooth functions v with $v^{(m)} := D^m v \in \mathcal{L}^2[0,1]$. Our analysis is based on *warping functions* h which are elements of the space $\mathcal{H} \subset \mathcal{W}^2[0,1]$ of all smooth, strictly increasing functions such that $h(0) = 0$, $h(1) = 1$, and $h'(t) > 0$ for all t. The *functional inverse* h^{-1} with the property $h^{-1}(h(t)) = (h^{-1} \circ h)(t) = t$ for all t is uniquely defined, and the *identity warping function* $\mathcal{I}$ given by $\mathcal{I}(t) = t$ for all t acts as the unit element $\mathcal{H}$ for functional composition. Common start and end points of h simplify the problem and are fairly natural in many applications. It is possible to modify this requirement in specific situations. Similar to Kneip and Ramsay (2008) higher order smoothness assumptions may also be imposed.

There exists numerous ways of representing warping functions. Examples are linear combinations of *warplets* as introduced by Claeskens et al. (2010) or using *I-Splines* from Ramsay (1988). We follow the representation used in Ramsay and Silverman (2005), where for $w \in \mathcal{W}^1[0,1]$ a warping function is defined as

$$h_w(t) = \frac{\int_0^t exp(w(u))du}{\int_0^1 exp(w(u))du}. \tag{3.2}$$

Note that for any constant $a \in \mathbb{R}$ the functions w and $w + a$ lead to the same warping function h_w. We will thus only consider functions $w \in \mathcal{W}_0^1[0,1]$, where

$\mathcal{W}_0^1[0,1]$ is the space of all $w \in \mathcal{W}^1[0,1]$ with $\int_0^1 w(t)dt = 0$. Then (3.2) defines a bijection from $\mathcal{W}_0^1[0,1]$ onto $\mathcal{H}$. Individual warping functions $h_i \equiv h_{w_i}$, $i = 1,\ldots,n$, can then equivalently be represented by the corresponding set $w_1,\ldots,w_n$ of essentially unconstrained $\mathcal{W}_0^1[0,1]$-functions. The latter are better suited for further statistical analysis, as for example FPCA.

Registration is driven by the succession of shape features, i.e. the points where the derivative of x_i is zero. For any smooth function x one can determine the set of all points $\tau^x = \{t|x'(t) = 0\}$. If x does not possess a constant segment, i.e. if there does not exist an interval $[a,b] \subset [0,1]$, $a < b$, such that $x(t) = x(s)$ for all $t,s \in [a,b]$, then the number $q(x) = |\tau^x|$ of points in this set is finite. One can then determine the corresponding locations $0 \leq \tau_1^x < \tau_2^x < \cdots < \tau_{q(x)}^x \leq 1$ and heights $x(\tau_1^x),\ldots,x(\tau_{q(x)}^x)$. Let $Q(x) = (x(0), x(\tau_1^x),\ldots,x(\tau_{q(x)}^x), x(1))^T \in \mathbb{R}^{p(x)}$ denote the corresponding $q(x)+2$-dimensional vector of heights of shape features (including starting and end points).

For simplicity we will assume that $\mathbb{P}(x_i'(0) = 0) = 0$ as well as $\mathbb{P}(x_i'(1) = 0) = 0$ such that a.s. $\tau_l^{x_i} \in (0,1)$ for all $l = 1,\ldots,q(x_i)$. Usually registration is only applied in the context of functions x_i possessing a typical succession of shape features which can be identified in each possible sample curve. In this paper we adopt a more general point of view. Registration may be useful for functional data possessing bounded shape variation, in the sense that the number of shape features to be found in individual sample curves does not exceed a certain bound $\mathbf{q} < \infty$. More precisely, further analysis will be based on the following assumption:

Assumption 1 *There exists a* $\mathbf{q} < \infty$ *such that* $\mathbb{P}(q(x_i) \leq \mathbf{q}) = 1$, *and furthermore*
$\mathbb{P}\left(x_i''(\tau_l^{x_i}) \neq 0 \;\; \textit{for all } l = 1,\ldots,q(x_i)\right) = 1.$

The important structural restriction here is the existence of the upper bound $\mathbf{q} < \infty$. The additional requirement $\mathbb{P}(x_i''(\tau_l^{x_i}) \neq 0$ for all $l = 1,\ldots,q(x_i)) =$

1, which assumes that a.s. all $x_i(\tau_l^{x_i})$, $l = 1, \ldots, q(x_i)$ are strict local minima/maxima, is a technical condition which simplifies analysis. Functions with flat parts as well as with additional inflection points only occur in a subset of $\mathcal{X}$ which has probability zero.

Assumption 1 seems to be a very natural condition in many applications in biomedicine, technics, chemometrics, etc. In practice, functions may be observed with error, and nonparametric estimates may show random wiggles. Problems of identifying "true" shape feature will be considered in Section 3.

Note that for any continuous x and any warping function h the resulting function $y = x \circ h$ has the same vector $Q(y) = Q(x)$ of heights of local extrema. This means that the registered curves y_i in (3.4) will exhibit the same visual shape (in terms of the succession of local extrema) as the original functions x_i. But for smooth functions $Q(x_i)$ is essentially the only structural feature of x_i which is invariant against strictly monotone transformations. It is thus the driving force of identifiability of any registration procedure.

Using (3.2) a *registration procedure* can then formally be defined as a measureable mapping $\mathcal{R}$ from $\mathcal{X}$ into $\mathcal{W}_0^1[0,1]$ which assigns a warping function $h_{\mathcal{R}(x)}$ to each $x \in \mathcal{X}$.

A basic insight which provides the basis of our approach now is that random functions satisfying Assumption 1 can always be registered to a finite dimensional linear function space. As usual, we will speak of a K-dimensional linear space $\mathcal{L}_K \subset \mathcal{W}^2[0,1]$ if there exist K orthonormal functions $\gamma_1, \ldots, \gamma_K \in \mathcal{W}^2[0,1]$ such that $\mathcal{L}_K = span\{\gamma_1, \ldots, \gamma_K\}$.

Proposition 1 *Under Assumption 1*

a) For some $K \leq \mathbf{q} + 2$ there exists a registration procedure $\mathcal{R} : \mathcal{X} \to \mathcal{W}_0^1[0,1]$ and a K-dimensional linear function space $\mathcal{L}_K$ such that with $w_i := \mathcal{R}(x_i)$,

$h_i := h_{w_i}$ *we obtain*

$$\mathbb{P}(x_i \circ h_i \in \mathcal{L}_K) = 1 \tag{3.3}$$

b) For some integer K *let* $\mathcal{L}_K = span\{\gamma_1, \ldots, \gamma_K\}$ *denote a* K*-dimensional linear function space generated by smooth, twice continuously differentiable basis functions* $\gamma_1, \ldots, \gamma_j$ *on* $[0,1]$*. If* $\mathbb{P}(Q(x_i) \in \{Q(\gamma) | \gamma \in \mathcal{L}_K\}) = 1$ *there then exists a registration procedure such that* $\mathbb{P}(x_i \circ h_i \in \mathcal{L}_K) = 1$*.*

c) There exists a registration procedure such that (3.3) *holds for* $K = 1$ *if and only if* $\mathbb{P}(\mathbf{q} = q(x_i)) = 1$ *as well as* $\mathbb{P}\left(Q(x_i) = aQ(x_j) \text{ for some } a \in \mathbb{R}\right) = 1$ *for* $i \neq j$*.*

Assertion a) of the proposition follows from Proposition 1 of Kneip and Ramsey (2008). The proposition tells us that already the number of peaks and valleys to be found in each curve x_i may provide an idea about an appropriate choice of K such that there exists a registration procedure and a suitable set of orthonormal basis functions $\gamma_1, \ldots, \gamma_k$, $span\{\gamma_1, \ldots, \gamma_K\} = \mathcal{L}_K$, such that with probability 1

$$y_i(t) := x_i(h_i(t)) = \sum_{j=1}^{K} a_{ij}\gamma_j(t) \tag{3.4}$$

for all $t \in [0,1]$, where $a_{ij} = \int_0^1 x_i(h_i(t))\gamma_j(t)dt$, $j = 1, \ldots, K$. In this qualitative model only the linear subspace $\mathcal{L}_K = span\{\gamma_1, \ldots \gamma_K\}$ can be identified, while there are many different possible choices of basis functions. Our approach relies on eigenfunctions of the second moment operator defined by $M_{\mathbf{y}}(y) = \mathbb{E}(\langle y_i, y\rangle y_i)$ for $y \in L^2[0,1]$. Under (3.4) the operator $M_{\mathbf{y}}(y)$, only possesses K nonzero eigenvalues, and a suitable basis $\gamma_1, \ldots, \gamma_K$ is given by the eigenfunctions corresponding to these K leading eigenvalues.

For random functions satisfying Assumption 1 there will thus exist a **minimal dimension $\mathbf{K_0}$** such that (3.3) holds for all $K \geq \mathbf{K_0}$, while there does not exist a registration procedure leading to (3.3) for $K < \mathbf{K_0}$.

By Proposition 1c) $\mathbf{K_0} = 1$ can only hold for structurally very simple random functions with a common set of $q(x_i) = \mathbf{q}$ shape features. Obviously, $x_i(h_i(t)) = a_i\gamma(t)$ can only hold if all possible vectors $Q(x_i)$ are proportional. Any shape feature of a registered function y_i is then aligned to the location of the corresponding feature of the template γ, i.e. $\tau^{y_i} = \tau^{\gamma}$. In contrast, as already illustrated by Figure 3-1, a suitable registration to a $K \geq 2$ dimensional space will usually not go along with an alignment of peaks.

Proposition 1b) shows that a structural analysis of observed realizations x_i may provide information about a suitable dimension and possible candidate spaces $\mathcal{L}_K$. This property will be exploited in our application to yeast genes in Section 3.5.

To illustrate this point consider a simple example. Assume that with probability 1 each sample function is a smooth periodic function with period length equal to 1, and assume that in each period every curve just possesses one local maximum and one minimum. Then Assumption 1 is satisfied with $\boldsymbol{q} = 2$, and $x_i(0) = x_i(1)$ a.s. Obviously any linear combination of the three functions $\gamma_1(t) \equiv 1$, $\gamma_2(t) \equiv \sin(2\pi t)$ and $\gamma_3(t) \equiv \cos(2\pi t)$ possesses the proper functional structure, and for any x_i one can a.s. find an unique element $y_i \in \mathcal{L}_3 := span\{\gamma_1, \gamma_2, \gamma_3\}$ with $Q(x_i) = Q(y_i)$. By Proposition 1b) there thus exists a registration procedure such that $y_i(t) := x_i(h_i(t)) = a_{i1} + a_{i2}\sin(2\pi t) + a_{i3}\cos(2\pi t)$ holds a.s. for all $t \in [0, 1]$. This also implies that for such random functions we have $\mathbf{K_0} \leq 3$.

On the other hand, the example shows that a solution of (3.4) will usually not be unique, since it is easy to construct alternative, $K = 3$ dimensional candidate spaces $\mathcal{L}_3^* \neq \mathcal{L}_3$ such that $\mathbb{P}(Q(x_i) \in \{Q(\gamma)|\gamma \in \mathcal{L}_3^*\}) = 1$.

3.2.2 Identifiability

The above example also shows that there are serious issues with identifiability. Warping functions as well as the space $\mathcal{L}_K$ may not be unique. A trivial non-identifiability consists in the fact that for an arbitrary function $g \in \mathcal{H}$ (3.4) remains valid if h_i and γ_j are replaced by $h_i^* = h_i \circ g$ and $\gamma_j^* := \gamma_j \circ g$, $j = 1, \ldots, K$. This effect can be eliminated by requiring that warping functions are **standardized** such that $\bar{w}(u) = \mathbb{E}(w_i(u)) = 0$ for all $u \in [0, 1]$. Note, that this is slightly different to the usual approach where the warping functions are standardized directly, such that $\mathbb{E}(h_i(u)) = \mathbb{E}(h_{w_i}(u)) = u$.

If $\mathbf{K_0} = 1$, then it is easily seen that by requiring $\mathbb{E}(w_i(u)) = 0$ there exists a unique registration procedure and a unique γ satisfying (3.4) for $K = 1$. But standardizing can only be shown to solve problems of identifiability in the case $K = \mathbf{K_0} = 1$. For $\mathbf{K_0} \leq K \geq 2$ there may exist different sets of standardized warping functions $h_i^* \neq h_i$ and different subspaces $\mathcal{L}_K^* \neq \mathcal{L}_K$ satisfying (3.4), respectively.

Since a complete statistical analysis will require to analyze warping functions h_i in addition to registered functions y_i, parsimony suggests to use the solution where the least amount of warping is necessary. In our representation this means that the functions $w_i(u)$ should be as close as possible to 0. This introduces an additional requirement for a suitable selection of warping functions for given dimension $K \geq K_0$:

- Under all possible registration procedures leading to (3.3) and $\bar{w}(u) = 0$ choose the solution such that the mean variance $\mathbb{E}(\int_0^1 (w_i(u))^2 du)$ of $w_i = \mathcal{R}(x_i)$, is minimal.

If for a given K already the original functions are K-dimensional, i.e. $\mathcal{X} \subset \mathcal{L}_K$ for some K dimensional linear space $\mathcal{L}_K$, then of course the solution with minimal

$\mathbb{E}(\int_0^1 (w_i(u))^2 du)$ is $w_i(u) = 0$ and thus $h_i(t) = t$ for all i (i.e. no warping at all). If the original (unregistered) sample itself is not low dimensional, then our approach is to determine the linear subspace where the least amount of warping is necessary. We want to note, however, that we do not have a formal proof of whether or not the condition of minimal $\mathbb{E}(\int_0^1 (w_i(u))^2 du)$ always leads to a unique solution.

By the minimal variance criterion there will exist a trade-off between dimensionality K and complexity of warping functions. Warping functions as well as registered functions depend on K, i.e. $y_i \equiv y_{K,i}$, $w_i \equiv w_{K,i}$, and $h_i \equiv h_{w_{K,i}}$. It is easily seen that $\mathbb{E}(\int_0^1 (w_{K,i}(u))^2 du) \to 0$ as $K \to \infty$.

Analyses with $K > \mathbf{K}_0$ may be of interest for clustering. The criterion of minimal variance of w_i will then tend to incorporate additional basis functions which define centers of clusters of phase variation, where phase variation within clusters is much lower than between clusters. In applications where a $K = 1$ dimensional model yields a good approximation, results may be comparable to those to be obtained by the "k mean" method of Sangalli et al. (2010). Roughly speaking, this approach assumes that each curve belongs to one of k different clusters, where within each cluster a one dimensional registration is possible. For the example of Figure 3-1 a detailed comparison between our method and the k-means approach is given in the online appendix.

3.2.3 The estimation problem for a sample of size n.

In practice, solutions will have to be estimated from an i.i.d. sample of n functions $x_1, \ldots, x_n$. For given K the aim is then to determine warping functions h_{w_i} such that (3.4) provides a good approximation for all $i = 1, \ldots, n$. The minimal variance condition has to be replaced by its empirical analogue $V_n(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^n \int_0^1 (w_i(u))^2 du$

More precisely, for given K we will consider the following minimization prob-

lem. Let δ_{ij} denote Kronecker's delta, and let $\langle a, b\rangle = \int_0^1 a(t)b(t)dt$. Furthermore, for an arbitrary (large) $0 < d < \infty$ let $\mathcal{W}_d^1[0,1] := \{w \in \mathcal{W}_0^1[0,1] |\, sup_{u\in[0,1]}|w'(u)| \leq d\}$. Determine $\mathbf{w} = (w_1, \ldots, w_n) \in (\mathcal{W}_d^1[0,1])^n$ with $\bar{w}(u) = 0$ such that

$$S_n(\mathbf{w}, K) = \min_{(\gamma_i):\langle\gamma_i,\gamma_j\rangle=\delta_{ij}} \frac{1}{n}\sum_{i=1}^{n} ||x_i(h_{w_i}(t)) - \sum_{j}^{K} \gamma_j(t)\langle\gamma_j, x_i(h_{w_i}(t))\rangle||^2 \qquad (3.5)$$

is minimal with respect to all possible $w_1, \ldots, w_n \in \mathcal{W}_d^1[0,1]$, and such that

$$V_n(\mathbf{w}) \leq V_n(\mathbf{w}^*) \quad \text{for all } \mathbf{w}^* \in (\mathcal{W}_d^1[0,1])^n \text{ with } S_n(\mathbf{w}, K) = S_n(\mathbf{w}^*, K) \qquad (3.6)$$

If for a selected dimension K the factor model (3.4) holds exactly, then $S_n(\mathbf{w}, K) = 0$. Otherwise, one has to find the solution $S_n(\mathbf{w}, K) > 0$ with the smallest L^2-approximation error. An algorithmic implementation is described in Section 3.4.

Introducing a bound $|w_i'(u)| \leq d$ and requiring $w_i \in \mathcal{W}_d^1[0,1]$ ensures a well-defined minimization problem even if $K \leq K_0$. Weak second derivatives of the resulting warping functions are then uniformly bounded, which means that functions h_{w_i} are selected from a compact subspace of $\mathcal{W}_0^2[0,1]$. This is important for **any** norm-based minimization since $\mathcal{W}_0^2[0,1]$ is not a closed space. If $K \leq K_0$, an infimum of (3.5) may be obtained at the the boundary of the closed hull of $\mathcal{W}_0^2[0,1]$. This boundary contains functions $\tilde{h}$ with jumps and monotone segments such that $Q(x \circ \tilde{h}) \neq Q(x)$. A possible tendency towards extreme warping function is known as the "pinching problem" (compare Ramsay and Li (1998)).

Figures 3-1 and 3-2 both show samples of random functions where (3.4) holds with $K = 2$. For the data of Figure 3-1 a $K = 1$ dimensional model yields a reasonable approximation. For suitable w_i in the interior of $\mathcal{W}_d^1[0,1]$ we obtain a small, although nonzero, value of $S_n(\mathbf{w}, 1)$. The situation is very different for the data of Figure 3-2. Fitting a $K = 1$ dimensional model leads to strong pinching,

which indicates that in this case $S_n(\mathbf{w}, 1)$ does not possess a local minimum for $\mathbf{w}$ in the interior of $\mathcal{W}_0^1[0,1]$.

3.3 Registration and the analysis of functional data

3.3.1 Registration versus FPCA

For analyzing functional data as displayed in Figure 3-2 most researcher would probably rely on standard functional principal component analysis (FPCA). But as illustrated by the figure an analysis based on representation (3.4) may lead to substantially different results. In order to see the point first recall that FPCA is based on the Karhunen-Loève decomposition

$$x_i(t) = \mu(t) + \sum_{j=1}^{\infty} b_{ij} g_j(t). \tag{3.7}$$

Here, $\mu = \mathbb{E}(x_i)$ and for $j = 1, 2, \ldots$ g_j is an eigenfunction corresponding to the $j-th$ largest eigenvalue λ_j of the covariance operator Γ, defined by $(\Gamma v)(t) = \int_0^1 \sigma(t,s)v(s)ds$, where $\sigma(t,s) := \mathbb{E}((x_i(t) - \mu(t))(x_i(s) - \mu(s))$ is the covariance function. Moreover, the coefficients $b_{ij} = \langle x_i - \mu, g_j \rangle$ are uncorrelated for different j, and $Var(b_{ij}) = \lambda_j$.

FPCA then relies on a finite dimensional approximation $x_i(t) \approx \tilde{x}_{\kappa,i}(t) := \mu(t) + \sum_{j=1}^{\kappa} b_{ij} g_j(t)$ for some suitable κ. But note that the functions $g_1, g_2, \ldots$ in (3.7) only depend on the covariance $\sigma(t,s)$. The exact distribution of x_i, and in particular shape features of possible realizations, will however additionally depend on higher order moments of the scores b_{ij}. Many structurally very different random processes may possess the same functional principal components $g_1, \ldots, g_\kappa$.

