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

Maria Grith,[1] Wolfgang K. Härdle,[1,3] Alois Kneip [2] and Heiko Wagner [2]

[1] Ladislaus von Bortkiewicz Chair of Statistics and C.A.S.E. - Center for Applied Statistics and Economics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany
[2] Institute for Financial Economics and Statistics, Department of Economics, Rheinische Friedrich-Wilhelms-Universität Bonn, Adenauerallee 24-26, 53113 Bonn
[3] Sim Kee Boon Institute for Financial Economics, Singapore Management University, 81 Victoria Street, Singapore 188065

This version: June 13, 2016

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 then 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.

Keywords: functional principal component, dual method, derivatives, high-dimensional spatial curves, state price densities

JEL codes: C13, C14, G13

⋆ The financial support from the German Research Foundation for the joint project no. 70102424 "Functional Principal Components for Derivatives and Higher Dimensions", between Humboldt-Universität zu Berlin and Rheinische Friedrich-Wilhelms-Universität Bonn, is gratefully acknowledged. We would like to thank as well the Collaborative Research Center 649 “Economic Risk” for providing the data and the International Research Training Group (IRTG) 1792 “High-Dimensional Non-Stationary Time Series Analysis”, Humboldt-Universität zu Berlin for additional funding.

1 Introduction

Over the last two decades, functional data analysis became a popular tool to handle data entities that are 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., **?**. 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., **?**, pricing kernel, e.g., **?** or the market price of risk, e.g., **?**. 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 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 **?**. 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 **?**, followed by subsequent studies that fostered the applicability of the method to samples of observed noisy curves. **?**, 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 **?**, **?**, **?**, **?** and **?**. To date, simple, one-dimensional spatial curves are well understood from both numerical and theoretical perspective. In the 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 **?** and **?**. 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 amount 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 **?** 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, **?** split the recorded data into smaller parts to make it manageable. The method called multilevel FPCA has been developed through previous studies, see **?**, **?**, and 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 spaces. We present two alternative methods to obtain the derivatives. The paper is organized as following: the theoretical framework, estimation procedure and statistical properties are derived through Section 2. Our empirical study in Section 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 Methodology

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.

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

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

$$X(t) = \sum_{r=1}^{\infty} \delta_r \gamma_r(t), \tag{2}$$

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

$$X^{(d)}(t) \stackrel{\text{def}}{=} \frac{\partial^d}{\partial t^d} X(t) = \frac{\partial^{d_1}}{\partial t_1^{d_1}} \cdots \frac{\partial^{d_g}}{\partial t_g^{d_g}} X(t_1,\ldots,t_g), \tag{3}$$

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 (1) and (2), we consider two approaches to model a decomposition for derivatives $X^{(d)}$. The first one can be 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. Then the principal components $\varphi_r^{(d)}$ are solutions to

$$\int_{[0,1]^g} \sigma^{(d)}(t,v)\varphi_r^{(d)}(v)dv = \lambda_r^{(d)}\varphi_r^{(d)}(t). \tag{4}$$

For nonderivatives, i.e., $|d| = 0$, we introduce the following notation $\varphi_r^{(0)}(t) \equiv \gamma_r(t)$. Similarly to (2), the decomposition of $X^{(d)}$ with principal components $\varphi_r^{(d)}(t)$ is

$$X^{(d)}(t) = \sum_{r=1}^{\infty} \delta_r^{(d)}\varphi_r^{(d)}(t), \tag{5}$$

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)

$$X^{(d)}(t) = \sum_{r=1}^{\infty} \delta_r\gamma_r^{(d)}(t), \tag{6}$$

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

$$\int_{[0,1]^g} \frac{\partial^d}{\partial v^d}\left(\sigma(t,v)\gamma_r(v)\right)dv = \lambda_r\gamma_r^{(d)}(t). \tag{7}$$