This effect is easily illustrated by an extreme, but analytically simple example. Consider classes of random functions with mean $\mu(t) \equiv 0$ and $\sigma(t,s) = \min\{t,s\}$.

Then $\lambda_j = \frac{1}{(j-0.5)^2\pi^2}$, $j = 1, 2, \ldots$, while corresponding orthonormal eigenfunctions are given by $g_j(t) = \sin((j - 1/2)\pi t)$, $j = 1, 2, \ldots$.

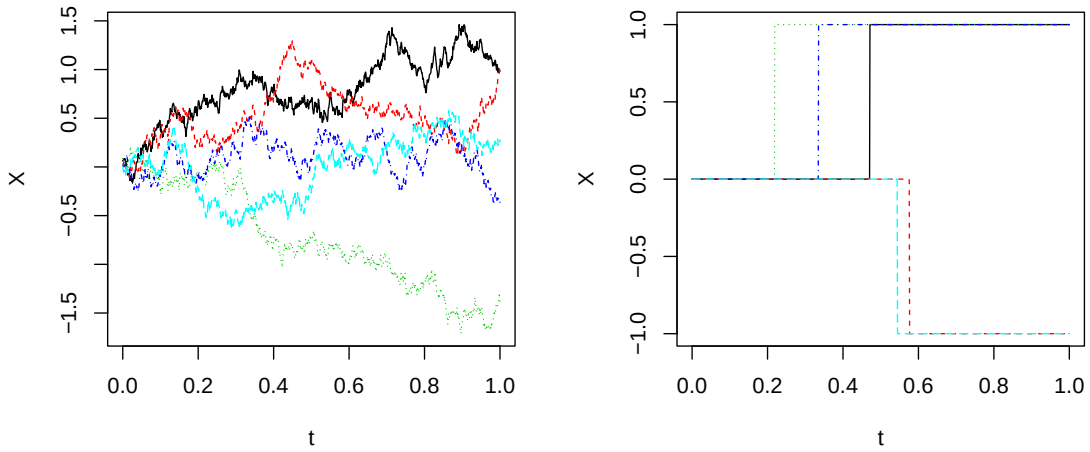


Figure 3-3: Sample of 5 random curves of the standard Brownian motion X^* (left) and of the jump process X (right)

If additionally the scores are independent normal variables, $b_{ij} \sim N(0, \lambda_j)$, then the resulting process x_i^* is a standard Brownian motion on $[0, 1]$. Possible sample path are displayed in left part of in Figure 3-3. The Brownian motion does not satisfy Assumption 1, for any $\mathbf{q} < \infty$ we then obtain $\mathbb{P}(q(x_i) > \mathbf{q}) > 0$. Any attempt of registration is obviously futile.

Sample paths of a very different process x_i are displayed in the right part of Figure 3-3. This process is defined as follows: For two independent random variables T_i and A_i, where $T_i \sim U(0, 1)$ and A_i is a binary variable with $\mathbb{P}(A_i = 1) = \mathbb{P}(A_i = -1) = \frac{1}{2}$, we have $x_i(t) = 0$ for $0 \leq t < T_i$ and $x_i(t) = A_i$ for $T_i \leq t \leq 1$. This also implies $\mu(t) = \mathbb{E}(x_i(t)) = 0$ and $\mathbb{E}\left(x_i(t)x_i(s)\right) = \min\{t, s\}$, and in view of (3.7) the only difference to a Brownian motion consists in a complicated, highly non-normal distribution of scores b_{ij}. Quite obviously, even for large κ FPCA will not provide a reasonable approximation of these function x_i. At the same time phase variation obviously constitutes the main source of variability of x_i. Even though x_i does not satisfy our smoothness condition, (3.4) holds for

$K = 1$, $x_i(h_i(t)) = a_i\gamma(t)$, $i = 1, \ldots, n$, where $a_i \in \{-1, 1\}$, $\gamma(t) = 0$ for $0 \leq t < \frac{1}{2}$ and $\gamma(t) = 1$ for $\frac{1}{2} \leq t \leq 1$, while $h_i(\frac{1}{2}) = T_i$. Since γ consist of two monotone segments, only the values $h_i(0) = 0$, $h_i(\frac{1}{2}) = T_i$, and $h_i(1) = 1$ are fixed, all other values of $h_i(t)$ are arbitrary. But for any reasonable interpolation scheme, $\{h_i\}$ will be a simple, one-dimensional family of functions.

In general, a stochastic process x_i is Gaussian if and only if the scores are independent normal variables, $b_{ij} \sim N(0, \lambda_j)$, $j = 1, 2, \ldots$. The following proposition shows that, unless $\mathcal{X}$ is already finite dimensional, Gaussian random functions cannot possess bounded shape variation.

Proposition 2 *Let $x_i^* \in \mathcal{W}^2[0,1]$, $m \geq 1$, be a Gaussian random function with bounded covariance operator Γ. If Γ has infinitely many non-zero eigenvalues, then $\mathbb{P}(q(x_i^*) > \mathbf{q}) > 0$ for any $\mathbf{q} < \infty$.*

The proposition essentially tells us that mean and covariance structure of a random process does not reflect shape information. For any process x_i with bounded shape variation and nontrivial covariance operator Γ there will exists a Gaussian process x_i^* with the same mean and the same covariance operator which does not satisfy Assumption 1. Both processes share structurally identical Karhunen-Loève decompositions (3.7) which only differ in higher moments of the distribution of scores. Here (3.4) offers a generally applicable nonlinear decomposition which provides an alternative to FPCA and explicitly exploits existing shape features.

- For $\kappa > 0$ FPCA is based on a variance decomposition which splits x_i into $\tilde{x}_{\kappa,i}$ and a residual function $r_{\kappa,i} = x_i - \tilde{x}_{\kappa,i} = \sum_{j=\kappa+1}^{\infty} b_{ij} g_j$. Shape features of x_i depend on the interplay between $\tilde{x}_{\kappa,i}$ and $r_{\kappa,i}$, and $\tilde{x}_{\kappa,i}$ alone may not properly reflect the shape of x_i.

- For $K \geq \mathbf{K}_0$ (3.4) decomposes x_i into $y_{K,i} := \sum_{j=1}^{K} a_{ij}\gamma_j(t)$ and a function $w_i \equiv w_{K,i}$. The functions $y_{K,i}$ possess the same sequences of shape features

as x_i, while remaining variability of x_i is explained by nonlinear "phase variation" quantified by $h_i \equiv h_{w_{K,i}}$. We then have $x_i = y_{K,i} \circ h^{-1}_{w_{K,i}}$ instead of $x_i = \tilde{x}_{\kappa,i} + r_{\kappa,i}$ for FPCA.

FPCA and a decomposition based on (3.4) will only coincide if the random functions x_i themselves are already low dimensional, i.e. if $\lambda_\kappa > 0$ while $\lambda_{\kappa+1} = 0$ for some $\kappa \geq 1$. Then $x_i = \mu(t) + \sum_{j=1}^{\kappa} b_{ij} g_j(t)$. If $\mu \in \text{span}\{g_1, \ldots, g_\kappa\}$, then (3.4) holds with $\kappa = K$, $\mathcal{L}_K = span\{g_1, \ldots, g_K\}$ and $w_i(t) \equiv 0$, $h_i(t) = t$. If $\mu \notin \text{span}\{g_1, \ldots, g_\kappa\}$, then (3.4) holds with $K \leq \kappa + 1$, $\mathcal{L}_K = span\{\mu, g_1, \ldots, g_K\}$ and $w_i(t) \equiv 0$, $h_i(t) = t$.

When using the decomposition based on (3.4) warping functions $h_{w_{K,i}}$ quantify phase variation. This may reflect an important (nonlinear) source of variability of x_i, and a serious statistical analysis will also have to extract information contained in the warping functions. Recall that the $w_{K,i}$ are essentially unconstrained random functions, which in turn define a corresponding covariance operator Γ_w. A simple way to extract information on variability of w_i is thus an FPCA using the eigenfunctions $\varphi_1, \ldots, \varphi_L$ corresponding to the L largest eigenvalues of Γ_w. The functions $w_{K,i}$ are then represented by

$$\tilde{w}_{K,L;i}(t) = \sum_{j=1}^{L} \vartheta_{ij} \varphi_j(t), \tag{3.8}$$

where $\vartheta_{ij} = \langle w_{K,i}, \varphi_j \rangle$. Recall that, by definition, $w_{K,i}$ has mean zero. The part of the warping function "explained" by FPCA is then given by $h_{\tilde{w}_{K,L;i}}$. For selected K, L the quality of approximating x_i by $y_{K,i} \circ h^{-1}_{\tilde{w}_{K,L;i}}$ may then be measured by

$$R_{K,L} = \frac{\mathbb{E}\left(\int_0^1 (x_i(t) - y_{K,i}(h^{-1}_{\tilde{w}_{K,L;i}}(t)))^2 dt\right)}{\mathbb{E}(\int_0^1 x_i(t)^2 dt)}. \tag{3.9}$$

This may also be compared to the results of an FPCA of x_i by computing $R_\kappa =$

$\frac{\mathbb{E}(\int_0^1 r_{\kappa,i}^2 dt)}{\mathbb{E}(\int_0^1 x_i(t)^2 dt)}$.

3.3.2 Identifying K_0 from noisy observations

In practice, the functions x_i will often not be directly observed, but one will have to deal with discrete, noisy observations contaminated with some error. Nonparametric estimates $\hat{x}_i$ of x_i may then show some random wiggles, and $S_n(\mathbf{w}, K)$ may be nonzero even for $K \geq K_0$.

In the following we will only consider a simple, standard error model: For T equidistant design points $t_1, \ldots, t_T \in [0,1]$ there are noisy observations Y_{il} such that

$$Y_{il} = x_i(t_l) + \epsilon_{il}, \quad i = 1, \ldots, n; l = 1, \ldots, T, \tag{3.10}$$

for i.i.d. zero mean error terms ϵ_{il} with finite variance $\sigma^2 > 0$ and $\mathbb{E}(\epsilon_{il}^4) < \infty$.

Assume that for all $i = 1, \ldots, n$ estimates $\hat{x}_i$ of x_i are determined by local linear estimators with bandwidth b and a continuous second order kernel function $\mathcal{K}$, where $\mathcal{K}$ has compact support $[-1,1]$ and $V(\mathcal{K}) \equiv \int_{-1}^1 \mathcal{K}(x)^2 dx < \infty$.

Registration now has to be based on these estimates, and for any $\mathbf{w} = (w_1, \ldots, w_n) \in (\mathcal{W}_d^1[0,1])^n$ let $\hat{S}_n(\mathbf{w}, K)$ be defined by (3.5) when replacing x_i by $\hat{x}_i$. For given $K \geq 1$ let $\hat{\mathbf{w}}_K = (\hat{w}_{K,1}, \ldots, \hat{w}_{K,n})$ denote a minimizer of $\hat{S}_n(\mathbf{w}, K)$ under the side condition (3.6). We will assume that the constant d in (3.5) is chosen such that (3.4) holds for warping functions $h_i \equiv h_{w_i}$ with $\sup_{u \in [0,1]} |w_i'(u)| \leq d$ a.s.

For given $K \geq 1$ let $\hat{h}_{K,i} := h_{\hat{w}_{K,i}}$, $i = 1, \ldots, n$ be the resulting warping functions, and let $\hat{a}_{K;ij}$ and $\hat{\gamma}_{K,j}$ denote the corresponding estimates of coefficients and basis functions in (3.4). Some basic consistency results are now given by the following theorem.

Theorem 1 *Under our setup additionally assume that* $\mathbb{E}(\sup_{t\in[0,1]} Dx_i''(t)^2) < \infty$. *We then obtain as* $n, T \to \infty$ *and* $b \to 0$, $Tb\log T \to \infty$.

a) There exists some $c > 0$ *such that for all* $K < \mathbf{K}_0$

$$\mathbb{P}\left(\hat{S}_n(\hat{\mathbf{w}}, K) \geq c\right) \to 1 \quad \textit{for all } K < \mathbf{K}_0, \tag{3.11}$$

while $\hat{S}_n(\hat{\mathbf{w}}, K) = O_P(b^4 + \frac{1}{Tb})$ *for all* $K \geq \mathbf{K}_0$.

b) If $b = o(T^{-1/5})$, *then for any constant* $A > 1$

$$\mathbb{P}\left(\hat{S}_n(\hat{\mathbf{w}}, K) \leq A\frac{\sigma^2 V(\mathcal{K})}{Tb}\right) \to 1 \quad \textit{for all } K \geq \mathbf{K}_0 \tag{3.12}$$

c) For all $K \geq \mathbf{K}_0$

$$\int_0^1 (x_i(\hat{h}_{K,i}(t)) - \sum_{j=1}^{K} \hat{a}_{K;ij}\hat{\gamma}_{K,j}(t))^2 dt = O_P(b^4 + \frac{1}{Tb}) \tag{3.13}$$

It is well-known the error variance σ^2 can be estimated consistently by nonparametric procedures. Theorem 1b) implies that the minimal dimension $\mathbf{K}_0$ is asymptotically identifiable by selecting the smallest K with $\hat{S}_n(\hat{\mathbf{w}}, K) \leq A\frac{\hat{\sigma}^2 V(\mathcal{K})}{Tb}$ for some $A > 1$ and a suitable nonparametric variance estimate $\hat{\sigma}^2$.

3.4 The algorithm

3.4.1 Implementation for fixed dimension K

When considering the minimization problem defined by (3.5) and (3.6) for fixed K, the values of $V_n(\mathbf{w})$ and $S_n(\mathbf{w}, K)$ are in interdependency with each other. The minimizing solution for $V_n(\mathbf{w})$ is given by $w_i(u) = 0$ for all i while in general this

does not minimize $S_n(\mathbf{w}, K)$. Related multi-objective minimization problems are important in many scientific fields, for example in economics and engineering, and are well studied in the literature (Ehrgott, 2000). So called Pareto optimal solutions can be determined using *weighted sum scalarization*. We follow this concept by replacing (3.5) and (3.6) by a single minimization problem in dependence of a parameter $1 > \nu > 0$: Determine $\mathbf{w} = (w_1, \ldots, w_n) \in L^2([0,1])^n$ such that

$$S_P^*(\mathbf{w}, K) := [(1-\nu)\frac{S_n(\mathbf{w}, K)}{S_0} + \nu V_n(\mathbf{w}))] \tag{3.14}$$

is minimal. Dividing $S_n(\mathbf{w}, K)$ by $S_0 := \frac{1}{N}\sum_{i=1}^{N} ||x_i(t)||^2$ eliminates possible scaling effects. In view of (3.5) and (3.6) our main interest is to minimize $S_n(\mathbf{w}, K)$, while reduction of phase variance is only secondary. This means that usually ν has to be very small. In the simulations and applications presented in this paper $\nu = 0.001$ was used throughout.

The parameter ν serves two purposes. The first one is of course to choose the solution with minimal variance of w_i among several possible candidate spaces. In this context a very small value of ν is clearly appropriate. The second role is regularization, i.e. excluding boundary solutions. As explained below, the functions w_i are approximated by spline functions. Thus $\nu > 0$ implicitly imposes bounds for the values $|w_i|$ and $|w_i'|$ of a possible solution of (3.14), since very large values can only be achieved by very large spline coefficients which in turn lead to large $V_n(\mathbf{w})$. In this context $\nu = 0.001$ still allows for fairly extreme warping functions, as can been seen for the $K = 1$ dimensional fit in Figure 2. Increasing ν would lead to smoother warping functions, but not to a better registration since a $K = 1$ dimensional model simply does not provide any reasonable approximation of these data. From this point of view a small value of ν might also be advantageous, since an inappropriate choice of K will then be highly visible. Our recommended choice

$\nu = 0.001$ generally delivered good results for further analysis. An automatic selection of $\nu > 0$ is an open task for further research.

We rely on nonlinear programming algorithms to determine a solution of (3.14) over a class of smooth warping functions. These algorithms are designed to minimize an objective function over a finite set of variables, but not over functions. To overcome this issue it is reasonable to require that for some $m > 0$, $w_i(u)$ is sufficient smooth such that $w_i(u) \in \mathcal{S}_{4,m+4}$ where $\mathcal{S}_{4,m+4}$ is the *B-spline space* of order 4 with $m + 4$ equidistant knots. To fasten up the computation we use a representation with m orthogonal splines $B_j(u),\ j = 1, \ldots, m$ based on cubic B-splines as described by Mason et al. (1993) which are additional normalized such that $\int_0^1 B_j(u)^2 du = 1$. This leads to a representation based on coefficents $c_i := (c_{i1}, \ldots, c_{im})$ given by $w_i(u) = \sum_1^m B_j(u)c_{ij}$. A warping function is therewith describable by

$$h(t, c_i) \equiv h_{w_i}(t) = \int_0^t exp(\sum_{j=1}^m B_j(u)c_{ij})du / \int_0^1 exp(\sum_{j=1}^m B_j(u)c_{ij})du. \qquad (3.15)$$

The choice of m determines the smoothness of the $w(u)$ functions. In the simulations choosing $m = 7$ leads to warping functions sufficient close to the true functions. From our experience even if the algorithm is used to process real data the gain by using a bigger m compared to the additional computation time can mostly be considered as negligible.

Since by construction $B_1, \ldots, B_m$ are orthonormal we get

$$V_n(\mathbf{c}) \equiv V_n(\mathbf{w}) = \frac{1}{mn} \sum_{i,j} (c_{ij}^2). \qquad (3.16)$$

To ensure that $\bar{w}(u) = 0$ and $\int_0^1 w_i(u)du = 0$ we impose the condition that $\mathbf{c} := (c_{ij})_{i=1,\ldots,n;j=1,\ldots,m}$ is such that $\frac{1}{n}\sum_{i=1}^n c_{ij} = 0,\ \forall j = 1, \ldots, m$ and $\frac{1}{m}\sum_{j=1}^m c_{ij} \int_0^1 B_j(u)du =$

$0,\ \forall i = 1, \ldots, n$.

Recall that $S_n(\mathbf{c}, K) \equiv S_n(\mathbf{w}, K) = \sum_{i=K+1}^{\infty} \lambda_i(\mathbf{c})$ with $\lambda(\mathbf{c})$ being ordered eigenvalues of the second-moment operator $M(x) = \frac{1}{n}\sum_{i=1}^{n} \langle x_i(h(c_i)), x\rangle x_i(h(c_i))$. To estimate these eigenvalues we use the duality relation from Härdle and Simar (2012) where eigenvalues can be computed very fast if n is small. The duality relation states that the eigenvalues $\lambda(\mathbf{c})$ corresponds to the eigenvalues of the $n \times n$ matrix $\mathbf{D}$ with elements

$$D_{ij}(\mathbf{c}) = \frac{1}{n}\int_0^1 x_i(h(t, c_i))x_j(h(t, c_j))dt, \quad i, j = 1, \ldots, n. \tag{3.17}$$

In our algorithm the integral is approximated by Riemann sums. If only discrete observations are available the curves are interpolated using linear interpolation.

The final task is then to determine the vector $\mathbf{c}$ of size nm satisfying the above conditions which minimizes

$$(1-\nu)\frac{\sum_{l=K+1}^{\infty} \lambda_l(\mathbf{c})}{S_0} + \nu V_n(\mathbf{c}). \tag{3.18}$$

This minimization problem is solved by using the "newuoa" algorithm developed by Powell (2006) and implented in "R" by Bates et al. (2014) which is able to handle a large amount of variables in endurable time. The algorithm is iterative and requires initial values for c_{ij}. Here, our standard choice is to start with $c_{ij} = 0$ for all i, j which corresponds to $h_{w_i}(t) = t$. If a complicated warping problem is given as is the case with simulation 3.6.3, it is useful to use different starting values as described in Section 3.4.2.

The optimal warping functions $h_i(t) = h(c_i, t)$ are computed with (3.15) using the final values of $\mathbf{c}$ minimizing (3.18). To get the optimal template decomposition we again rely to the duality relation. Let $\theta_j,\ j = 1, \ldots, n$ be the ordered

eigenvectors of $D(\mathbf{c})$ then

$$\gamma_j(t) = \frac{1}{\sqrt{n\lambda_j(\mathbf{c})}} \sum_{i=1}^{n} \theta_{ij} x_i(h(c_i, t)) \tag{3.19}$$

$$a_{ij} = \theta_{ij} \sqrt{n\lambda_j(\mathbf{c})}. \tag{3.20}$$

If only discrete observations are available, again linear interpolation of x_i is used to get continuous $\gamma_j(t)$ if necessary.

3.4.2 Determining a suitable dimension K

In practice, the minimal dimension $\mathbf{K}_0$ of model (3.4) is usually unknown. But it follows from Proposition 1 and the discussion of Section 3.2 that usually the functional structure of the sample curves provide some information about a *maximal* value K_{max} of K. Additionally note that in the above algorithm the computational complexity of the nonlinear minimization problem (3.18) only depends on the number of warping coefficients c_{ij} but not on the value of K. We thus propose to proceed as follows:

1) Determine a maximal dimension $K_{max} \leq \mathbf{p}$.

2) For $K = K_{max}, K_{max}-1, K_{max}-2, \ldots, 1$ successively calculate K-dimensional approximations of (3.4) by solving (3.18) for the given K. If $K = K_{max}$ choose $c_{ij} = 0$ for all i, j as initial values for the nonlinear minimization algorithm, while for $K < K_{max}$ use the coefficients of the $K+1$-dimensional solution as initial values for the algorithm.

For example, to handle the growth data in Section 3.5.1 $K_{max} = 4$ was used. We want to emphasize that recursive updating of starting values for c_{ij} in step 2) is of particular importance. Iterative, nonlinear minimization algorithms sometimes

tend to get stuck in local minima if initial values are too far from actual solutions. Complexity of warping functions decreases as K increases, and hence $c_{ij} \equiv 0$, i.e. $h_{w_i}(t) = t$, may be a good starting point for $K = K_{max}$. This will not necessarily be true for $K \ll K_{max}$, but then the warping functions obtained from a $K+1$-dimensional approximation may serve as a reasonable first guess. Therefore even in the case of a pre-specified, fixed dimension K accuracy of numerical results may improve when following step 2) until the desired dimension K is reached.

Based on steps 1) and 2) one may then determine a minimal dimension by determining the smallest K such that $S_n(\mathbf{w}, K)$ is sufficiently close to zero. In practice data will usually consist of discrete observations contaminated with some type of error, and $S_n(\mathbf{w}, K) > 0$ even if the true functions x_i satisfy (3.4). When assuming the simple error model of Section 3.3.2, we may estimate $\mathbf{K}_0$ by choosing the smallest dimension K such that $\hat{S}_n(\hat{\mathbf{w}}, K) \leq A\frac{\hat{\sigma}^2 V(\mathcal{K})}{Tb}$ for some $A > 1$. For more complicated error models the criterion may be modified accordingly.

We want to emphasize, however, that even for $K < \mathbf{K}_0$ a K-dimensional approximation may be useful for further analysis if $S_n(\hat{\mathbf{w}}, K)$ attains small, although nonzero value. This is, for example, the case for the data of Figure 3-1, where $\mathbf{K}_0 = 2$ but the mean of the registered functions for the $K = 1$ dimensional fit provides a reasonable "structural mean" of the sample. As already mentioned above, a completely inappropriate choice of K will often be highly visible (see Figure 3-2).

3.5 Applications

3.5.1 Berkley Growth Data

The well-known Berkeley growth data (Tuddenham and Snyder (1954)) contain height measurements for 93 children (38 boys and 54 girls), with 31 measurements

taken over a time span of 18 years. In this section the growth acceleration functions (second derivatives) are analyzed. The curves are estimated non parametrically using a monotone smoothing procedure as described in Ramsay and Silverman (2005). Growth over the first years of life can be considered as unstructured, thus the first two and a half years are exclude after smoothing for further analyze.

Human growth exhibits considerable phase variation, and in a number of previous studies different registration procedures have been applied to analyze growth velocity functions (see, for example, Sangalli et al. (2010) or Srivastava et al. (2011)). We use the Berkeley growth data in order to show exemplary how to perform statistics using our method. In particular we will present a method for classifying growth curves with respect to sex. From biology it is known that growth features of boys and girls differ in timing and amplitude. These preconditions are comparable to the data of Figure 3-1 where two groups with slightly different structure were simulated.

For comparison a registration with $K = 1$ and $K = 2$ is carried out. In Figure 3-4 it can be observed that by going from $K = 2$ to $K = 1$ the information about sex of a child seems to move from the amplitude to the warping space. When choosing $K = 1$, the pubertal growth spurts of boys and girls are matched at one single peak, while with $K = 2$ the basis is automatically chosen such that the main peaks of boys and girls are separated. The opposite is true for the warping functions. With $K = 2$ the resulting warping curves look similar for most boys and girls, while for $K = 1$ warping functions of girls are usually below those of the boys.

The following statistical analysis relies on the concepts developed at the end of Section 3.3.1. The idea is to determine parsimonious models with high explanatory power by using few functional components which are either taken from amplitude functions y_i or warping functions w_i. We want to note that Poss and Wagner

(2014) where able to improve their analysis of juggling curves by including an FPCA of w in their model.

To classifying growth curves we rely on a logit approach. By Figure 3-4 we deduce that there is a interdependency between the representation of information in the warping and the amplitude space. To take this into account a logit model is fitted in dependence of individual scores obtained from the decompositions of the registered curves y_i and warping functions w_i.

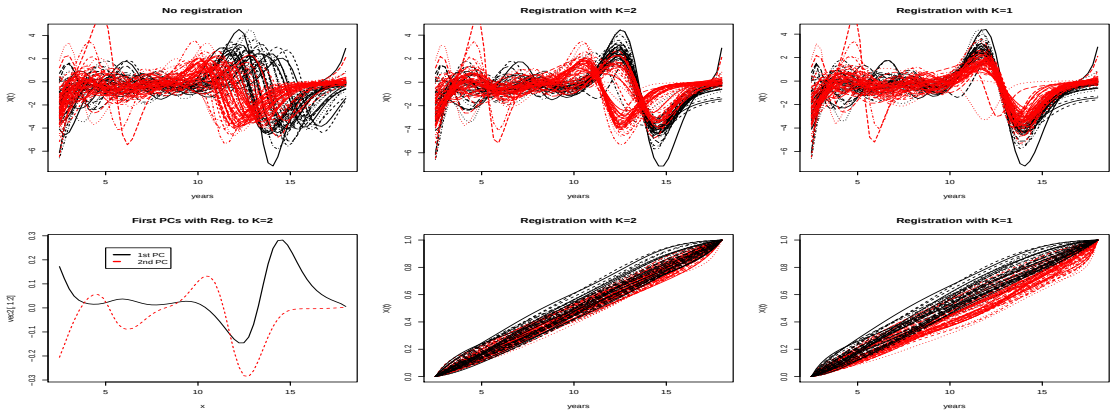


Figure 3-4: The upper left figures shows smoothed second derivative of the observed unregistered curves with girls colored red and boys black. The upper middle and upper right figure shows a registration using $K = 2$ and $K = 1$ accordingly while the figures beneath the corresponding warping functions. The lower left figure exhibits the fist two components $\hat{\gamma}_{K,1}, \hat{\gamma}_{K,2}$ of an decomposition of the registered curves with $K = 2$. The main puberty growth peak is clearly visible.

More precisely, we apply our method with $K = 1$ and $K = 2$ to the second derivative $x_i(t)$ of the growth curves to get estimators $\hat{w}_{i,K}$ and $\hat{y}_{i,K}$. For different values of $\tilde{K}$ and L, we then proceed as follows in order to fit a logit model as introduced by Berkson (1944):

- Code outcome binary, $Z_i = 1$ (girls) $Z_i = 0$ (boys)
- Decomposition of $\hat{y}_{i,K}(t) \approx \hat{y}_{K,\tilde{K};i}(t) = \sum_{j=1}^{\tilde{K}} \hat{a}_{K,ij}\hat{\gamma}_{K,j}(t)$ and $\hat{w}_{i,K}(u) \approx \hat{w}_{K,L;i}(t) = \sum_{j=1}^{L} \hat{\vartheta}_{K,ij}\hat{\varphi}_{K,j}(u)$

- Fit the logit model

$$P(Z_i = 1, K, L, \tilde{K}) = \frac{exp(f_{K,L,\tilde{K}}(i))}{1 + exp(f_{K,L,\tilde{K}}(i))}, \tag{3.21}$$

$$f_{K,L,\tilde{K}}(i) = \theta_0 + \sum_{j=1}^{\tilde{K}} \hat{a}_{K,ij}\theta_j + \sum_{j=\tilde{K}+1}^{\tilde{K}+L} \hat{\vartheta}_{K,i(j-\tilde{K})}\theta_j \tag{3.22}$$

In the presence of noisy observations or if $K < \mathbf{K}_0$ a decomposition based on (3.4) with $\hat{y}_{K,\tilde{K};i}(t)$ or an approximation of the registered curves using FPCA with $\hat{y}_{i,K}(t) \approx \hat{y}_{K,\kappa,i}(t) = \hat{\mu}(t) + \sum_{j=1}^{\kappa} \hat{b}_{ij}\hat{g}_j(t)$ can result in different outcomes. For comparison we therefore also include models based on FPCA of $\hat{y}_{i,K}(t)$. Scores of the principal components are then used instead of $\hat{a}_{K,ij}$.

To evaluate the performance of the different approaches we rely on cross-validated prediction errors. Separately for each child i, we determine $P_{-i}(Z_i = 1, K, L, \tilde{K})$ by fitting the respective model to the remaining observations. Let $\hat{Z}_i = 1$ if and only if $P_{-i}(Z_i = 1, K, L, \tilde{K}) \geq 1/2$, then the percentage of correct assignments is given by $COR_{K,L,\tilde{K}} = \frac{1}{n}\sum_{i=1}^{n} I(Z_i = \hat{Z}_i)$ while the mean squared prediction error $MSPE_{K,L,\tilde{K}} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - P_{-i}(Z_i = 1, K, L, \tilde{K}))^2$.

In order to quantify the quality in approximating the original functions x_i for different choices of L and $\tilde{K} = \kappa$ we rely (3.9), but cross-validation is additionally applied to minimize the influence of random fluctuations. We therefore compute $R_{K,L,\tilde{K}} := (\sum_{i=1}^{n} \int h'_{w_{K,L;i}}(t)(x_i(h_{w_{K,L;i}}(t)) - y_{K,\tilde{K};-i}(t))^2 dt)/(\sum_{i=1}^{n} \int x_i(t)^2 dt)$. Here, $y_{K,\tilde{K};-i}(t)$ is based on the basis functions (3.4) when leaving the i-th curve out, while $y_{K,\kappa;-i}(t)$ is based on mean and principal components when leaving the i-th curve out.

Table 3.1 shows that already a model based on the first component of the unregistered curves provide a fairly accurate classification, however a better classification is possible when registering to $K = 2$. The ability of the unregistered

Table 3.1: Model comparison. $L = 0$ is defined such that the sum in (3.22) vanishes. It is visible by the bold expressions that the $K = 2$ model is better suited to identify the groups using one or two variables, while the $K = 1$ model gives the best low dimensional representation using two or three components. For comparison a standard FPCA with $\kappa = \tilde{K}$ for registered and unregistered curves is included.

					FPCA of registered curves with $\kappa = \tilde{K}$						Decomposition based on (3.4)					
		No registration			$K = 1$			$K = 2$			$K = 1$			$K = 2$		
$\tilde{K}$	L	$MSPE$	COR	R	$MSPE$	COR	R	$MSPE$	COR	R	$MSPE$	COR	R	$MSPE$	COR	R
3	0	0.111	0.828	0.205	0.123	0.817		0.099	0.828		0.131	0.806		0.091	0.849	
2	1				0.086	0.871	**0.191**	0.095	0.882	0.201	0.087	0.871	0.206	0.085	0.892	0.250
1	2				0.141	0.796	0.267	0.092	0.903	0.260	0.090	0.860	0.258	0.082	0.892	0.496
2	0	0.106	0.839	0.260	0.120	0.849		0.105	0.860		0.127	0.839		0.101	0.839	
1	1				0.136	0.806	0.273	0.089	**0.903**	0.270	0.086	0.882	0.257	**0.080**	**0.903**	0.506
1	0	0.109	0.839	0.444	0.254	0.581		0.101	0.860		0.152	0.806		0.099	0.849	

model to identify the groups does not improve using more components. The logit model applied to the FPCA of the unregistered curves is always inferior to both registered models based on (3.4) using $L > 0$, even if fewer variables are considered.

When considering approximation of x_i, with rising κ the explanatory power of the unregistered FPCA quite obviously improves. Approximation qualities of unregistered FPCA with $\kappa = 2$ and registration with $K = \tilde{K} = 1$, $L = 1$ are almost identical, while using 3 components an FPCA of the registered curves with $\tilde{K} = 2, L = 1$ slightly outperforms the unregistered FPCA.

3.5.2 Yeast Genes

A yeast cell contains approximately 6000 genes. To save energy depending on the actual task only a few genes are active. Activations are not directly observable, but one can measure RNA or protein related to a specific gene over time. Different biological methods are then used to identify the activity of a certain gene.

The data used in this section comes from the α factor–based synchronization experiment conducted by Spellman et al. (1998), who measured the gene expression of all 6178 genes of a yeast cell during two cell cycles. The experiment lasted 2 hours where time series of cDNA micro-arrays were gathered over 18 equally space time points. Two cycles were observed because this allows to identify the active genes

due to the presence of periodicity. Curves belonging to active genes are suspected to have one peak and one valley within in each period. We follow the approach of Zhao et al. (2004) by discarding all times series with missing observations, which leaves 4489 genes.

Using techniques based on Fourier transforms, Spellman et al. (1998) identifies 612 out of these 4489 genes as being periodic and thus active during the cell division. Each of the selected genes is assigned to one of five so called "'phase groups"' termed $G1, S, G2, M$, and $M/G1$, which possess important substantial interpretation (in the provided data file the actual groups are named slightly differently and given by $G1, M/G1, G2/M, S/G2, S$). Assignment is based on data analytic tools together with some biological information.

In the following we will concentrate on two important questions: identification of active genes due to periodicity, and classification into phase groups. As already indicated by Zhao et al. (2004), the identification provided by Spellman et al. (1998) is not completely convincing, since some of the 612 genes are clearly not periodic. Analysis is complicated by the fact that such micro-array data contains a large amount of noise.

Using functional data analysis a statistical approach of identifying periodic genes is given by Zhao et al. (2004). In a first step observations are rescaled and considered as functions on the interval $[0, 4\pi]$. A Fourier transform is done, keeping only even frequencies, and FPCA with $\kappa = 2$ is used for dimension reduction. Principal components look much like sin and cos functions. This motivates the final step of the analysis which consists in determining the parameters minimizing the L_2- distances between $x_i(t)$ and $\beta_{i1}sin(t) + \beta_{i2}cos(t)$. The resulting coefficients are then intended to measure the periodicity. The idea is, that if $||\beta_i||$ is close to zero there is no periodicity present. To identify periodic curves due to the size of the coefficients without applying any warping bears the risk to select the wrong

curves since high coefficients can have several origins. Thus, in the presented results it is visible that still some non-periodic curves are chosen by the method.

Based on a pre-selected subset of 90 genes with clearly periodic trajectories, Leng and Müller (2006a) and Leng and Müller (2006b) study classification of active genes into the five groups mentioned above. Division into phase groups traditionally aims to reflect differences in the time ordering of the dynamics of gene coefficients with the same genetic pathway. Leng and Müller (2006b) show that time ordering may be quantified by global time shifts. These time shifts are estimated by minimizing suitably standardized pairwise L^2 distances between shifted curves.

Our approach now consists in a registration-based structural analysis of the data. Trajectories may be represented by functions on $[0, 4\pi]$. Existing work assumes that active genes correspond to 2π-periodic functions centered around 0, possessing one peak and one valley within each period. More precisely, in the notation of Section 3.2.1 one may assume that (up to error) *shape features* of such functions satisfy $Q(x_i) \in \{Q(\alpha_1 \sin + \alpha_2 \cos) |\ \alpha_1, \alpha_2 \in \mathbb{R}\}$. On the other hand, there is no substantial reason to expect that trajectories can be exactly described via sin and cos functions. But Proposition 1b) then tells us that there then exists warping functions h_i such that

$$x_i(h_i(t)) = a_{i1} sin(t) + a_{i2} cos(t). \tag{3.23}$$

for some real-valued coefficients a_{i1} and a_{i2}. Any such function is 2π-periodic if and only if additionally $h_i(t) + 2\pi = h_i(t + 2\pi)$ for all $t \in [0, 2\pi]$.

The general framework of registering to low-dimensional linear subspace discussed in this paper can be readily adapted to incorporate the structural infor-

mation given by (3.23). For samples of periodic functions one may fit a $K = 2$ dimensional model with **pre-specified** basis functions $\gamma_1(t) = \sqrt{2}\sin(t)$ and $\gamma_2(t) = \sqrt{2}\cos(t)$. But for mixed samples with periodic and aperiodic function, additional basis functions accounting for the structure of aperiodic genes have to be determined nonparametrically. Some preliminary analysis showed that adding two additional function γ_3, γ_4 seems to be sufficient to capture all important effects. This translates into fitting a $K = 4$ dimensional model with two pre-specified basis functions.

The minimization is then similar to (3.14) and given by an alternation of (3.5) with

$$S_n(\mathbf{w}, 4) = \min_{(\gamma_i):\langle\gamma_i,\gamma_j\rangle=\delta_{ij}} \frac{1}{n}\sum_{i=1}^{n} ||x_i(h_{w_i}(t)) - \sum_{j=1}^{4}\gamma_j(t)\langle\gamma_j, x_i(h_{w_i}(t))\rangle||^2 \quad (3.24)$$

$$\text{s.t } \gamma_1(t) = \sqrt{2}\sin(t), \gamma_2(t) = \sqrt{2}\cos(t). \quad (3.25)$$

For periodic functions the warping functions have to reflect the periodicity of the curves as well, in particular $h_{w_i}(t)-t \approx h_{w_i}(t+2\pi)-(t+2\pi),\ t \in (0, 2\pi)$. Therefore also the minimum variance criteria is modified with

$$V_n(\mathbf{w}) = n^{-1}\sum_{i=1}^{n}\int_0^{2\pi}\left(h_{w_i}(t) - h_{w_i}(t+2\pi) + 2\pi\right)^2 dt. \quad (3.26)$$

Each exactly 2π-periodic function with the appropriate shape features should go along with $a_{i3} = a_{i4} = 0$ as well as with a warping function possessing the above property. Distance from 2π-periodicity may then be measured by the following "periodicity index":

$$per_i = \frac{\sum_{j=3}^{4} a_{ij}^2}{\sum_{j=1}^{2} a_{ij}^2} + \rho\int_0^{2\pi}\left(h_i(t) - h_i(t+2\pi) + 2\pi\right)^2 dt. \quad (3.27)$$

We shorten the notation for the optimal solution $\mathbf{w}^*$ with $a_{ij} := \langle \gamma_j^*, x_i(h_{w_i^*}(t)) \rangle$ and $h_i = h_{w_i^*}$. $\rho = median(\frac{\sum_{j=3}^{4} a_{ij}^2}{\sum_{j=1}^{2} a_{ij}^2})$ is used to scale between periodicity reflected by the warping and the amplitude space. If $per_i \approx 0$ then a curve is approximately periodic.

In a first step the described method is applied to the 612 curves selected as "periodic" by Spellman et al. (1998). Note than only for these genes the data also contain information about the corresponding group affiliation. Observations are slightly smoothed using a local polynomial smoother and a direct plug-in method to choose the bandwidth. Resulting model fits are displayed in Figure 3-5, while in Figure 3-6 we additionally display each curve in dependence of its affiliation to one of the phase groups. A number of individual genes lead to high values of per_i, indicating aperiodic structures. In the plots we thus introduce a grouping according to the individual values of the periodicity index. The upper quartile Q_3 of the empirical distribution of per_i is used as threshold. When considering the lower part of Figure 3-6, it is clearly seen that many function with $per_i > Q_3$ are non-periodic.