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 $a_{ri} \in \mathbb{R}$ such that $\varphi_r^{(d)}(t) = \sum_{i=1}^{\infty} a_{ri}\gamma_i^{(d)}(t)$. However, if we consider a truncation of (2) after a finite number of components this is no longer true in general. An advantage of using $\varphi_r^{(d)}(t)$ instead of $\gamma_r^{(d)}(t)$ is that the decomposition of the covariance function of the derivatives give orthonormal basis that fulfill the best basis property, such that for any fixed $L \in \mathbb{N}$ and every other orthonormal basis system φ'

$$E||X^{(d)} - \sum_{r=1}^{L}\left\langle X^{(d)}, \varphi_r^{(d)}\right\rangle\varphi_r^{(d)}|| \leq E||X^{(d)} - \sum_{r=1}^{L}\left\langle X^{(d)}, \varphi_r'\right\rangle\varphi_r'||. \tag{8}$$

This guarantees that by using $\varphi_r^{(d)}(t),\ r = 1,\ldots,L$ 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 rates of convergence. However, if we assume a L-dimensional function space from the beginning, which is equivalent to a factor model setup, this advantage vanishes, because it is possible to derive a basis system with the same features using $\text{span}(\gamma^{(d)})$. In particular, this can be achieved by deriving the function space of $\gamma_r^{(d)}(t),\ r = 1,\ldots,L$ and performing a spectral decomposition of the finite-dimensional function space to get an orthonormal basis system fulfilling (8).

2.2 Sample inference

Let $X_1,\ldots,X_N \in L^2([0,1]^g)$ be an i.i.d. sample of smooth curves with continuous covariance function. For the sample of N observed curves, the empirical approximation of the covariance function is given by the sample counterpart

$$\hat{\sigma}^{(d)}(t,v) = \frac{1}{N}\sum_{i=1}^{N} X_i^{(d)}(t)X_i^{(d)}(v) \tag{9}$$

and of the covariance operator by

$$\hat{\Gamma}_N^{(d)}\hat{\varphi}_r^{(d)}(t) = \int_{[0,1]^g} \hat{\sigma}^{(d)}(t,v)\hat{\varphi}_r^{(d)}(v)\,dv, \tag{10}$$

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 **?** or **?**. The loadings corresponding to each realization X_i can be estimated via the empirical eigenfunctions as $\hat{\delta}_{ri}^{(d)} = \int_{[0,1]^g} X_i^{(d)}(t)\hat{\varphi}_r^{(d)}(t)\,dt$.

2.3 The model

In most applications, the curves are only observed at discrete points and the data is corrupted by additive noise. To model these issues, we assume that each curve in the sample is observed at independent randomly-distributed points $t_i = (t_{i1},\ldots,t_{iT_i})$, $t_{ik} \in [0,1]^g$, $k = (1,\ldots,T_i)$ from a continuous distribution with density f such that $\inf_{t\in[0,1]^g} f(t) > 0$. Our model is then given by

$$Y_i(t_{ik}) = X_i(t_{ik}) + \varepsilon_{ik} = \sum_{r=1}^{\infty} \delta_{ri}\gamma_r(t_{ik}) + \varepsilon_{ik}. \tag{11}$$

For each curve i, ε_{ik} are i.i.d. random variables, $\mathsf{E}\left[\varepsilon_{ik}\right] = 0$ and $\mathrm{Var}\left(\varepsilon_{ik}\right) = \sigma_{i\varepsilon}^2$ and ε_{ik} is independent of X_j, $j = 1,\ldots,N$.

2.4 Estimation procedure

1. 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 **?** to estimate density functions and later adapted by **?** for general functions. Let $\nu = (\nu_1,\ldots,\nu_g)^\top$, $\nu_i \in \mathbb{N}^+$, $|\nu| < \rho \leq m$ a and $M^{(\nu)}$ be the dual matrix of $\hat{\sigma}^{(\nu)}(t,v)$ from (9) consisting of entries

$$M_{ij}^{(\nu)} = \int_{[0,1]^g} X_i^{(\nu)}(t)X_j^{(\nu)}(t)\,dt. \tag{12}$$

Let $p_r^{(\nu)} = (p_{1r}^{(\nu)},\ldots,p_{Nr}^{(\nu)})$ be the eigenvectors and $l_r^{(\nu)}$ the corresponding ordered eigenvalues of matrix $M^{(\nu)}$. In particular, the cases $\nu = d$ relates to equation (4) and gives an empirical version of (5) by

$$\hat{\varphi}_r^{(d)}(t) = \frac{1}{\sqrt{l_r^{(d)}}}\sum_{i=1}^{N} p_{ir}^{(d)}X_i^{(d)}(t)\,,\ \hat{\lambda}_r^{(d)} = \frac{l_r^{(d)}}{N} \text{ and } \hat{\delta}_{ri}^{(d)} = \sqrt{l_r^{(d)}}\,p_{ir}^{(d)}. \tag{13}$$