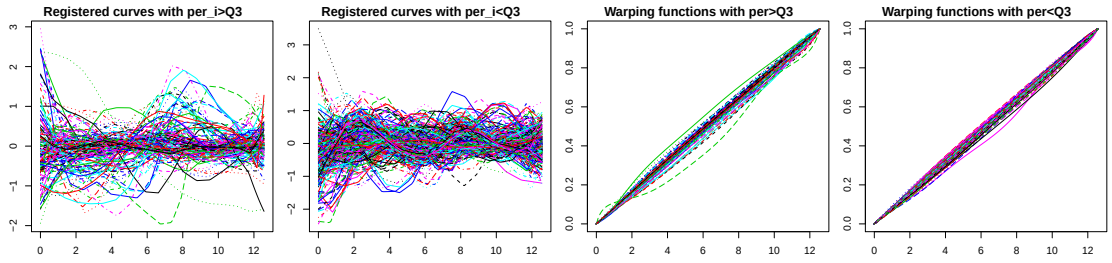


Figure 3-5: The two left pictures are the results of the registration to $span(\gamma_1, \ldots, \gamma_4)$ for the subgroup of 612 genes selected by Spellman et al. (1998). The left figure shows the curves with $per_i > Q_3$ and alongside with $per_i \leq Q_3$. The two right figures are the corresponding warping functions.

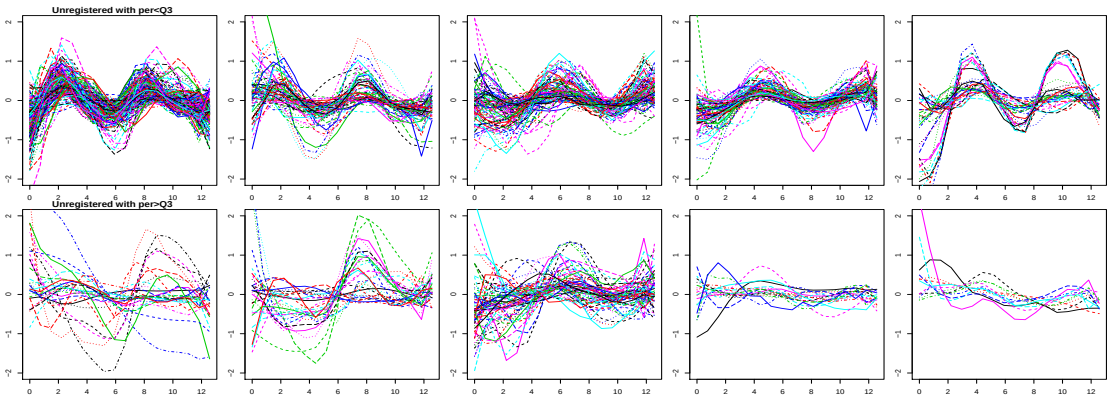


Figure 3-6: Registered functions from Figure 3-5 in dependence of phase groups; from left to right: $G2/M \rightarrow S/G2 \rightarrow S \rightarrow G1 \rightarrow M/G1$. The upper figures show the curves with $per_i \leq Q_3$, the lower with $per_i > Q_3$.

Following the ideas of Leng and Müller (2006b), we can now study the question whether group affiliation is connected to global shifts of an underlying functional structure. In the context of our analysis shifts may be due to amplitude variation as well as warping. Note that that

$$a_{i1}sin(t) + a_{i2}cos(t) = \beta_i sin(t + \phi_i), \tag{3.28}$$

$$\text{where } \phi_i := \begin{cases} arccos(\frac{a_{i2}}{\beta_i}) & \text{for } a_{i1} \geq 0 \\ 2\pi - arccos(\frac{a_{i2}}{\beta_i}) & \text{for } a_{i1} < 0 \end{cases}, \beta_i := \sqrt{a_{i1}^2 + a_{i2}^2}. \tag{3.29}$$

The coefficients a_{i1} and a_{i2} in (3.23) therefore provide information about a global shift ϕ_i of the basic sinusoidal structure. Although in the given context the role of warping functions mainly consists in quantifying functional relationships which cannot be exactly modelled by trigonometric functions, warping may obviously results in shifts of corresponding shape features. To approximate additional "global" shifts, linear approximations of the estimated warping functions with $d_i^* = argmin_d||h_i(t) - t - d||$ are used. This then leads to individual phase coefficients $\hat{s}_i := (\phi_i + d_i^*)mod(2\pi)$, which are determined for each of the 459 curves

Table 3.2: Five-number summary plus mean of the of the angles given by s_i grouped by the "Phase Group". Note that since we have to deal with angles we use the circular counterpart where one rotation is $360\,^\circ$ see for example Jammalamadaka et al. (2001).

	G1	M/G1	G2/M	S/G2	S
Min.	102.670	281.670	339.130	64.690	89.780
1st Qu.	257.770	312.030	52.480	141.860	184.260
Median	279.550	339.360	105.370	156.020	194.750
Mean	275.820	339.420	95.800	157.390	193.370
3rd Qu.	297.140	359.930	129.370	180.540	206.150
Max.	347.380	116.370	220.420	246.830	236.860

(out of the 612 selected by Spellman et al. (1998)) satisfying $per_i \leq Q_3$. In Table 3.2 it is clearly seen that the resulting phase coefficients (expressed in angles) yield very pronounced, clearly separated clusters for the different cell cycle phases $G2/M \to S/G2 \to S \to G1 \to M/G1$.

In a final step of our analysis the gained information is used to search the whole set of 4489 curves for periodic curves that may be overlooked before, and to assign them automatically to a phase group. All functions are registered to the pre-specified space given by $span(\gamma_1^*, \ldots, \gamma_4^*)$ obtained in first step, and per is then computed for each curve. To determine the group affiliation a prediction using multinomial logistic regression is carried out. The procedure is similar to (3.22) but allows for more than two groups. The logistic model is calibrated using the first two scores together with d_i^* and the group affiliation of the prior 459 selected curves. The results for the 612 curves with the smallest per_i and $a_{i1}^2 + a_{i2}^2 > median(a_1^2 + a_2^2)$ are displayed in Figure 3-7.

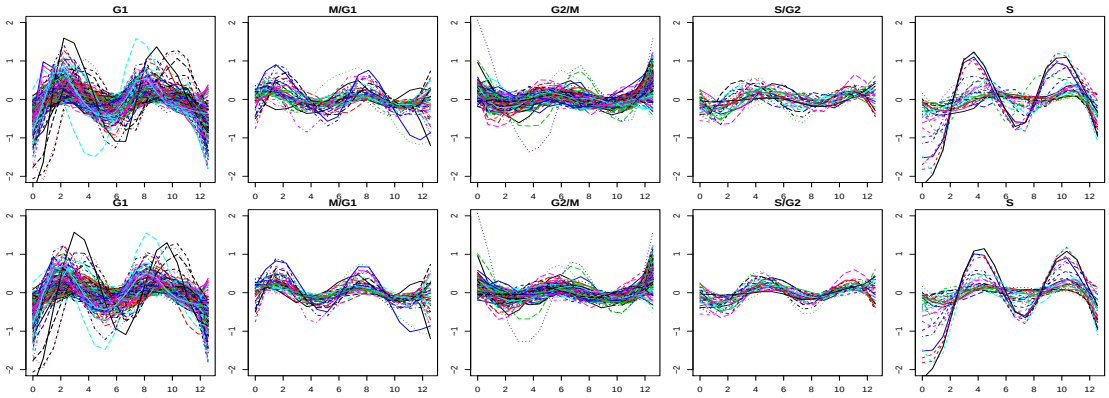


Figure 3-7: The Figure shows the 612 curves out of the complete sample of 4489 genes with the smallest *per* score. Group affiliations are obtained with an automatic clustering approach using a multinomial logistic regression. The upper figures show the unregistered functions, while the lower provide the corresponding registered curves.

In summary, we believe that the structural analysis presented in this section may constitute a promising path to achieve the goal of correctly identifying active genes. But any substantial progress will require additional biological input. In particular, any notion of "significance" will have to be based on some type of model for the random error contaminating micro-array data.

3.5.3 Aneurisk Data

Our method is used to analyze the AneuRisk65 data set Aneurisk-Team (2012). This data set contains the centerline of either the left or the right Internal Carotid Artery (ICA) of 65 Patients depending where an aneurism was suspected. One aim of the anuerisk project is to explore the role of vessel morphology to determine the pathogenesis of cerebral aneurysms. To detect the differences in the vessel morphology a registration is needed, because lengths and also shapes of blood vessels differ from person to person. Studies in this direction where published, for example, by Sangalli et al. (2009) using a $K = 1$ method, and by Sangalli et al. (2010) using the k mean approach with $k = 2$.

Our main goal is to link the position of the aneurysm to the vascular geometry of the ICA. For our analysis we dropped the patients where no aneurism was found, leaving 58 patients. The different length of the ICAs are normalized by scaling the data to $t = (0, 1)$. The data can be interpreted as a function $x_i : \mathbb{R} \to \mathbb{R}^3,\ i = 1 \ldots, 58$, where three dimensions $x(t) = [x_x(t), x_y(t), x_z(t)]$ are necessary to describe the spatial coordinates of an ICA. To model a low dimensional representation of the data we refer to chapter 8.5 of Ramsay and Silverman (2005) where a multivariate FPCA is described. It is straightforward to adopt the procedure to our needs. For three dimensional functions $\xi : \mathbb{R} \to \mathbb{R}^3$ we define an inner product by $\langle \xi_1, \xi_2 \rangle = \int_0^1 \sum_{m=(x,y,z)} \xi_{1,m}(u)\xi_{2,m}(u)du$.

Analogous to (3.1) a K dimensional representation of the registered curves are given by

$$x_i(h_i(t)) = y_i(t) \approx \sum_{j=1}^{K} [\gamma_{j,x}(t), \gamma_{j,y}(t), \gamma_{j,z}(t)]a_{ij}$$

where $a_{ij} = \langle y_i, \gamma_j \rangle$. For this application (3.5) was adjusted to work with three spatial coordinates using the corresponding squared norm $||\xi||^2 = \sum_{m=(x,y,z)} ||\xi_m||^2$ and

$$S(\mathbf{w}, K) = \min_{(\gamma_i):\langle \gamma_i, \gamma_j \rangle = \delta_{ij}} \sum_{i=1}^{N} ||x_i(h_{w_i}(t)) - \sum_{j=1}^{K} \gamma_j(t)\langle \gamma_j, x_i(h_{w_i}) \rangle||^2$$

In most approaches the data is additionally a priori adjusted for observing left or right ICA by flipping the sign of the x-coordinate. By introducing an additional dimension our method is capable to model such differences automatically, therefore there is no need to flip the sign manually. In addition one advantage of the above described method is that it does not only captures correlations in a single spacial direction, but also between the spacial coordinates. By using the data as raw as possible, there is a chance that other structural difference of left and right ICA

which does not only effect a simple coordinate flip in the x-direction but also other structural features possibly in other spacial directions becomes visible.

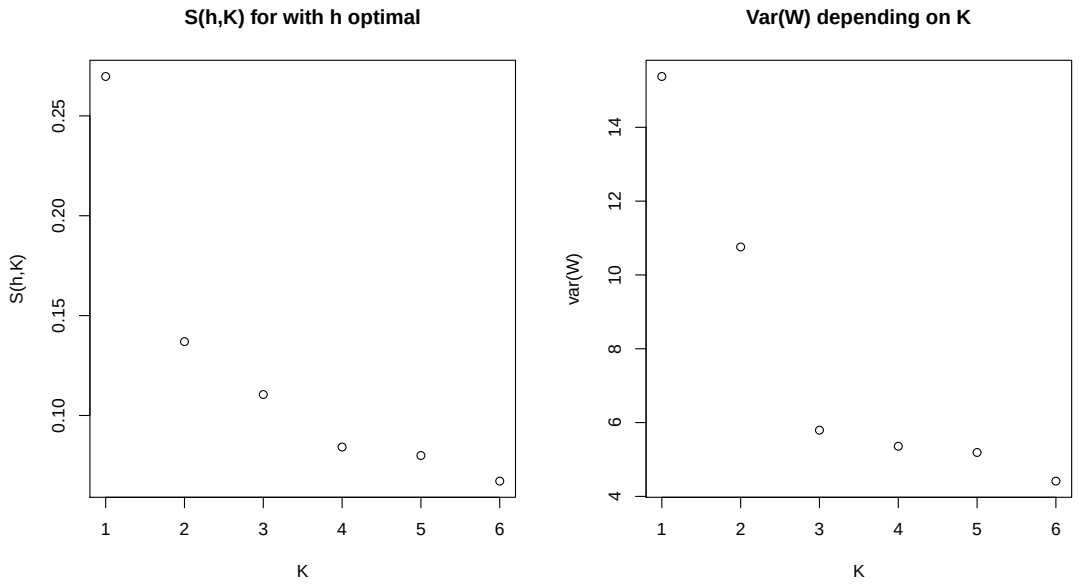


Figure 3-8: The left figure shows $S(\mathbf{w}, K)$ for different K, the right picture provides the corresponding values of $V(\mathbf{w})$ $(\equiv V(\mathbf{w}, K))$

The Aneurisk data is slightly smoothed, and there does not seem to exist a straightforward statistical model for the structure of the measurement error. Therefore the strategy presented in Remark 3.4.2 does not apply here, instead we use the graphical method proposed in Section 3.4.2 in order to determine a dimension K which compresses the information in as few components as possible. Looking at Figure 3-8 we argue that $K = 3$ is the best choice here. The gain in $S(\mathbf{w}, K)$ by using more than $K = 3$ dimensions is negligible, while the complexity of the warping does not decrease significantly. On the other hand using less than $K = 3$ dimensions comes at the cost of more complex warping functions. The outcome of the registration with $K = 3$ is shown in Figure 3-9.

H_0	$F_{U,1} = F_{L,1}$	$F_{U,2} = F_{L,2}$	$F_{U,3} = F_{L,3}$
p-value)	0.2778	0.03562	0.0111
H_0	$F_{r,1} = F_{l,1}$	$F_{l,2} = F_{r,2}$	$F_{l,3} = F_{r,3}$
p-value	$9.743e^{-08}$	$6.584e^{-13}$	0.6153

Table 3.3: $F_{U,j} := F(a_{ij}|G_{1,i} = U)$, $F_{L,j} := F(a_{ij}|G_{1,i} = L)$, $F_{r,j} := F(a_{ij}|G_{2,i} = r)$, $F_{l,j} := F(a_{ij}|G_{2,i} = l)$ denote the conditional probability distribution functions given different values of the binary variables G_1 and G_2.

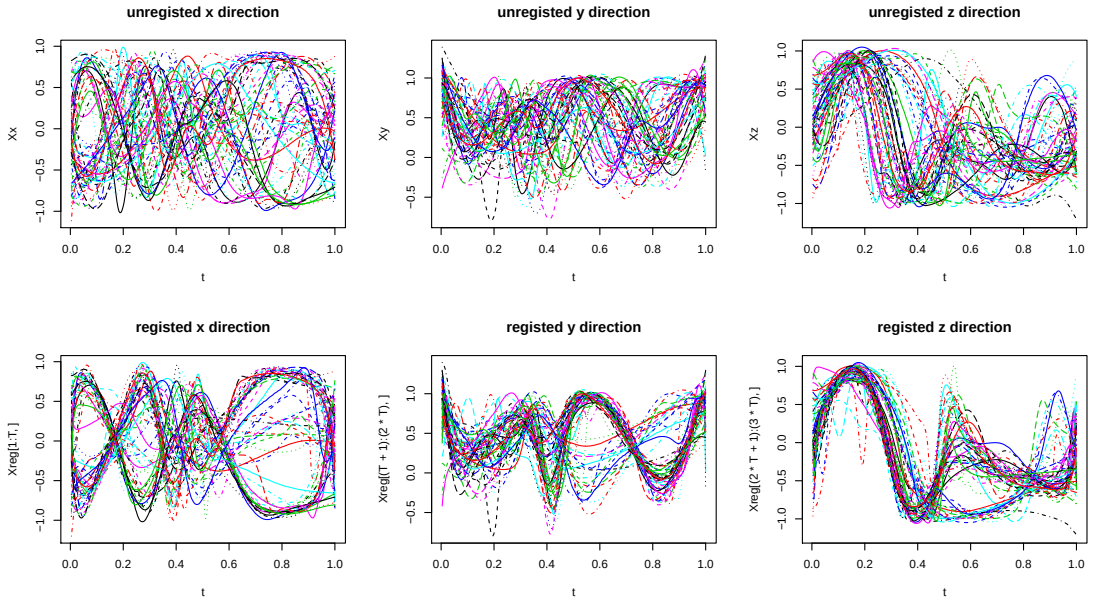


Figure 3-9: Upper pictures unregistered, Lower pictures registration to $K = 3$

For further analysis two groupings are introduced, $G_1 = \{U, L\}$ to describe if the aneurism is observed in the upper (U) or in the lower (L) ICA, and $G_2 = \{r, l\}$ which is used to code if the left (l) or right (r) ICA is observed. To check if components γ_j, $j = 1, \ldots 3$ are related to specific groups, we look for significant differences in the distributions of the respective scores a_{ij}, $j = 1, \ldots, k$. The results of a corresponding two-sided Kolmogorov-Smirnov tests are presented in Table 3.3.

It turns out that using a 5% significance level we can state that the 1st compo-

nent is connected to orientation of the ICA, while the 3rd component is related to the position of the aneurism. The 2nd component is linked to both characteristics.

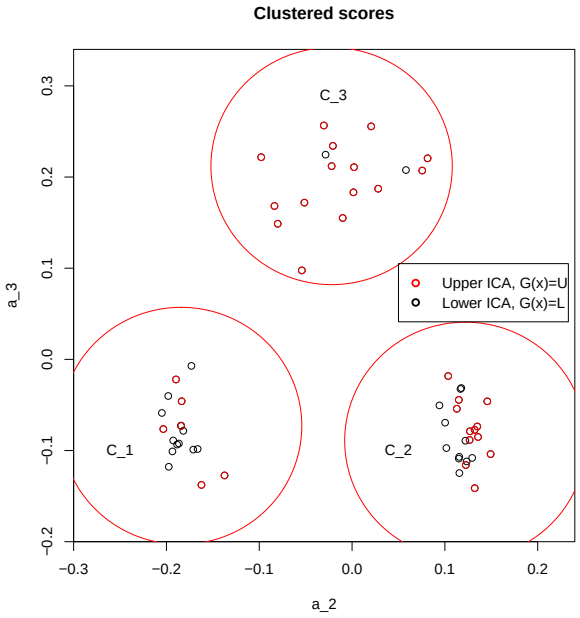


Figure 3-10: Scores a_{i2} are plotted against a_{i3} and clustered depending on the location. The color codes if upper (red) or lower (black) ICA was observed.

Cluster	$P_j(U)$	Risk Group
C_1	0.3333	Low
C_2	0.5217	Medium
C_3	0.8823	High

Figure 3-11: Risk classification depending on group affiliation $P_j(U) := P(G_{1i} = U | C(x_i) = C_j)$.

The "upper aneurysm" is the most dangerous one, it is of big interest for clinical reasons to identify the aneurysm position automatically. In Figure 3-10 the scores a_{i2} of the second component are plotted against the scores a_{i3} of the third one. We observe that these scores are clustered into 3 groups $C = \{C_1, C_2, C_3\}$. These clusters are related to the position of the aneurism and can be used to determine risk groups depending on the cluster affiliation as shown in Table 3-11.

3.6 Simulations

In 3.6.1 the construction of Figure 3-1 is discussed in detail and clustering capability is compared with Sangalli et al. (2010). 3.6.2 gives a detailed description of

Figure 3-2 and a Monte Carlo simulation to inspect small sample behavior under the presence of noisy observations is carried out. In 3.6.3 our method is compared to the FR-Metric approach by Srivastava et al. (2011) where 3.6.3 is a replicate of "Simulated Data 4" while 3.6.3 is very similar to "Simulated Data 3". Besides 3.6.3 illustrates the behavior of $V(\mathbf{w})$ for different choices of K.

3.6.1 Detailed description of Figure 3-1 and comparison with the "k-mean" approach

The introductory example of Figure 3-1 consists of a sample of $n = 48$ random functions $x_i(t)$, $i = 1, \ldots, n$, generated by the following process: For independent random variables $z_{1i} \sim \mathbf{N}(0, 0.2)$, $z_{2i} \sim \mathbf{N}(0, 0.1)$ and $z_{3i} \sim \mathbf{N}(0, 0.3)$. Let $t \in [0, 1]$, we define

$$y_i(t) = \begin{cases} (1.4 + z_{1i})sin(2\pi h_i(t)) + (1 + z_{2i})sin(2\pi h_i(t)^2) & \text{for } i = 1, \ldots, 24 \\ (2.1 + z_{1i})sin(2\pi h_i(t)) + (-1 + z_{2i})sin(2\pi h_i(t)^2) & \text{for } i = 25, \ldots, 48 \end{cases}$$

discretized at 128 equidistant points and $x_i = y_i \circ h_i^{-1}$. The warping functions are given by

$$h_i(t) = \frac{\exp(tz_{3i}) - 1}{\exp(z_{3i}) - 1}.$$

The data generating process implies that (3.1) holds with $K = 2$. The right upper part of Figure 3-12 shows the "true" data generating functions y_i. Using $\nu = 0.0001$ and 12 basis functions to approximate $w(u)$, with $K = 2$ the algorithm described in Section 3.4 recovers a two-dimensional linear function space with $S(\mathbf{w}, 2)/S_0 = 4.10 \cdot 10^{-6}$. This error is mainly due to linear interpolation of the curves as well as to some bias introduced by approximating the warping functions using a finite dimensional spline basis. Registered functions determined by the

algorithm are shown in the left, middle part of Figure 3-12. The two clusters introduced by the differences in the definition of y_i for $i \leq 24$ and $i > 24$, are clearly visible in the scores (a_{i1}, a_{i2}) of the fitted model, as shown in the lower left part of Figure 3-12. Indeed, for all functions belonging to the first group ($i \leq 24$) we have $a_{i2} < 0$, while $a_{i2} > 0$ for all functions belonging to the second group ($i > 24$).

Not shown here, using our algorithm with $K = 1$ leads to results very similar to the Fisher-Rao metric registration shown in Figure 3-1. Unlike a registration with $K = 2$ the corresponding registered curves can no longer be described by two components and $S(\mathbf{w}, 1)/S_0 = 4.48 \cdot 10^{-4}$.

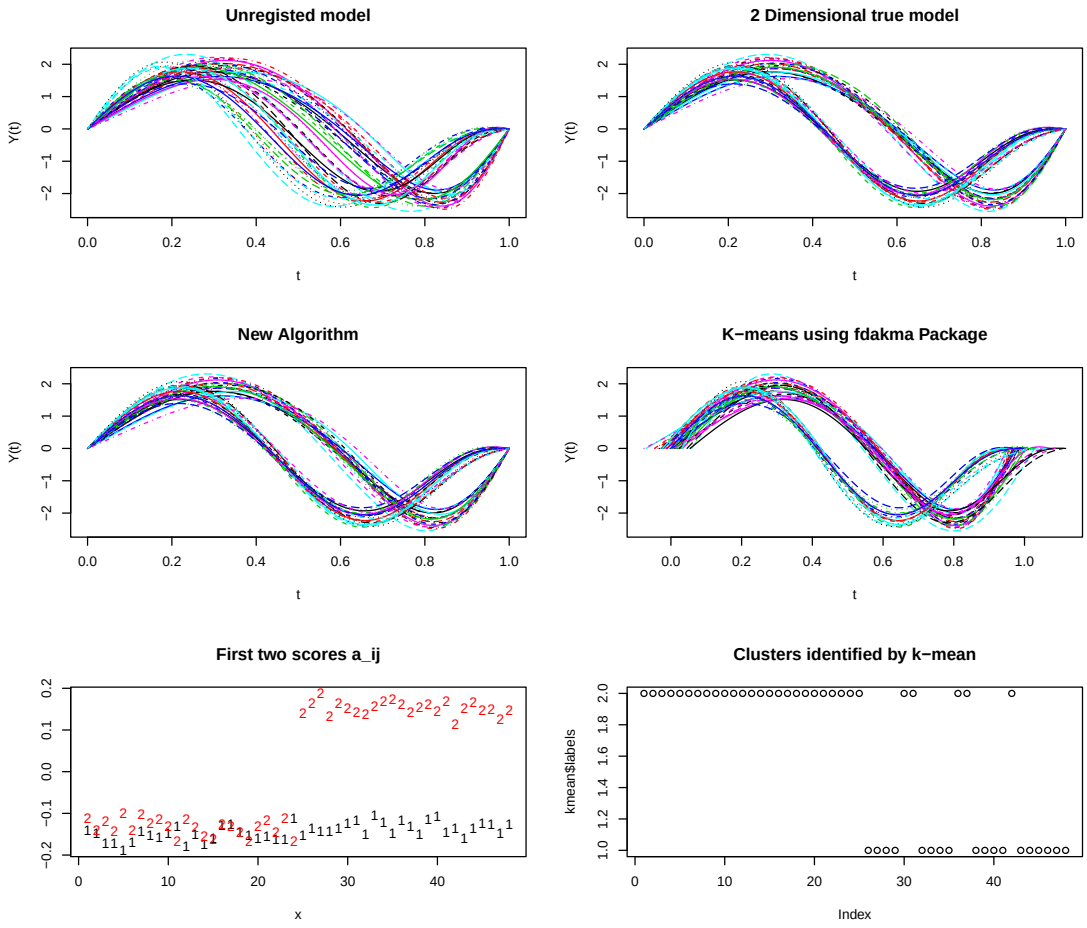


Figure 3-12: The left middle part show the registered curves determined by our algorithm; the lower left part shows the scores a_{i1} and a_{i2} of the fitted model in dependence of indices $i = 1, \ldots, 48$ of the 48 functions. The right middle part show the registered curves calculated by by the k-means algorithm using the R-package fdakma; the lower part part show the cluster affiliations determined by k-means in dependence of indices $i = 1, \ldots, 48$ of the 48 functions.

As already discussed in Section 3.2.2, our method may be reasonably for clustering purposes and to be compared to the k-means approach of Sangalli et al. (2010) (with $k = 2$). A Monte Carlo Simulation with 100 repetitions results in an average correct classification rate of 99.60% compared to 93.15% using the "k-mean" approach with the "d0.pearson" option as similarity measure. We want to note, however, that only an approximate solution of (3.1) can be obtained by the

"k-mean" approach. One reason is that the data generating process, implies that even within each cluster two components are necessary to capture amplitude variation. Another important point is that the "k-mean" algorithm as implemented in the R-package "fdakma" by Parodi et al. (2015) only relies on a very simple shift-scale model for warping functions. Consequently, resulting registered curves are no longer defined over identical domains. This effect is clearly seen in the right, middle part of Figure 3-12 for a given sample. The lower left part of Figure 3-12 shows that k-means is able to cluster the curves quite accurately, only six misclassifications are to be observed.

3.6.2 Detailed description of Figure 3-2 and Monte Carlo simulation

To derive the curves in Figure 3-2 a two dimensional factor model using the second and fourth Legendre-polynomial is simulated, i.e. we generated $i = 1, \ldots, 15$ curves over the interval $t \in [-1, 1]$ discretized at 101 equidistant points by:

$$x_i(t) = a_{i1}\frac{1}{2}(3h_i(t)^2 - 1) + a_{i2}\frac{1}{8}(35h_i(t)^3 - 30h_i(t)^2 + 3),$$

the warping functions h_i are given by

$$h_i(t) = 2\frac{\exp(z_i(t+1)/2) - 1}{\exp(z_i) - 1} - 1, \quad z_i \neq 0 \quad \text{and } t \text{ otherwise}$$

with a_{i1}, a_{i2}, z_i are iid. $\mathbf{N}(0, 1)$.

Here, the generated true curves do not share the same peak locations or even the same amount of peaks. Hence, any registration procedure which does not register the given sample to a two dimensional space, but instead tries to register to $K = 1$, will necessary fail to give an exact low dimensional representation.

Using our algorithm with $K = 2$ we obtain the registration shown in the lower right picture of Figure 3-2 with $S(\mathbf{w}, K)/S_0 = 4.06 \cdot 10^{-5}$. To get the extreme warping shown in the the upper right picture using $K = 1$ it is necessary to let the "newoua" algorithm iterate several thousand times since converging to an extreme warping is harder to achieve for the algorithm.

Further a simulation with 1000 repetitions using discretizations with $T = 101$ and $T = 201$ points is carried out. The performance of the registration with an additional error is evaluated due to $\tilde{x}_i(t_j) = x_i(t_j) + \epsilon_{ij},\ \epsilon_{ij} \sim \mathbf{N}(0, 0.1),\ i = 1, \ldots 15,\ j = 1, \ldots T$. As described in Section 3.3.2 the noisy curves are pre-smoothed using a local polynomial smoother using the "locpoly" function from the R package "KernSmooth". To determine the bandwidth we use a direct plug-in method implemented with "dpill". The smoothed curves are labeled as $\hat{x}_i$. To evaluate the performance we use $\tilde{S}(K, f) := n^{-1} \sum_{i=1}^{n} \int_0^1 (f_i(\hat{h}_{K,i}(t)) - \sum_{j=1}^{K} \hat{a}_{K;ij} \hat{\gamma}_{K,j}(t))^2 dt$ where $\hat{h}_{K,i}, \hat{\gamma}_{K,j}$ minimize (3.14) in dependence of f and K. Using our criteria we suggest a critical value $A\frac{\hat{\sigma}^2 V(\mathcal{K})}{Tb} \approx 0.0045$ for $T = 101$ and $A\frac{\hat{\sigma}^2 V(\mathcal{K})}{Tb} \approx 0.0023$ for $T = 201$ with $A = 1.1$. This leads to a choice of $\mathbf{K}_0 = 2$ which corresponds to the true value.

Table 3.4: It can be verified that with increasing T, $\tilde{S}(K, f)$ decreases for reliable choices for K given by $K = 2$ or $K = 3$ while $\tilde{S}(K = 1, f) \approx 0.1$ independent of T or the presence of noise. While for $K = 2$ or $K = 3$ the variance and inter quartile range is very small which means that the algorithm almost always works very well, for $K = 1$ the results are more fluctuating.

	T=101						T=201					
	$f = x$			$f = \hat{x}$			$f = x$			$f = \hat{x}$		
	$K = 3$	$K = 2$	$K = 1$	$K = 3$	$K = 2$	$K = 1$	$K = 3$	$K = 2$	$K = 1$	$K = 3$	$K = 2$	$K = 1$
$10 \cdot mean$	0.0001	0.0001	1.116	0.019	0.036	1.075	0.0001	0.0001	1.103	0.014	0.023	1.063
$10^3 \cdot variance$	0.00000	0.00000	2.420	0.0003	0.001	2.114	0.00000	0.00002	2.150	0.0001	0.0002	2.008
$10^2 \cdot iqr$	0.001	0.001	6.362	0.060	0.106	6.147	0.001	0.001	6.285	0.042	0.060	6.057

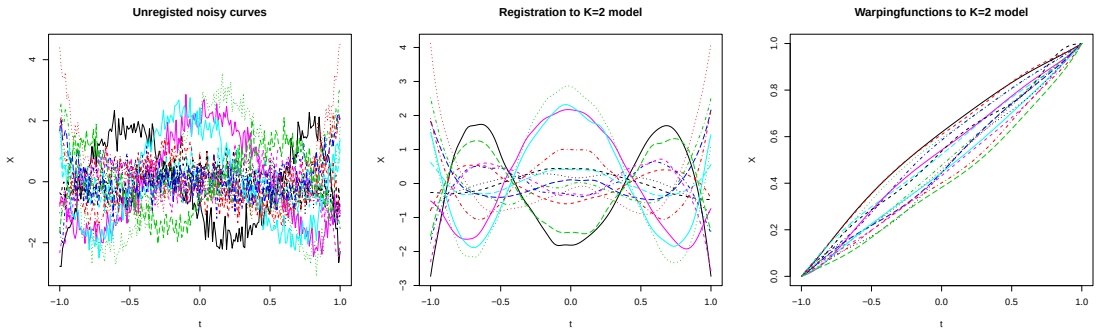


Figure 3-13: The figure displays the same curves as Figure 3-2 but with additional noisy as described in Section 3.6.2. The middle picture shows the pre-smoothed curves and registered curves using a local polynomial smoother and our algorithm. The right picture shows the corresponding warping functions.

3.6.3 Comparison with the FR-Metric approach for $\mathbf{K}_0 = 1$

For $\mathbf{K}_0 = 1$ Srivastava et al. (2011) did an elaborated simulation study where the FR-Metric approach is compared to various other registration methods using $ls = n^{-1} \sum_{i=1}^{n} \frac{\int (y_i(t) - (n-1)^{-1} \sum_{j=1}^{n-1} y_j(t))^2 dt}{\int (x_i(t) - (n-1)^{-1} \sum_{j=1}^{n-1} x_j(t))^2 dt}$, $sls = n^{-1} \sum_{i=1}^{n} \frac{\int (y_i'(t) - (n-1)^{-1} \sum_{j=1}^{n-1} y_j'(t))^2 dt}{\int (x_i'(t) - (n-1)^{-1} \sum_{j=1}^{n-1} x_j'(t))^2 dt}$ and $pc = \frac{\sum_{i \neq j} cc(y_i(t), y_j(t))}{\sum_{i \neq j} cc(x_i(t), x_j(t))}$ where $cc(f, g)$ is the pairwise Pearson's correlation between functions. Here the FR-Metric approach basically outperforms each of the other approaches. In this Section replicates of "Simulated Data 4" and "Simulated Data 3" by Srivastava et al. (2011) are constructed. We use ls, sls and pc to compare our registration outcome using $K = 1$ to the FR-Metric approach, results are given in Table 3.5. We can conclude that our method had a comparable performance which is a good result since the FR-Metric approach is possibly the best $K = 1$ registration method.

Table 3.5: Comparison between our approach and the FR-Metric approach using ls, sls and pc. The left columns relate to Section 3.6.3 while the right columns belong to 3.6.3.

	Simulation 3.6.3		Simulation 3.6.3	
	K=1 Reg.	FR-Metric	K=1 Reg.	FR-Metric
ls	0.019	0.019	0.004	0.004
sls	0.018	0.020	0.003	0.002
pc	14.517	14.514	21.319	21.319

Replicate of "Simulated Data 4"

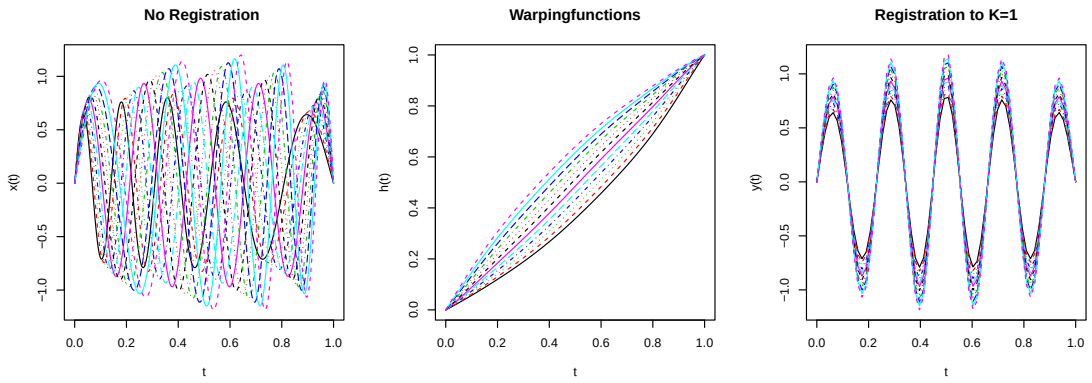


Figure 3-14: Alignment using our algorithm, this $\mathbf{K}_0 = 1$ simulation is very demanding for most algorithms because there is a tendency to stuck in a local minima and match the wrong peaks.

The curves shown by Figure 3-14 correspond to "Simulated Data 4" from Srivastava et al. (2011). These curves are constructed by simulating 12 curves with $a = (0.8, 0.8333, \ldots, 1.2)$ over the interval $t \in [0, 9]$ discretized at 512 equidistant points given by

$$x_i(t) = a_i(1 - (h_i(t)/9 - 0.5)^2)sin(\pi h_i(t)),\ i = 1, \ldots, 12.$$

The warping functions are given with $z = (-1.2, -1, \ldots, 1, 1.2)$ by

$$h_i(t) = 9\frac{e^{z_i t/9} - 1}{e^{z_i} - 1}, \quad z_i \neq 0 \quad \text{and } t \text{ otherwise.}$$

Even though this simulation has only one component it is very demanding for most algorithm. The challenge is that the initial peak locations overlap and algorithms that use a local registration approach will likely stuck in a local minima. Using our approach we avoid this issue by starting a registration using $K_{max} = 4$ and gradually count down to $K = 1$ as described in Section 3.4.2. The resulting registration outcome is shown in Figure 3-14 with $S(\mathbf{w}, K)/S_0 = 1.01e^{-4}$.

Replicate of "Simulated Data 3"

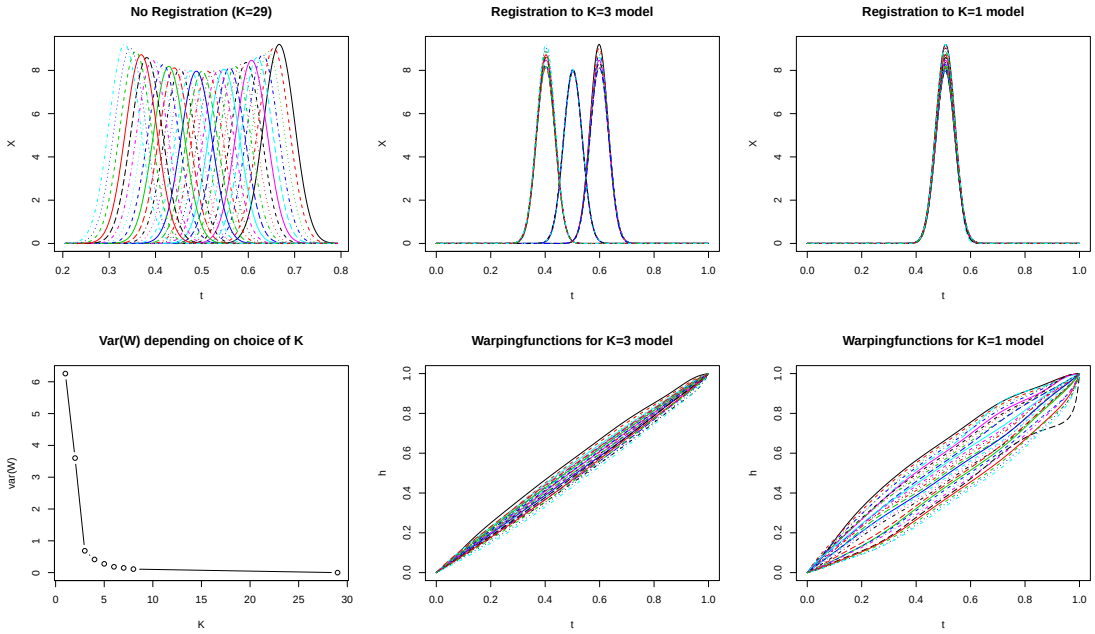


Figure 3-15: A registration using different choices of K is carried out. At the lower left figure the connection between K and the complexity of the warping is documented.