Important for the representation given in equation (6) are the eigenvalues and eigenvectors of $M^{(0)}$ denoted by $l_r \overset{\text{def}}{=} l_r^{(0)}, p_r \overset{\text{def}}{=} p_r^{(0)}$ and $\hat{\gamma}_r(t) = \hat{\varphi}_r^{(0)}(t)$ respectively. It is straightforward to derive

$$\hat{\gamma}_r^{(d)}(t) = \frac{1}{\sqrt{l_r}}\sum_{i=1}^{N} p_{ir}X_i^{(d)}(t). \tag{14}$$

2. Quadratic integrated regression — Before deriving estimators of $M^{(0)}$ and $M^{(d)}$ using the model from section 2.3 we outline some results needed to construct these estimators. Consider a generic curve $Y(t_l)$ observed at points $l = 1,\ldots,T$ generated as in equation (11). We use the following notation: $k = (k_1,\ldots,k_g)^\top$, $k_l \in \mathbb{N}^+$, $|k| = \sum_{l=1}^g k_g$, $t^k = t_1^{k_1} \times \ldots \times t_g^{d_g}$, $k! = k_1! \times \ldots \times k_g!$ and $a \circ t = (a_1 t_1,\ldots,a_g t_g)^\top$, for $t \in \mathbb{R}^g$ and $a \in \mathbb{R}^g$.

Consider a multivariate local polynomial estimator $\hat{\beta}(t) \in \mathbb{R}^\rho$ that solves

$$\min_{\beta(t)} \sum_{l=1}^{T} \left[Y(t_l) - \sum_{0 \le |k| \le \rho} \beta_k(t)(t_l - t)^k \right]^2 K_b(t_l - t). \tag{15}$$

K_b is any non-negative, symmetric and bounded multivariate kernel function and $b = (b_1,\ldots,b_g)$ is the diagonal of a $g \times g$ bandwidth matrix, whose off-diagonal entries are zero. In our empirical study, we take K_b to be a product of g univariate Epanechnikov kernel functions as described, for example, in **?**.

As noted by **?** the solution of the minimization problem (15) can also be represented using a weight function W_ν^T, see Appendix 5.2, which define

$$\hat{X}_b^{(\nu)}(t) = \nu!\hat{\beta}_\nu(t) = \nu! \sum_{l=1}^{T} W_\nu^T \left((t_l - t) \circ (b^{-1})^\top \right) Y(t_l), \tag{16}$$

where the argument of W_ν^T depends on the distance between point t_l and location t.

Local polynomial regression estimators are better suited to estimate integrals 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 **?**. We propose the following estimator for the squared integrated functions $\int_{[0,1]^g} X^{(\nu)}(t)^2 dt$

$$\begin{aligned} \theta_{\nu,\rho} = \int_{[0,1]^g} \nu!^2 \sum_{k=1}^{T} \sum_{l=1}^{T} W_\nu^T \left((t_k - t) \circ (b^{-1})^\top \right) W_\nu^T \left((t_l - t) \circ (b^{-1})^\top \right) Y(t_l) Y(t_k) dt \\ -\nu!^2 \hat{\sigma}_\varepsilon^2 \int_{[0,1]^g} \sum_{k=1}^{T} W_\nu^T \left((t_k - t) \circ (b^{-1})^\top \right)^2 dt. \end{aligned} \tag{17}$$

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

Lemma 2.1 *Under Assumptions 5.1-5.4, X is $m \ge 2|\nu|$ times differentiable, the local polynomial regression has order ρ with $|\nu| < \rho \le m$ and $|\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2| = \mathcal{O}_P(T^{-1/2})$ then as $T \to \infty$ and $\max(b)^{\rho+1} b^{-\nu} \to 0$, $\frac{\log(T)}{T \max(b)^g} \to 0$ as $T b_1 \times \ldots \times b_g b^{4\nu} \to \infty$,*

$$\begin{aligned} \mathsf{E}\left[\theta_{d,\rho} \right] - \int_{[0,1]^g} X^{(\nu)}(t)^2 dt &= \mathcal{O}_p(\max(b)^{\rho+1} b^{-\nu}) \\ \mathrm{Var}(\theta_{d,\rho}) &= \mathcal{O}_p\left(\frac{1}{T^2 b_1 \times \ldots \times b_g b^{4\nu}} + \frac{1}{T} \right). \end{aligned} \tag{18}$$

The proof of Lemma 2.1 is given in Appendix 5.2.