The curves displayed in the upper left of Figure 3-15 are similar to "Simulated Data 3" from Srivastava et al. (2011), constructed by simulating 29 differently scaled and shifted beta distributions. To construct the curves we use an equally spaced $a = (0.25, \ldots, 0.75)$ together with an interval $t \in [0, 1]$ discretized at 512

equidistant points, $s(m) = 0.75 - 0.5m,\ z(m) = 5\frac{(s(m)-0.5)^2}{2} + 1$, then

$$x_i(t) = z(a_i) f\left(s(a_i) + (1.5t - 0.75), 50, 50\right),\ i = 1, \ldots, 29$$

where $f(x, \alpha, \beta) = B(\alpha, \beta)^{-1} x^{\alpha-1}(1-x)^{\beta-1}$ and $B(\alpha, \beta)$ is the betafunction.

Figure 3-15 examine different registrations where K decreases from $K = 8$ to $K = 1$. It can be verified that higher K results in less complex warping functions while with $K = 1$ or $K = 2$ a high $V(\mathbf{w})$ has to be accepted.

3.7 Proofs

3.7.1 Proof of Proposition 1

Proposition 1a) is an immediate consequence of Proposition 1 of Kneip and Ramsay (2008). For proving Proposition 1 b) first note that by assumption a realization x_i of the random process a.s. satisfies $x_i'(\tau_{i,l}) = 0$, $x_i''(\tau_{i,l}) \neq 0$ for all $l = 1, \ldots, q(x_i)$ and there a.s. exists a $y_i \in \mathcal{L}_K$ with $Q(x_i) = Q(y_i)$. Since by definition y_i is twice continuously differentiable and $Q(y_i)$ contains all values of y_i where y_i' is zero, there necessarily exist exactly $q(x_i)$ points $\tau_{y_i,l}$, $l = 1, \ldots, q(x_i)$ with $y_i'(\tau_{y_i,l}) = 0$, and necessarily also $y_i''(\tau_{y_i,l}) \neq 0$. With $\tau_{i0} = \tau_{y_i,0} := 0$ and $\tau_{i,q(x_i)+1} = \tau_{y_i,q(x_i)+1} := 1$ the function x_i and y_i are therefore strictly monotone in each of the segments $[\tau_{i,l}, \tau_{i,l+1}]$ and $[\tau_{y_i,l}, \tau_{y_i,l+1}]$, $l = 0, \ldots, q(x_i)$, respectively. Hence, for each $l = 0, \ldots, q(x_i)$ there exists a strictly monotonic function $z_{x_i;l}^{-1} : [x_i(\tau_{i,l}), x_i(\tau_{i,l+1})] \to [\tau_{i,l}, \tau_{i,l+1}]$ such that $z_{x_i;l}^{-1}[z_{x_i;l}(t)] = t$ for all $t \in [\tau_{i,l}, \tau_{i,l+1}]$. Defining then $h_{il} : [\tau_{y_i,l}, \tau_{y_i,l+1}] \to [\tau_{i,l}, \tau_{i,l+1}]$ by $h_{il}(t) = z_{x_i;j}^{-1}[y_i(t)]$, the above construction implies that with $h_i(t) = \sum_{l=0}^{q(x_i)} h_{il}(t) I(t \in [\tau_{y_i,l}, \tau_{y_i,l+1}])$ we obtain $x_i(h_i(t)) = y_i(t)$ for all $t \in [0, 1]$. h_i is a strictly monotonic function, and twice continuous differentiability of x_i and y_i translates into $h_i \in \mathcal{W}^2[0, 1]$. Note that $h_i'(t) = \frac{y_i'(t)}{x_i'[h_i(t)]}$ for $t \notin$

$\{\tau_{y_i,1},\ldots,\tau_{y_i,q(x_i)}\}$, and $h_i(\tau_{y_i,l}) = \lim_{t\to\tau_{y_i,l}} \frac{y_i'(t)}{x_i'[h_i(t)]} = (\frac{y_i''(\tau_{y_i,l})}{x_i''(\tau_{i,l})})^{1/2}$. This proves assertion 1b).

It remains to show assertion 1c). If $\mathbb{P}(\boldsymbol{q} = q(x_i)) = 1$ as well as $\mathbb{P}(Q(x_i) = aQ(x_j)$ for some $a \in \mathbb{R}) = 1$, then for an arbitrary realization x_j and $\mathcal{L}_1 := \{v|\ v = ax_j$ for some $a \in \mathbb{R}\}$ we have $\mathbb{P}(Q(x_i) = Q(\gamma)|\gamma \in \mathcal{L}_1) = 1$. It then follows from Proposition 1b) that (3.3) holds for $K = 1$. On the other hand, if $x_i(h_i(t)) = a_i\gamma(t)$ a.s. for all $t \in [0,1]$, then $\mathbb{P}(\boldsymbol{q} = q(x_i)) = 1$ for $q = q(\gamma)$. Furthermore, $\mathbb{P}(Q(x_i) = aQ(\gamma)$ for some $a \in \mathbb{R}) = 1$, and hence also $\mathbb{P}(Q(x_i) = aQ(x_j)$ for some $a \in \mathbb{R}) = 1$.

3.7.2 Proof of Proposition 2

Select an arbitrary integer $\mathbf{q}$. Since the eigenfunctions $\gamma_1, \gamma_2, \ldots$ of Γ are a sequence of orthornormal functions, there exists a an integer J such that the function γ_J has $\mathbf{q}^* > \mathbf{q} + 1$ zero crossings in the interior of $[0,1]$. Let $\mu_{b,J} := \mu + b\gamma_J$. If b is sufficiently large, then also $\mu_{b,J}$ has at least $\mathbf{q}^*$ zero crossings. Choose some $a > 0$. There then exists a $b_a < \infty$ and some $t_1,\ldots,t_{\mathbf{q}^*+1}$ such that $|\mu_{b_a,J}(t_j)| \le a$, $j = 1,\ldots,\mathbf{q}^*+1$, as well as $sign(\mu_{b_a,J}(t_j)) = -sign(\mu_{b_a,J}(t_{j+1}))$, $j = 1,\ldots,\mathbf{q}^*$.

We have $x_i = \mu_{b_{ij},J} + \tilde{x}_i$, where $\tilde{x}_i := \sum_{j,j\neq J} b_{ij}\gamma_j$. But obviously x_i has at least $\mathbf{q}^*-1$ shape features in $(0,1)$ if $sign(x_i(t_j))) = -sign(x_i(t_{j+1}))$ for all $j = 1,\ldots,\mathbf{q}^*$. This is necessarily true if $b_{ij} > b_a$ and if at the same time $(\tilde{x}_i(t_1),\ldots,\tilde{x}_i(t_{\mathbf{q}^*+1}))^T \in (-a,a)^{\mathbf{q}^*+1}$. By assumption, the vector $(\tilde{x}_i(t_1),\ldots,\tilde{x}_i(t_{\mathbf{q}^*+1}))^T$ follows a multivariate normal distribution and is independent of $b_{ij} \sim N(0,\lambda_j))$. Hence, the proposition is an immediate consequence of

$$\mathbb{P}(q(x_i) \ge \mathbf{q}) \ge \mathbb{P}\left(b_{ij} \ge b_a\right) \cdot \mathbb{P}\left((\tilde{x}_i(t_1),\ldots,\tilde{x}_i(t_{\mathbf{q}^*+1})) \in (-a,a)^{\mathbf{q}^*+1}\right) > 0$$

3.7.3 Proof of Theorem 1

For $x, y \in \mathcal{L}^2[0,1]$ let $\langle x, y\rangle := \int_0^1 x(t)y(t)dt$, and $\|x\|_2^2 := \int_0^1 x(t)^2 dt$.

Select some integer K. Since $\mathcal{W}_d^1[0,1]$ is a compact space, for all orthonormal $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K) \in (\mathcal{W}^1[0,1])^K$

$$s(x_i, w, \boldsymbol{\gamma}) := \|x_i \circ h_w - \sum_{j=1}^{K} \langle x_i \circ h_w, \gamma_j\rangle \gamma_j\|_2^2$$

attains a minimum $w_{i,\boldsymbol{\gamma}}$ over all $w \in \mathcal{W}_d^1[0,1]$, and $\mathcal{R}_{\boldsymbol{\gamma}} : x_i \to w_{i,\boldsymbol{\gamma}}$ is a measurable operator. For $c_0 > 0$ let $\mathcal{W}_{c_0;1}^2[0,1] := \{y \in \mathcal{W}^2[0,1] |\ \|y\|_2 = 1,\ sup_{u\in[0,1]}|y''(u)| \leq c_0\}$, and note that by assumption $K < \mathbf{K}_0$ implies $\mathbb{E}(s(x_i, w_{i,\boldsymbol{\gamma}}, \gamma)) > 0$ **for all** $\boldsymbol{\gamma} \in (\mathcal{W}_{c_0;1}^2[0,1])^K$. By the Arzela–Ascoli theorem $(\mathcal{W}_{c_0;1}^2[0,1])^K$ is a compact function space (with respect to supremum as well as L_2-metrics), and therefore a minimum $r(c_0, K) > 0$ of $\mathbb{E}(s(x_i, w_{i,\gamma}, \boldsymbol{\gamma})) > 0$ is attained for some element $\boldsymbol{\gamma} \in (\mathcal{W}_{c_0;1}^2[0,1])^K$. We can thus conclude that if c_0 is sufficiently large,

$$r(c_0, K) := \min_{\gamma \in (\mathcal{W}_{c_0;1}^2[0,1])^K} \mathbb{E}(s(x_i, w_{i,\boldsymbol{\gamma}}, \boldsymbol{\gamma})) > 0 \qquad \text{for } K < \mathbf{K}_0, \tag{3.30}$$

$$\min_{\gamma \in (\mathcal{W}_{c_0;1}^2[0,1])^K} \mathbb{E}(s(x_i, w_{i,\boldsymbol{\gamma}}, \boldsymbol{\gamma})) = 0 \qquad \text{for } K \geq \mathbf{K}_0, \tag{3.31}$$

Now consider minimizing $S_n(\mathbf{w}, K)$ with respect to the true functions x_i, $i = 1, \ldots, n$. For given $\mathbf{w}$ corresponding functions $\gamma_j \equiv \gamma_{j,\mathbf{w}}$ can then be determined as eigenfunctions of the empirical second moment operator, i.e. for an eigenvalue l_j the functions γ_j satisfies $l_j\gamma_j(t) = \frac{1}{n}\sum_{i=1}^n \langle x_i \circ h_{w_i}, \gamma_j\rangle x_i(h_{w_i}(t))$. This implies that γ_j is twice differentiable, and it follows from our assumptions on the derivatives of x_i and w_i that there exist some $c_0 < \infty$ such that with probability tending to 1

$\sup_u |\gamma_j''(u)| \leq c_0$ as $n \to \infty$. Therefore, with probability tending to 1

$$\min_{\mathbf{w} \in (\mathcal{W}_d^1[0,1])^n} S_n(\mathbf{w}, K) = \min_{\boldsymbol{\gamma} \in (\mathcal{W}_{c_0;1}^2[0,1])^K} \frac{1}{n} \sum_{i=1}^n s(x_i, w_{i,\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \tag{3.32}$$

For $y \in (\mathcal{W}_{c_0;1}^2[0,1])^K$ and $\delta > 0$ let $U_\delta(y) := \{z \in (\mathcal{W}_{c_0;1}^2[0,1])^K | \ \|y_j - z_j\| \leq \delta \ \text{ for all } j = 1, \ldots, K\}$. Since $(\mathcal{W}_{c_0;1}^2[0,1])^K$ is compact, for any $\delta > 0$ there exists some $m(\delta) < \infty$ and functions $\boldsymbol{\gamma}^l \in (\mathcal{W}_{c_0;1}^2[0,1])^K$ such that $(\mathcal{W}_{c_0;1}^2[0,1])^K \subset \bigcup_{l=1}^{m(\delta)} U_\delta(\boldsymbol{\gamma}^l)$. The triangle inequality implies that for any $l = 1, \ldots, m(\delta)$, all $y \in U_\delta(\boldsymbol{\gamma}^l)$, and each $w \in \mathcal{W}_d^1[0,1]$ we have

$$\begin{gathered}
(\tfrac{1}{n} \textstyle\sum_{i=1}^n s(x_i, w, y))^{1/2} \geq \\
(\tfrac{1}{n} \textstyle\sum_{i=1}^n \|x_i \circ h_w - \sum_{j=1}^K \langle x_i \circ h_w, y_j \rangle \gamma_j^l\|_2^2)^{1/2} \\
-(\tfrac{1}{n} \textstyle\sum_{i=1}^n \sum_{j=1}^K \langle x_i \circ h_w, y_j \rangle^2 \|\gamma_j^l - y_j\|_2^2)^{1/2}.
\end{gathered}$$

There obviously exists a constant $D_0 < \infty$ such that $\sup_{t \in [0,1]} |(h_w^{-1})'(t)| \leq D_0$ for all $w \in \mathcal{W}_d^1[0,1]$. Therefore $\|x_i \circ h_w\|_2^2 = \int_0^1 (h_w^{-1})'(t) x_i(t)^2 dt \leq D_0 \|x_i\|_2^2$. For any w we thus obtain $\langle x_i \circ h_w, y_j \rangle^2 \leq D_0 \|x_i\|_2^2$, and therefore

$$\min_{\boldsymbol{\gamma} \in (\mathcal{W}_{c_0;1}^2[0,1])^K} (\frac{1}{n} \sum_{i=1}^n s(x_i, w_{i,\boldsymbol{\gamma}}, \boldsymbol{\gamma}))^{1/2} \geq \tag{3.33}$$

$$\min_{l=1,\ldots,m(\delta)} (\frac{1}{n} \sum_{i=1}^n s(x_i, w_{i,\boldsymbol{\gamma}^l}, \boldsymbol{\gamma}^l))^{1/2} - (K \frac{1}{n} \sum_{i=1}^n D_0 \|x_i\|_2^2)^{1/2} \delta \tag{3.34}$$

At the same time, $\mathbb{E}(\|x_i\|_2^2) < \infty$, as well as $\mathbb{E}(s(x_i, w_{i,\boldsymbol{\gamma}^l}, \boldsymbol{\gamma}^l)) < \infty$ for all $l = 1, \ldots . m(\delta)$. Furthermore, for any $l = 1, \ldots, m(\delta)$, $s(x_1, w_{1,\boldsymbol{\gamma}^l}, \boldsymbol{\gamma}^l), \ldots, s(x_n, w_{n,\boldsymbol{\gamma}^l}, \boldsymbol{\gamma}^l)$ are i.i.d. random variables, and the law of large numbers such implies that for each

$0 < \delta^* < \infty$

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} s(x_i, w_{i,\boldsymbol{\gamma}^{\boldsymbol{l}}}, \boldsymbol{\gamma}^{\boldsymbol{l}}) - \mathbb{E}(s(x_i, w_{i,\boldsymbol{\gamma}^{\boldsymbol{l}}}, \boldsymbol{\gamma}^{\boldsymbol{l}}))| \leq \delta^* \text{ for all } l = 1, \ldots, m(\delta) \right) \to 1$$

as $n \to \infty$. Since $\frac{1}{n}\sum_{i=1}^{n} \|x_i\|_2^2 \to \mathbb{E}(\|x_i\|_2^2)$ a.s. and since δ, δ^* are arbitrary, we can conclude from (3.30), (3.32), (3.33), and (3.35) that

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{\mathbf{w}\in(\mathcal{W}_{d[0,1])^n}^1} S_n(\mathbf{w}, K) \geq \tilde{c}_K \right) = 1 \quad \text{for all } K < \mathbf{K}_0 \tag{3.35}$$

$$\text{and each } 0 < \tilde{c}_K < r(c_0, K) \tag{3.36}$$

Now consider the local linear estimator $\hat{x}_i$ of x_i. The estimator can be written in the form $\hat{x}_i(t) = \sum_{l=1}^{T} v(t, t_l, b) Y_{il}$, where for any t, l, b the the weights $v(t, t_l, b)$ can be computed from the kernel function. We then have $\hat{x}_i(t) = r_1(x_i, b; t) + r_{2i}(b; t)$, where $r_1(x_i, b; t) = \sum_{l=1}^{T} v(t, t_l, b) x_i(t_l)$ and $r_{2i}(b; t) = \sum_{l=1}^{T} v(t, t_l, b)\epsilon_{il}$. Under our assumptions $r_2(b; t)$ is independent of x_i and . Since error terms are homoscedastic and design points are equidistant, standard arguments (see e.g. Fan and Gijbels (1996)) can now be used to show that there exist constants $D_1, D_2 < \infty$ such that for all sufficiently large T

$$\sup_{t\in[0,1]} \mathbb{E}((X_i(t) - r_1(x_i, b; t))^2) \leq b^4 D_1 \sup_{t\in[0,1]} E|x_i''(t)^2|, \tag{3.37}$$

$$\sup_{t\in[0,1]} \mathbb{E}(r_{2i}(b; t)^2) \leq \frac{D_2}{Tb}, \tag{3.38}$$

$$\sup_{t\in[0,1]} \mathbb{E}((\hat{x}_i(t) - x_i(t))^2) = \sup_{t\in[0,1]} \left(\mathbb{E}((X_i(t) - r_1(x_i, b; t))^2) + \mathbb{E}(r_{2i}(b; t)^2) \right) \tag{3.39}$$

$$\leq b^4 D_1 \sup_{t\in[0,1]} E|x_i''(t)^2| + \frac{D_2}{Tb}, \tag{3.40}$$

while as $Tb\log T \to \infty$

$$\sup_{t\in[b,1-b]} |\mathbb{E}(r_{2i}(b;t))^2 - \frac{\sigma^2 V(\mathcal{K})}{Tb}| = o(1) \tag{3.41}$$

By the triangle inequality it follows that for any $\mathbf{w} \in (\mathcal{W}_d^1[0,1])^n$

$$S_n(\mathbf{w},K)^{\frac{1}{2}} - (\frac{1}{n}\sum_{i=1}^n \int_0^1 (\hat{x}_i(h_{w_i}(t)) - x_i(h_{w_i}(t)))^2 dt)^{\frac{1}{2}} \tag{3.42}$$

$$\leq \hat{S}_n(\mathbf{w},K)^{\frac{1}{2}} \leq S_n(\mathbf{w},K)^{1/2} + (\frac{1}{n}\sum_{i=1}^n \int_0^1 (\hat{x}_i(h_{w_i}(t)) - x_i(h_{w_i}(t)))^2 dt)^{\frac{1}{2}}. \tag{3.43}$$

Since $\int_0^1 (\hat{x}_i(h_w(t)) - x_i(h_w(t)))^2 dt = \int_0^1 (h_w^{-1})'(t)(\hat{x}_i(t) - x_i(t))^2 dt \leq D_0 \int_0^1 (\hat{x}_i(t) - x_i(t))^2 dt$, we can thus infer from (3.40) that

$$\sup_{\mathbf{w}\in(\mathcal{W}_d^1[0,1])^n} |\hat{S}_n(\mathbf{w},K)^{\frac{1}{2}} - S_n(\mathbf{w},K)^{\frac{1}{2}}| = O_P(b^2 + \frac{1}{\sqrt{Tb}}) \tag{3.44}$$

$$\min_{\mathbf{w}\in(\mathcal{W}_d^1[0,1])^n} \hat{S}_n(\mathbf{w},K)^{\frac{1}{2}} = \min_{\mathbf{w}\in(\mathcal{W}_d^1[0,1])^n} S_n(\mathbf{w},K)^{\frac{1}{2}} + O_P(b^2 + \frac{1}{\sqrt{Tb}}) \tag{3.45}$$

Assertion a) of Theorem 1 are now immediate consequences of (3.31), (3.35), and (3.45), while Assertion c) follows from (3.44) together with $\sup_{\mathbf{w}\in(\mathcal{W}_{d[0,1]}^1)^n} |\hat{S}_n(\mathbf{w},K)^{\frac{1}{2}} = O_P(b^4 + \frac{1}{Tb})$ for $K \geq \mathbf{K}_0$.

It remains to prove Assertion b). Since $K \geq \mathbf{K}_0$ there exist some $\mathbf{w}_K \in (\mathcal{W}_d^1[0,1])^K$ with $S_n(\mathbf{w}_K,K) = 0$. Let $h_{K,1} := h_{w_{K,1}}, \ldots, h_{K,n} := h_{w_{K,n}}$ be the resulting warping functions. Therefore,

$$\hat{S}_n(\hat{\mathbf{w}},K) = \min_{\mathbf{w}\in(\mathcal{W}_{d[0,1]}^1)^n} \hat{S}_n(\mathbf{w},K) \tag{3.46}$$

$$\leq \hat{S}_n(\mathbf{w}_K,K) \leq \tfrac{1}{n}\textstyle\sum_{i=1}^n \int_0^1 (\hat{x}_i(h_{K,i}(t)) - x_i(h_{K,i}(t)))^2 dt \tag{3.47}$$

Note that $\int_0^1 (h_{K,i}^{-1})'(t)dt = h_{K,i}^{-1}(1) = 1$ for all i. By (3.38) - (3.41), by the in-

dependence of $r_{2i}(b;t)$ from x_i and h_i, and by our additional assumption on the bandwidth sequence we thus obtain

$$\begin{aligned}\mathbb{E}\left(\int_0^1 (\hat{x}_i(h_{K,i}(t)) - x_i(h_{K,i}(t)))^2 dt\right) &= \mathbb{E}\left(\int_0^1 (h_{K,i}^{-1})'(t)(\hat{x}_i(t)) - x_i(t))^2 dt\right) \\ &= \mathbb{E}\left(\int_0^1 (h_{K,i}^{-1})'(t)\mathbb{E}(r_{2i}(b;t)^2) dt\right) + o(\frac{1}{Tb}) \\ &= \frac{\sigma^2 V(\mathcal{K})}{Tb} + o(\frac{1}{Tb})\end{aligned}$$

Since $\mathbb{E}(\epsilon_{ij}^4) < \infty$ it is easily verified that $\mathbb{E}(r_{2i}(b;t)^4) = O(\frac{1}{T^2b^2})$,
and thus $\frac{1}{n}\sum_{i=1}^n Tb \int_0^1 (h_{K,i}^{-1})'(t)(r_{2i}(b;t)^2) dt \to_P \sigma^2 V(\mathcal{K})$. Hence, as $n, T \to \infty$

$$\begin{aligned}Tb\frac{1}{n}\sum_{i=1}^n \int_0^1 (\hat{x}_i(h_{K,i}(t)) - x_i(h_{K,i}(t)))^2 dt = \frac{1}{n}\sum_{i=1}^n Tb \int_0^1 (h_{K,i}^{-1})'(t)(r_{2i}(b;t)^2) dt + o_P(1) \\ = \sigma^2 V(\mathcal{K}) + o_p(1),\end{aligned}$$

which together with (3.46) leads to the desired result.

Chapter 4

Analysis of juggling data: Registering data to principal components to explain amplitude variation

Abstract

The paper considers an analysis of the juggling dataset based on registration. An elementary landmark registration is used to extract the juggling cycles from the data. The resulting cycles are then registered to functional principal components. After the registration step the paper then lays its focus on a functional principal component analysis to explain the amplitude variation of the cycles. More results about the behavior of the juggler's movements of the hand during the juggling trials are obtained by a further investigation of the principal scores.

4.1 Introduction

Functional Principal Component Analysis (FPCA) approximates a sample curve $f(t)$ as a linear combination of orthogonal basis functions $\gamma_j(t)$ with coefficients θ_j:

$$f(t) \approx \sum_{j=1}^{L} \gamma_j(t)\theta_j. \tag{4.1}$$

The principal components γ_j have the best basis property: for any fixed number L of orthogonal basis functions, the expected total squared lose is minimized. The choice of L is up to the operator, depending what accuracy is needed. It is often possible to describe the essential parts of the variations of functional data by looking only at a usually very small set of principal components and the corresponding principal scores θ_j.

However, if the curves have phase variation, even the most elementary tools of any data analysis like the pointwise mean or variance will not be able to describe the data adequately Ramsay and Silverman (2005). In such a case not only are more principal components needed to describe the same amount of variation in the data, but also further analysis based on principal components will become more difficult to interpret. In order to analyze the juggling data, we use a registration procedure introduced by Kneip and Ramsay (2008) in which the principal components are the features which are aligned. The juggling data is a nice application, because the data set contains many problems that have to be solved using different strategies.

After registering the data in Section 4.2, we perform a FPCA on the individual juggling cycles in Section 4.2.1. In Section 4.2.2 we examine the evolution of the scores of the juggling cycles over the trials where we additionally take the information from the warping functions into account. Section 4.3 summarizes our findings.

4.2 Registering the juggling data

During our analysis we are especially interested in the juggling cycles. We will use the following notation: for $t \in [0,1]$ let $f(t) = (f_x(t), f_y(t), f_z(t))$ be the spatial coordinates of a typical juggling cycle, $\mu(t) = \mathbb{E}(f(t))$ their structural mean and $\gamma_j(t) = (\gamma_{x,j}(t), \gamma_{y,j}(t), \gamma_{z,j}(t))$ be a typical principal component. We refer to chapter 8.5 of Ramsay and Silverman (2005) for an instruction on how to calculate the principal components in our multivariate case in practice. Referred to Ramsay et al. (2014), a juggling cycle is observed on the "clock time scale" which is the "juggling time" t transformed by a warping function h. As usual, we assume h to be an element of the space $\mathcal{H}$ of strictly increasing continuous functions. We hence observe

$$f[h(t)] = \mu[h(t)] + \sum_{j=1}^{\infty} \gamma_j[h(t)]\theta_j, \tag{4.2}$$

where $\theta_j = \int_0^1 \gamma_{x,j}(u)f_x(u) + \gamma_{y,j}(u)f_y(u) + \gamma_{z,j}(u)f_z(u)\,du$.

Note that by stating equation (2.37), we met the natural assumption that time and therefore also the warping function has to be the same in all three directions by introducing a common h function for all three spatial dimensions. In contrast to Ramsay et al. (2014) where the tangential velocity function is used to avoid the problem of facing three spatial dimensions at once, we will work in the original three dimensional coordinate system. By doing so we hope to find effects which are only observable within the raw data. We approach the registration of the cycles with a two stage procedure by performing what we call "macro" and "micro" warping. By macro warping we mean a very basic registration. The purpose of this registration step is to normalize the overall juggling speed such that we can properly extract the cycles from each trial. We adjusted the data for the different numbers of cycles per trial by trimming each trial down to the first 10 juggling cycles. In order to preserve as much information of the cycles as possible for further

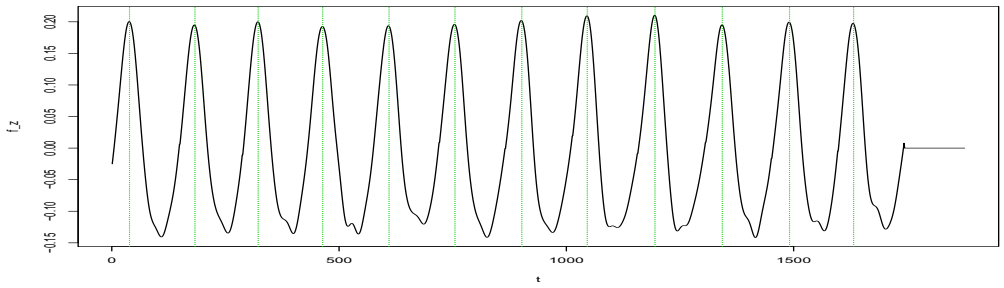


Figure 4-1: A random trial along the x direction together with the chosen landmarks.

analysis, we chose the simplest possible landmark registration which consists only of one landmark per cycle located at the local maxima occurring along the z-direction and a linear interpolation of the h function between. Since we only select one landmark per cycle, identifying it can be done very quickly.

The next step is to cut of all cycles at the landmarks such that we end up with a set of data consisting of a total of 100 cycles. This cropping implies that each of the cycles starts when one of the balls leaves the hand of the juggler to go up in the air in a high arc as seen in Figure 4-1.

During the "micro" step, we register all 100 cycles simultaneously. By doing this we perform a very precise warping on the cycles. This is in fact a more difficult task than the "macro" warping part, because a lot of different features in the cycle curves have to be taken into account. To clarify this point we displayed a random sample of 20 cycles in Figure 4-2.

It is seen from Figure 4-2 that the data needs more than just one principal component to be explained accurately. For example, by looking at the first half of this random sample along the x direction (left plot in the figure), we see variation which is obviously not induced by phase variation. Also a closer look at the middle part in the z direction (right plot) reveals a lot of variation which can not be explained by amplitude variation of a single component. Situations where we encounter more complex amplitude variations are well suited for the registration

method presented in Kneip and Ramsay (2008). This procedure has another advantage because it allows to control the intensity of the micro warping due to the smoothing parameter in equation (16) of Kneip and Ramsay (2008).

The method can be easily adapted to the multivariate case. Let D be the derivative operator, then a straightforward modification of equation (15) of Kneip and Ramsay (2008) now becomes

$$SSE(\widetilde{h}) = \int_0^1 \sum_{k=(x,y,z)} \{f_k(u) - f_k[h^{-1}(u)] - Df_k[h^{-1}(u)]\widetilde{h}(u)\}^2 \, du \tag{4.3}$$

which has to be minimized over $\widetilde{h} \in \mathcal{H}$. Finding a common warping function for multivariate data can easily be handled by using (4.3) for the SSE part occurring in the procedure of Kneip and Ramsay (2008).

The result of our alignment is shown as the black curves in Figure 4-2 where we registered the curves to 3 principal components. We observe that after the warping procedure the main features along all directions are well aligned. By looking at the first half of the left plot of Figure 4-2 one can observe the complexity of the juggling cycles along the x direction: If the cycles would belong to a one dimensional space (i.e. all cycles were random shifts from a mean curve), then all features would have been aligned. However, a more complex model underlies the data along this

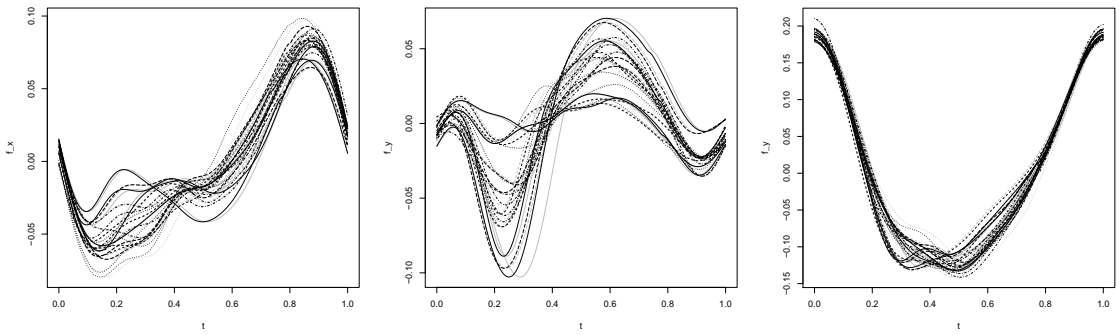


Figure 4-2: The figure shows a random sample of 20 cycles for the x, y and z direction. Registered curves are displayed black, corresponding unregistered curves grey.

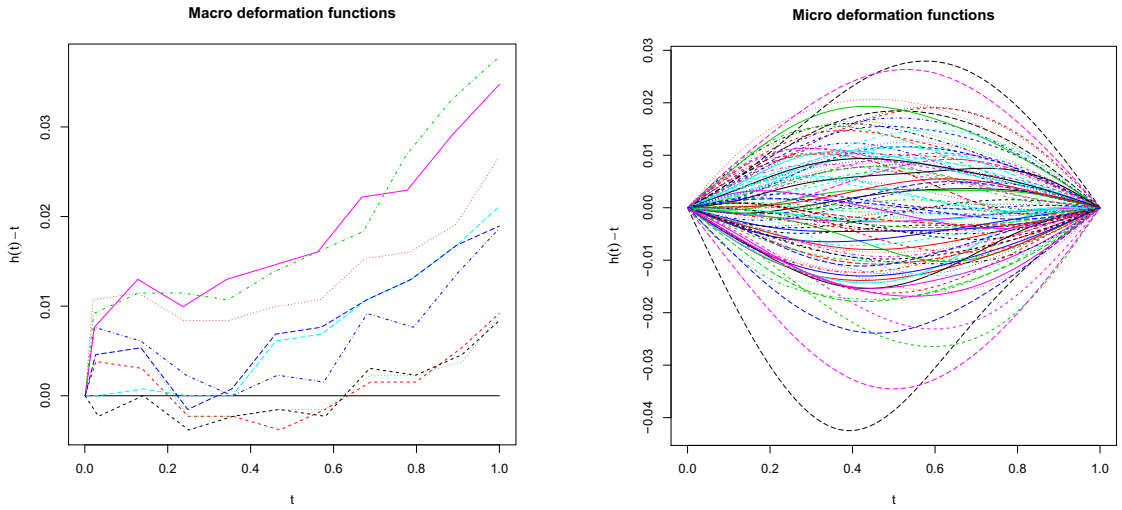


Figure 4-3: The deformation functions estimate during the macro- and microwarping.

direction and any attempt to force the data to fit in a simpler model will destroy the intrinsic features of the data; the alleged shift we are observing after the registration is in fact a part of the data. The warping functions for our alignment are displayed in Figure 4-3 through the deformation functions $h(t) - t$ obtained from the macro and micro step. Note that the deformation functions for the macro step do not end at a value of 0 since we only displayed the part of the warping functions corresponding to the first 10 cycles within the trials.

4.2.1 Analyzing the principal components

After the preprocessing steps we get suitable data to perform a FPCA. We chose to use three components to represent the data, which explain more than 80 percent of the total variance. The impact of the three principal components on each of the spatial directions of the data is displayed in Figure 4-4 where we also pictured the effect of adding and subtracting a multiple of each of the principal components to max-normalized mean curves. A closer look at Figure 4-4 reveals that the first component mainly explains the amplitude variation of the y direction while the

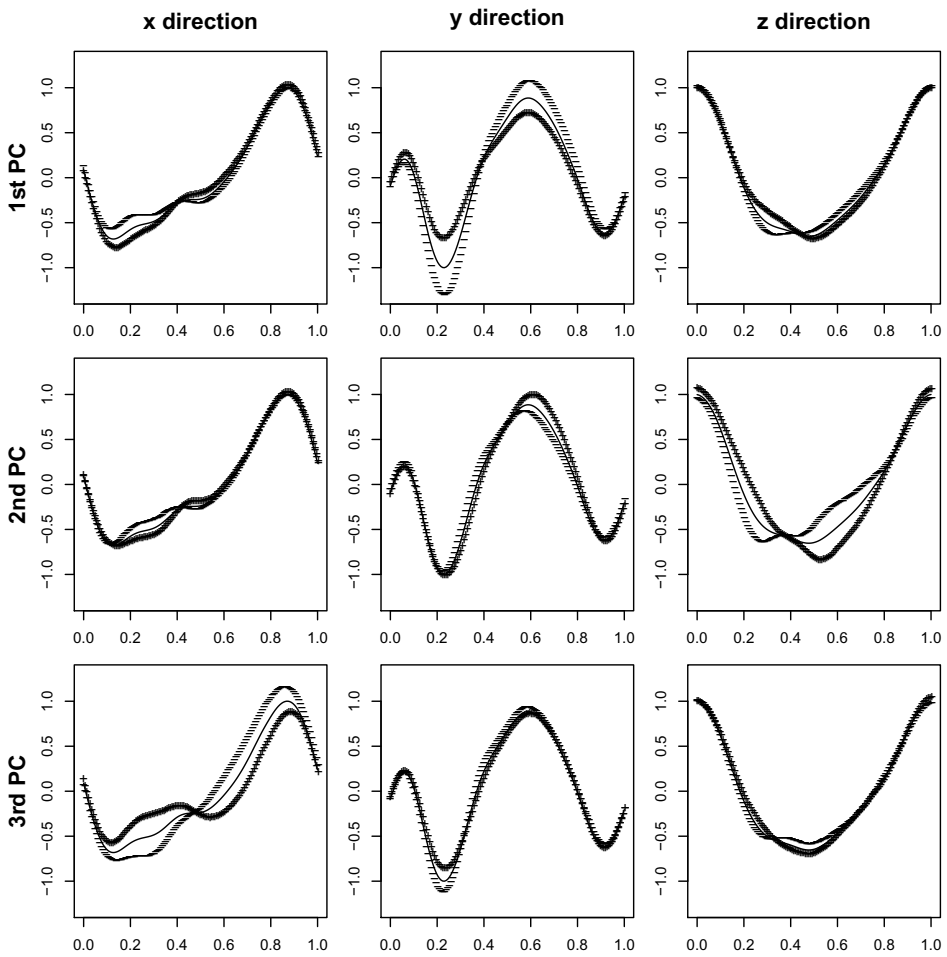


Figure 4-4: The Figure shows the effect of adding or subtracting a multiple of each of the principal components to the scaled mean curves. The columns are the spatial directions x,y,z and the rows represent the first, second and third principal component respectively.

Table 4.1: Variation of the j-th principal component due to the l-th spatial direction

	Spatial direction		
Principal Component	x	y	z
1st	0.117	**0.793**	0.091
2nd	0.053	0.185	**0.762**
3rd	**0.851**	0.100	0.049

second component explains mainly the z direction and the third component the x direction. While the effect of the first component of the movement of the jugglers hand along the x and z direction only accounts for a small shift in the beginning of the movement (the catch phase) it has an important impact for the variation across the y direction. By looking at the impact of the first component along the y direction we can see that, if the ball coming in at low arch during the catch phase is juggled right in front of the juggler, then he will overcompensate for this movement by throwing the next ball from a much greater distance to himself. Such an compensation effect can also be seen for the second component along the z direction and for the the third component along the x direction. While for the y direction the latter two components mainly adjust for the two bumps, which are influenced by the first component, individually.

The importance of the components for the three directions is summarized in Table 4.1, where we capture the variability in the j-th principal component which is accounted for by the variation in the l-th direction. More formally: for a typical principal component γ we necessarily have $\int_0^1 \gamma_x^2(u)\,du + \int_0^1 \gamma_y^2(u)\,du + \int_0^1 \gamma_z^2(u)\,du = 1$. And hence each of the summands can be interpreted to give the proportion of the variability of the component which is accounted for by the spatial direction. It is seen from the table that the y direction contributes 80% of the variation of the first component while the z and x direction can be accounted for the variation of the second and third component respectively. These values reveal that the

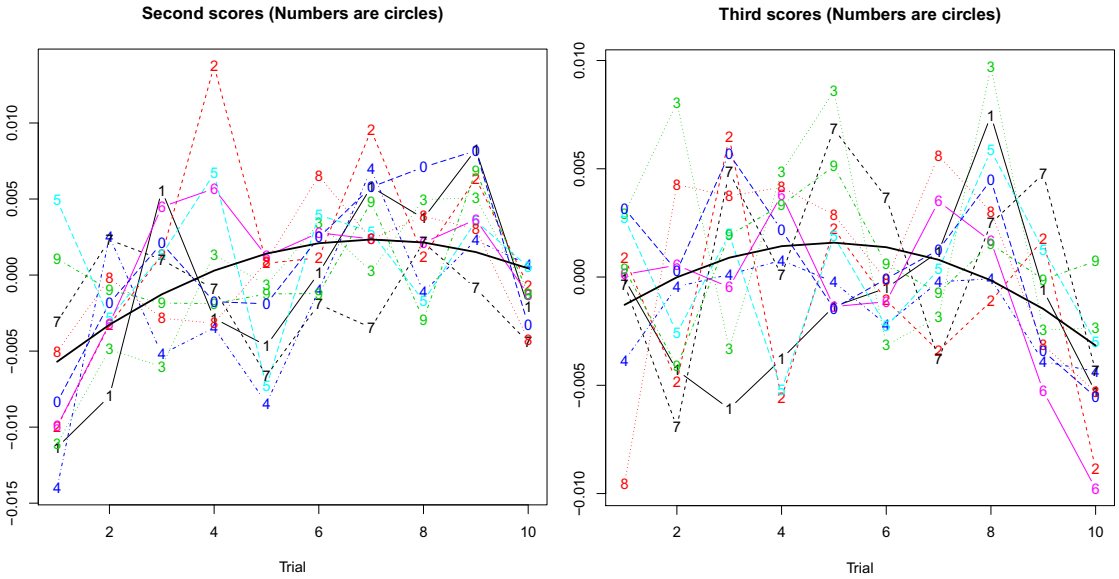


Figure 4-5: The figure shows the evolution of the scores for the cycles corresponding to the second and third principal component over the ten trials. The solid line represents the estimated regression function when we impose a quadratic model.

directions are somewhat independent in the way that each principal component represents mainly a single direction. These observations where only possible by keeping the data multivariate and not analyzing the tangential velocity function.