3. Estimation of $M^{(0)}$ and $M^{(d)}$ — The curves Y_i in equation (11) are assumed to be observed at different random points. We consider uniformly sampled points

$t_1,\ldots,t_T \in [0,1]^g$ with $T = \min_{i\in 1,..,N} T_i$ and replace the integrals in (17) with the Riemann sums

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

where $w_\nu^T(t_{ik},t_{jl},b) := T^{-1}\sum_{m=1}^{T} W_\nu^T\left((t_{ik}-t_m)\circ(b^{-1})^\top\right) W_\nu^T\left((t_{jl}-t_m)\circ(b^{-1})^\top\right)$. 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, which gives an estimate at a common grid, and has been analyzed in Lemma (2.1). The other one is from approximating the integral in (17) with a sum in the above equation, for the observed curves given in equation (11). The error of the integral approximation is of order $T^{-1/2}$, see Appendix (5.3). By requiring that $\hat{\sigma}_{i\varepsilon}$ used for the diagonal correction is also $T^{-1/2}$ consistent, we get $T^{-1/2}$ consistency for all terms of $\hat{M}^{(\nu)}$.

Proposition 2.2 *Under the requirements of Lemma 2.1*

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

By Proposition 2.2 estimating the dual matrix of $\sigma^{(d)}$ gives an asymptotic higher bias and also a higher variance then estimating the dual matrix for σ. This effect becomes even worse in high-dimensions. However, by using local polynomial regression with large ρ one can still get parametric rates within each method.

Remark 2.3 *Under the assumptions of Lemma 2.1 and using Proposition 2.2 we can derive estimators for $M^{(\nu)}$, which attain parametric rates. If $m \geq \rho \geq \frac{g}{2} - 1 + 3\sum_{l=1}^{g}\nu_l$, $b = T^{-\alpha}$ 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 $m \geq \rho$ - and higher bandwidth to achieve the same parametric convergence rate as the estimator for $M^{(0)}$.

In Lemma 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 2.3. By Assumption 5.2, in the univariate case a simple class of estimators for $\sigma_{i\varepsilon}^2$, which archive the desired convergence rate are given by successive differentiation, see **?** and **?**. However, as pointed out in **?** difference estimators are no longer consistent if $g \geq 4$ due to the curse of dimensionality. A possible solution is to generalize the kernel based variance estimator proposed by **?** to higher dimensions with

$$\hat{\sigma}_{\varepsilon,i}^2 = \frac{1}{\nu_i}\sum_{l=1}^{T_i}\left(Y_i(t_{il}) - \sum_{k=1}^{T_i} w_{ilk} Y(t_{ik})\right)^2, \tag{19}$$

where $w_{ilk} = K_{r,H}(t_{il}-t_{ik}) / \sum_{k=1}^{T_i} K_{r,H}(t_{il}-t_{ik})$ and $\nu_i = T_i - 2\sum_l w_{ilk} + \sum_{l,k} w_{ilk}^2$ and $K_{r,H}$ is a g-dimensional product kernel of order r with bandwidth matrix H. **?** 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 (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 (12). This estimator then is given by

$$\tilde{M}^{(0)}_{ij} = \begin{cases} \frac{1}{T}\sum_{l=1}^{T} Y_i(t_l)Y_j(t_l) \text{ if } i \neq j \\ \frac{1}{T}\sum_{k=1}^{T} Y_i(t_l)^2 - \hat{\sigma}^2_{i\varepsilon} \text{ if } i = j \end{cases}. \tag{20}$$

In Appendix (5.3) we verify that if a common random grid is observed the convergence rate of $\tilde{M}^{(0)}_{ij}$ does not depend on g.

When working with more than one spatial dimension, in practice the 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 the data is recorded we recommend using a common random grid which keeps computing time as well as required space to the store the data to a minimum but still gives parametric convergence rates for the estimator of $M^{(0)}_{ij}$. If $T \gg N$ equation (20) gives a straightforward explanation why from a computational point of view the choice of the dual matrix is preferable to derive the eigendecomposition of the covariance operator, because taking sums has a computational cost that is linear.

4. Estimating the basis functions — In the following, we use again the notations $\nu = d$ if we refer to the estimation of (5) and $\nu = (0, \ldots, 0)^\top$ for (6). A spectral decomposition of $\hat{M}^{(\nu)}$ is applied to obtain the eigenvalues $\hat{l}^{(\nu)}_r$ and eigenvectors $\hat{p}^{(\nu)}_r$ for $r, j = 1, \ldots, N$. This gives straightforward empirical counterparts $\hat{\lambda}^{(\nu)}_{r,T} = \hat{l}^{(\nu)}_r / N$ and $\hat{\delta}^{(\nu)}_{rj,T} = \sqrt{\hat{l}^{(\nu)}_r}\,\hat{p}^{(\nu)}_{rj}$.

To estimate $\varphi^{(d)}_r$ and $\gamma^{(d)}_r$, a suitable estimator for $X^{(d)}_i$, $r, j = 1, \ldots, N$ is needed. We use a local polynomial kernel estimator, denoted $\hat{X}^{(d)}_{i,h}$, similarly to (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 is that the bias and variance can be derived analytically. For the univariate case these results can be found in **?** and for the multivariate case in **?**. We summarize them in terms of order of convergence below

$$\begin{aligned} \mathsf{E}\left[X^{(d)}_j(t) - \hat{X}^{(d)}_{j,h}(t)\right] &= \mathcal{O}_p(\max(h)^{p+1}h^{-d}) \\ \mathrm{Var}\left(\hat{X}^{(d)}_{j,h}(t)\right) &= \mathcal{O}_p\left(\frac{1}{Th_1 \times \ldots \times h_g h^{2d}}\right). \end{aligned} \tag{21}$$

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

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

$$\text{Var}\left(\frac{1}{\sqrt{l_r^{(v)}}}\sum_{i=1}^{N} p_{ir}^{(v)}\hat{X}_{i,h}^{(d)}(t)\right) = \frac{1}{l_r^{(v)}}\sum_{j=1}^{N}\left(p_{jr}^{(v)}\right)^2 \text{Var}\left(\hat{X}_{j,h}^{(d)}(t)\right) + \mathcal{O}_p\left(\frac{1}{NTh_1\times\ldots\times h_g h^{2d}}\right)$$
$$= \mathcal{O}_p\left(\frac{1}{NTh_1\times\ldots\times h_g h^{2d}}\right).$$

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 2.4 *Under the requirements of Lemma 2.1, Assumptions 5.6 and 5.7, Remark 2.3, and for* $\inf_{s\neq r}|\lambda_r - \lambda_s| > 0$, $r,s = 1,\ldots,N$ *and* $\max(h)^{p+1}h^{-d} \to 0$ *with* $NTh_1\ldots h_g h^{2d} \to \infty$ *as* $T, N \to \infty$ *we obtain*

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

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

A proof of Proposition 2.4 is provided in Appendix 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 3.1.

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) after L basis functions. If a true factor model is assumed, the basis representation to reconstruct $X^{(d)}$ is arbitrary in sense that

$$X^{(d)}(t) = \sum_{r=1}^{L}\delta_r\gamma_r^{(d)}(t) = \sum_{r=1}^{L_d}\delta_r^{(d)}\varphi_r^{(d)}(t). \tag{22}$$

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 $\gamma_r^{(d)}(t) = \sum_{s\neq r} a_{sr}\gamma_s^{(d)}(t)$.

Thus, based on the estimates from Section 2.4 the estimated derivatives of the functions are given by

$$\hat{X}_{i,FPCA_1}^{(d)}(t) \overset{\text{def}}{=} \sum_{r=1}^{L}\hat{\delta}_{ir,T}\hat{\gamma}_{r,T}^{(d)}(t) \;\approx\; \hat{X}_{i,FPCA_2}^{(d)}(t) \overset{\text{def}}{=} \sum_{r=1}^{L_d}\hat{\delta}_{ir,T}^{(d)}\hat{\varphi}_{r,T}^{(d)}(t). \tag{23}$$

Proposition 2.5 *Let* $NT^{-1} \to 0$, *together with the requirements of Proposition 2.4 the true curves can be reconstructed with*

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

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

A proof of Proposition (2.5) is given in Appendix (5.5). Compared with the convergence rates of the individual curves estimators $\hat{X}^{(d)}_{j,h}$, see (21), the variance of our estimators reduces not only in T but also in N, because equations (13) and (14) can be interpreted as an average over N curves for only a finite number of L components. The intuition behind is, that only components are truncated which 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 2.3, 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 strange results. To learn more about these issues, we check the performance of both approaches in a simulation study in Section 3.2 using different sample sizes.