4.2.2 Analyzing the principal scores

If we perform activities like juggling several times, we expect something like a learning effect to happen. For a juggler this effect could be measured by the behavior of his hands along the directions, i.e. as the juggler gets more and more used to the juggling, one would expect the movements to be more efficient or at least the executions of the movements should become more homogeneous. By performing a FPCA we prepare our data for further statistical analysis which support us to answer such claims. This analysis will be performed on the scores.

Figure 4-5 shows the evolution of the scores corresponding to the second and third principal component over the ten trials. A typical principal score θ can

be modeled as a function depending on trial $k = 1, \ldots, 10$ and number of cycle $i = 1, \ldots, 10$. Figure 4-5 suggests that a polynomial regression model can capture the main message of the data. i.e. we assume

$$\theta(i, k) = \alpha_0 + \alpha_1 k + \alpha_2 k^2 + \epsilon_i. \tag{4.4}$$

Table 4.2 contains the coefficients resulting from this regression. Before we interpret the results, recall that the first component explains mostly the y direction which is on one hand less complex in terms of its variability and on the other hand is less important for a juggler. Indeed, one could imagine a perfect juggling machine which would keep this direction constant such that a juggling cycle could be described by looking solely at the x and z directions. Now, the non-significant coefficients in the first row of Table 4.2 indicate that the movement across the y direction can not be explained by the trials. This is reasonable as one would expect that an experienced juggler mainly focuses about the movement in the other two directions and any variation of his movement along the y direction from a constant value should be random.

By the significance of the coefficients of the regressions for the scores corresponding to the second and third principal component, we can conclude that there exists indeed an evolution of the scores over the trials which can essentially be described by our regression. This evolution can be regarded as some kind of a

Table 4.2: Least squares coefficients obtained from a quadratic regression of the scores on the trials. Significance codes are added in parentheses where 0 '***'; 0.001 '**'; 0.01 '*'; 1 ' '

Scores	Parameter Estimates		
	α_0	α_1	α_2
1st	-0.0040 ( )	0.0014 ( )	-0.0001 ( )
2nd	-0.0086 (***)	0.0031 (***)	-0.0002 (***)
3rd	-0.0029 (*)	0.0018 (**)	-0.0002 (***)

"learning effect". For example, in Figure 4-5 we can see that the scores will have a small value at the peak of our regression function, implying that in this area the variation of the movement of the jugglers hand is not very high and has to be close to the mean curve. This can be seen as an improvement in his juggling skills. Interestingly, the slope of the regression function decreases at the end. While this effect is subsidiary for the second principal score and could be seen as a nuisance from the simple quadratic model, it is apparent in the evolution of the scores corresponding to the third component.

Recall that the second component mainly quantifies the variation of the jugglers hand movement along the z-direction, which captures the up- and downwards movement of his hand. A negative score in the beginning of the trials indicates that he lunges out too far before throwing the ball up in the air. As the regression function for the scores of the second component approaches values close to zero, the "learning effect" becomes visible: getting used to the juggling in the later trials, he performs almost identical movements along this direction.

If we take a more precise look at the regression function of the scores corresponding to the third component, an interpretation is somewhat more complicated as we experience a significant downward slope at the last trials. Maybe the juggler gets fatigued or the behavior is caused by some kind of a psychological effect, i.e. the concentration of the juggler decreases as he knows that he only has to perform a few more trials and gets more impatient.

Taking a look at the time frame around 0.2–0.5 of the the bottom left panel of Figure 4-4, we see that a particular small value of the third component implies that his hand for catching the ball coming in from a low arch is comparable moved towards the other hand. Possibly e is learning to simplify the process of catching the ball coming in from low arch. Unfortunately this implies that he has to wind up more in order to throw the ball leaving in high arch.

We were further interested in an analysis of the warping functions themselves which was the reason to perform only a very basic "macro" warping. In this special kind of data set it is not reasonable to assume that the warping function is only a nuisance parameter because the speed of juggling might have an effect on the manner of the juggling.

To check this hypothesis we performed some further analysis on the warping functions. Note that we can not perform a FPCA on the warping functions directly, because we can not guarantee that the resulting curves are still elements of $\mathcal{H}$, i.e. strictly monotonic functions. Instead we pursue the following way out. It is well known from Ramsay and Silverman (2005) that any function $h \in \mathcal{H}$ can be represented as

$$h(t) = \int_0^t e^{W(u)}\, du,$$

where $W(t) = log[Dh(t)]$ itself is an unrestricted function. In order to analyze the warping functions h appropriately, we can use the unrestricted functions $W(t)$. We approximate $W(t)$ by using the first two principal components which explain more than 95 Percent of the variations in $W(t)$ and define by $\theta_{W,1}$, $\theta_{W,1}$ a typical scores corresponding to these two components. In Table 4.3 we computed the correlation between the scores of W and θ.

We can determine that the speed a juggling cycle is performed with has nearly no influence on the first component of a cycle. But this speed does have an effect on the second and third component which explain mostly the x and z direction. Obviously, this effect is occurs mainly through the first component of W.

Another interesting result occurs by computing the correlation between the scores of the principal components of W and and the residuals resulting from the polynomial regression in (4.4). It reveals a significant amount of correlation between these variables, i.e. a not negligible part of the residuals from (4.4) can be explained by the juggling speed of the cycles. Moreover, running a regression of

Table 4.3: The table shows the correlation between the scores corresponding to the first two components of W and the scores corresponding to the first three components of the juggling cycles

Scores of W	Scores of the cycles		
	θ_1	θ_2	θ_3
$\theta_{W,1}$	-0.0120	0.3044	-0.2351
$\theta_{W,2}$	-0.0122	0.0355	0.0013

the scores of the warping function W on the trials showed no significant coefficient. From this we can conclude that, what we identified as a learning effect, has no significant impact on the warping for a specific cycle. We hence can identify two effects which influence the scores of a juggling cycle. The first is due to learning and the second is a result which is related to the specific warping. The effects are modeled by augmenting equation (4.4) by

$$\theta(i,k) = \alpha_0 + \alpha_1 k + \alpha_2 k^2 + \beta_1 \theta_{W,1,i} + \beta_2 \theta_{W,2,i} + \epsilon_i, \tag{4.5}$$

where $\theta_{W,j,i}$ is the score of the i-th cycle corresponding to the j-th principal component of the function W. Estimated coefficients are given in Table 4.4, from where it can be seen that neither the speed the juggling cycles are performed with, nor the trials have an impact on the movement of the jugglers hand along the y direction. Moreover, it can be seen that there is a connection between the scores of a juggling cycles and the speed of the juggling.

Table 4.4: The table shows the results from an Regression of the cycle scores on the trial number, squared trial number as well as the scores from W with corresponding coefficients β_1 and β_2. Significance codes are added in parentheses where 0 '***'; 0.001 '**'; 0.01 '*'; 1 ' '

Scores	Parameter Estimates				
	α_0	α_1	α_2	β_1	β_2
1st	-0.0042 ()	0.0014 ()	-0.0001 ( )	-0.0009 ()	0.0000()
2nd	-0.0081 (***)	0.0030 (***)	-0.0002 (***)	0.0034 (**)	0.0025 ()
3rd	-0.0033 (*)	0.0019 (***)	-0.0002 (***)	-0.0027 (*)	-0.0009 ()

4.3 Summary

We analyzed the juggling data by combining two registration methods. First we used an elementary landmark registration in order to crop the individual juggling cycles, which were the focus of our analysis. In order to perform a refined warping of the juggling cycles in a second step, we generalized the registration method from Kneip and Ramsay (2008) to the multivariate nature of the data. We analyze the registered data by performing a FPCA using three principal components where we observed that each of the components essentially quantified the variation across a single spatial direction.

More specific information about the behavior of the juggler is contained in the scores which we studied in dependence on the trials. By doing so, we were able to identify some kind of learning effect over the trials. The movement of the jugglers hand for throwing a ball up in the air levels out over the trials. After applying an alignment procedure one should not forget about the warping functions. Interpreting the warping functions can not only be a very interesting task for itself, but they can contain important additional information which can be helpful to analyze the data.

Bibliography

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Aneurisk-Team (2012). AneuriskWeb project website, http://ecm2.mathcs.emory.edu/aneuriskweb. Web Site.

Bahra, B. (1997). Implied risk-neutral probability density functions from option prices: theory and application. Bank of England working papers 66, Bank of England.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221.

Banz, R. W. (1978). Prices for State-contingent Claims : Some Estimates and Applications. *The Journal of Business*, 51(4):653–672.

Bates, D., Mullen, K. M., Nash, J. C., and Varadhan, R. (2014). *minqa: Derivative-free optimization algorithms by quadratic approximation*. R package version 1.2.3.

Bates, D. S. (2006). Maximum likelihood estimation of latent affine processes. *Review of Financial Studies*, 19(3):909–965.

Benko, M., Härdle, W., and Kneip, A. (2009). Common functional principal components. *The Annals of Statistics*, 37(1):1–34.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311.

Bliss, R. R. and Panigirtzoglou, N. (2002). Testing the stability of implied probability density functions. *Journal of Banking & Finance*, 26(2-3):381–422.

Bollen, N. P. B. and Whaley, R. E. (2004). Does Net Buying Pressure Affect the Shape of Implied Volatility Functions? *Journal of Finance*, 59:711–753.

Bookstein, F. (1978). *The Measurement of Biological Shape and Shape Change*. Springer.

Bookstein, F. (1997). *Morphometric Tools for Landmark Data: Geometry and Biology*. Geometry and Biology. Cambridge University Press.

Bosq, D. (2000). *Linear processes in function spaces*. Springer.

Breeden, D. T. and Litzenberger, R. H. (1987). Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51:621–651.

Brigo, D. and Mercurio, F. (2002). Lognormal-mixture dynamics and calibration to market volatility smiles. *International Journal of Theoretical and Applied Finance*, 5:427–446.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

Cardot, H., Mas, A., and Sarda, P. (2007). Clt in functional linear regression models. *Probability Theory and Related Fields*, 138(3-4):325–361.

Claeskens, G., Silverman, B. W., and Slaets, L. (2010). A multiresolution approach to time warping achieved by a bayesian prior posterior transfer fitting strategy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):673–694.

Constantinides, G. M. and Lian, L. (2015). The Supply and Demand of S&P 500 Put Options. *SSRN Electronic Journal*.

Cont, R. and da Fonseca, J. (2002). Dynamics of implied volatility surfaces. *Quantitative Finance*, 2(1):45—60.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12:136–154.

Di, C., Crainiceanu, C., Caffo, B., and Punjabi, N. (2009). Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1):458.

Ehrgott, M. (2000). *Multicriteria Optimization*. Lecture notes in economics and mathematical systems. Springer.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1997). Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, 49(1):79–99.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4):2008–2036.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Number 66 in Monographs on statistics and applied probability series. Chapman & Hall, London [u.a.].

Fengler, M., Härdle, K. W., and Villa, C. (2003). The dynamics of implied volatilities: A common principal components approach. *Review of Derivatives Research*, 6:179–202.

Fengler, M. R., Härdie, W., and Schmidt, P. (2002). *Applied Quantitative Finance: Theory and Computational Tools*, chapter The Analysis of Implied Volatilities, pages 127–144. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis, theory and practice*. Springer.

Gârleanu, N., Pedersen, L. H., and Poteshman, A. M. (2009). Demand-based option pricing. *Review of Financial Studies*, 22:4259–4299.

Gasser, T. and Kneip, A. (1995). Searching for structure in curve sample. *Journal of the American Statistical Association*, 90(432):1179–1188.

Gervini, D. and Gasser, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika*, 92(4):801–820.

Geweke, J. and Porter-Hudak, S. (1983). The Estimation and Application of Long Memory Time Series Models. *Journal of Time Series Analysis*, 4(4):221–238.

Grith, M., Härdle, W. K., and Park, J. (2013). Shape invariant modeling of pricing kernels and risk aversion. *Journal of Financial Econometrics*, 11:370–399.

Grith, M., Härdle, W. K., and Schienle, M. (2012). *Handbook of Computational Finance*, chapter Nonparametric Estimation of Risk-Neutral Densities, pages 277–305. Springer Verlag.

Gu, J., Li, Q., and Yang, J.-C. (2015). Multivariate local polynomial kernel estimators: Leading bias and asymptotic distribution. *Econometric Reviews*, 34(6-7):978–1009.

Hadjipantelis, P. Z., Aston, J. A. D., Müller, H. G., and Evans, J. P. (2015). Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin chinese. *Journal of the American Statistical Association*, 110(510):545–559. PMID: 26692591.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of The Royal Statistical Society Series B*, 68(1):109–126.

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika*, 77(2):pp. 415–419.

Hall, P., Müller, H.-G., and Yao, F. (2009). Estimation of functional derivatives. *The Annals of Statistics*, 37(6A):3307–3329.

Härdle, K. W. and Lopez-Cabrera, B. (2012). The implied market price of weather risk. *Applied Mathematical Finance*, 19(1):59–95.

Härdle, W. and Simar, L. (2012). *Applied Multivariate Statistical Analysis*. Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24.

Jackwerth, J. C. (1999). Option-implied risk-neutral distributions and implied binomial trees: a literature review. *Journal of Derivatives*, 2:66–82.

Jammalamadaka, S., Sengupta, A., and Sengupta, A. (2001). *Topics in Circular Statistics*. Series on multivariate analysis. World Scientific.

Kneip, A. and Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20(3):1266–1305.

Kneip, A., Li, X., MacGibbon, K. B., and Ramsay, J. O. (2000). Curve registration by local regression. *Canadian Journal of Statistics*, 28(1):19–29.

Kneip, A. and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165.

Kneip, A. and Utikal, K. J. (2001). Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association*, 96(454):519–542.

Leng, X. and Müller, H.-G. (2006a). Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22(1):68–76.

Leng, X. and Müller, H.-G. (2006b). Time ordering of gene coexpression. *Biostat*, 7(4):569–584.

Liu, B. and Müller, H.-G. (2009). Estimating derivatives for samples of sparsely observed functions, with application to online auction dynamics. *Journal of the American Statistical Association*, 104(486):704–717.

Lo, A. W. (1991). Long-Term Memory in Stock Market Prices. *Econometrica*, 59:1279–1313.

Majer, P., Mohr, P., Heekeren, H., and Härdle, K. W. (2015). Portfolio decisions and brain reactions via the cead method. *Psychometrika*, pages 1–23.

Mallows, C. (1973). Some comments on c_p. *Technometrics*, 15:661–675.

Mas, A. (2002). Weak convergence for the covariance operators of a hilbertian linear process. *Stochastic Processes and their Applications*, 99(1):117 – 135.

Mas, A. (2008). Local functional principal component analysis. *Complex Analysis and Operator Theory*, 2(1):135–167.

Mason, J., Rodriguez, G., and Seatzu, S. (1993). Orthogonal splines based on b-splines with applications to least squares, smoothing and regularisation problems. *Numerical Algorithms*, 5(1):25–40.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Ser. Anal*, 17:571–599.

Mixon, S. (2002). Factors explaining movements in the implied volatility surface. *Journal of Futures Markets*, 22(10):915–937.

Munk, A., Bissantz, N., Wagner, T., and Freitag, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society Series B*, 67(1):19–41.

Parodi, A., Patriarca, M., Sangalli, L., Secchi, P., Vantini, S., and Vitelli, V. (2015). *fdakma: Functional Data Analysis: K-Mean Alignment*. R package version 1.2.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2:559–572.

Petersen, A. and Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a hilbert space. *Ann. Statist.*, 44(1):183–218.

Poss, D. and Wagner, H. (2014). Analysis of juggling data: Registering data to principal components to explain amplitude variation. *Electron. J. Statist.*, 8(2):1825–1834.

Powell, M. (2006). The newuoa software for unconstrained optimization without derivatives. In Di Pillo, G. and Roma, M., editors, *Large-Scale Nonlinear Optimization*, volume 83 of *Nonconvex Optimization and Its Applications*, pages 255–297. Springer US.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441.

Ramsay, J. O. (1996). Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(3):495–508.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):365–375.

Ramsay, J. O., Gribble, P., and Kurtek, S. (2014). Description and processing of functional data arising from juggling trajectories. *Electron. J. Statist.*, 8(2):1811–1816.

Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd edition.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1):233–243.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22(3):1346–1370.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49.

Sangalli, L. M., Secchi, P., Vantini, S., and Veneziani, A. (2009). A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, 104(485):37–48.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010). k-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233.

Silverman, B. W. (1995). Incorporating Parametric Effects into Functional Principal Components Analysis. *Journal of the Royal Statistical Society*, 57(4):673–689.

Slaets, L., Claeskens, G., and Hubert, M. (2012). Phase and amplitude-based clustering for functional data. *Computational Statistics & Data Analysis*, 56(7):2360 – 2374.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. of the Cell*, 9:3273–3297.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011). Registration of Functional Data Using Fisher-Rao Metric. http://arxiv.org/abs/1103.3817.

Staicu, A.M.and Crainiceanu, C. M. and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics*, 11(2):177–194.

Tucker, J. D. (2014). *fdasrvf: Elastic Functional Data Analysis*. R package version 1.4.2.

Tuddenham, R. and Snyder, M. (1954). *Physical Growth of California Boys and Girls from Birth to Eighteen Years*. Publications in child development. University of California Press.

von Neumann, J., Kent, R. H., Bellinson, H. R., and Hart, B. I. (1941). The mean square successive difference. *The Annals of Mathematical Statistics*, 12(2):pp. 153–162.

Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *Annals of Statistics*, 25(3):1251–1276.

Wang, K. and Gasser, T. (1998). Asymptotic and bootstrap confidence bounds for the structural average of curves. *The Annals of Statistics*, 26(3):972–991.

Wang, K. and Gasser, T. (1999). Synchronizing sample curves nonparametrically. *The Annals of Statistics*, 27(2):439–460.

Weron, R. (2002). Estimating long-range dependence: finite sample properties and confidence intervals. *Physica A: Statistical Mechanics and its Applications*, 312(1-2):285–299.

Zhao, X., Marron, J. S., and Wells, M. T. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica*, 14:789–808.

Zipunnikov, V., Caffo, B. C., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. M. (2011). Functional principal component model for high-dimensional brain imaging. *NeuroImage*, 58(3):772–784.

Heiko Wagner wurde am 09. November 1982 in Euskirchen geboren. Nach der Schulzeit an der kath. Grundschule in Arloff und später am Städt. Sankt Michael Gymnasium in Bad Münstereifel schloss Heiko Wagner 2002 mit dem Abitur ab. Im Anschluss absolvierte er einen neun neunmonatigen Wehrdienst in Kastellaun und Bonn.

Im Anschluss begann er 2003 das Studium der Volkswirtschaftslehre an der Universität Bonn welches 2009 mit einem Diplom abgeschlossen wurde. Anschließend begann er das Promotionsstudium im selben Fach in der statistischen Abteilung des "Instituts für Finanzmarktökonomie und Statistik" unter Prof. Alois Kneip.