3 Application to state price densities implied from option prices

In this section we analyses the state price densities (SPD) implied by the stock index option prices. As state dependent contingent claims, they 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 conditions. 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 **?**, **?** and **?**. 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 **?** and **?** 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

$$C(k,\tau) = \exp(-r_\tau \tau) \int_0^\infty (s_\tau - k)^+ q(s_\tau,\tau)\, ds_\tau, \tag{24}$$

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

$$q(s_\tau,\tau) = \exp(r_\tau\tau)\left.\frac{\partial^2 C(k,\tau)}{\partial k^2}\right|_{k=s_\tau}. \tag{25}$$

Let s_0 be the asset price at the moment of pricing and assume it to be fixed. Then $F = \exp(r_\tau\tau)s_0$ is called the forward price. Suppose that the call price is homogeneous of degree one in the strike price. Then it holds that

$$C(k,\tau) = FC(k/F,\tau). \tag{26}$$

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

$$\frac{\partial^2 C(k,\tau)}{\partial k^2} = \frac{1}{F}\frac{\partial^2 C(m,\tau)}{\partial m^2}. \tag{27}$$

Using our previous notations, in our application $X(t)$ will refer to $C(m,\tau)/F$ and its second derivative with respect to the moneyness, $X^{(2,0)}$ will specify the density of future returns $q(s_\tau/s_0,\tau) = s_0 q(s_\tau,\tau)$. In practice, when analyzing a sample of curves, it is preferable to work with densities of future assets returns rather than prices because the underlying functions become location invariant.

3.1 Implementation

1. 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

$$\overline{M}_{ij}^{(\nu)} = \hat{M}_{ij}^{(\nu)} - \frac{1}{T}\sum_{k=1}^T \left(\bar{X}^{(\nu)}(t_k)\hat{X}_{i,b}^{(\nu)}(t_k) + \bar{X}^{(\nu)}(t_k)\hat{X}_{j,b}^{(\nu)}(t_k) - \bar{X}^{(\nu)}(t_k)^2\right). \tag{28}$$

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 2.4, the $\frac{1}{T}\sum_{k=1}^T \bar{X}^{(\nu)}(t_k)^2$ term can be improved accordingly to Lemma 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.

To get the best rates of convergence for $\hat{M}^{(d)}$ according to Remark 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 here $\rho = 1 > 0$ 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 2.4 the dominating error depends mainly on the choice of h. When estimating state price densities $t_{ik} = (\tau_{ik}, m_{ik})$ and we determine the bandwidth with

$b_{i\tau} = \left((\max(\tau_i) - \min(\tau_i))|\tau_i|\right)^{\alpha}$ for the maturity direction where $|\tau_i|$ is the cardinality of the set $\tau_i = \{\tau_{i1}, \ldots \tau_{iT_i}\}$. The bandwidth for the moneyness direction is chosen likewise. In the estimation of $\hat{M}^{(d)}$ we set $\alpha = -1/10$ and $\alpha = -1/3$ for $\hat{M}^{(0)}$.

2. Bandwidth selection h — The choice of bandwidths h is a crucial parameter for the quality of the estimators. To obtain a rough approximation of the optimal bandwidths, 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 2.4.

The theoretical optimal univariate asymptotic bandwidth for h in direction $j = 1,2$ for the r-th basis is given by

$$h_{jr,opt} = C_{d,p,j}(K) \left[T^{-1} \frac{\int_0^1 \sum_{i=1}^N (p_{ir}^{(\nu)})^2 \sigma_{\varepsilon,i}^2(t_j) f_i(t_j)^{-1} dt_j}{\int_0^1 \left\{ \sum_{i=1}^N p_{ir}^{(\nu)} X_i^{(p+1)}(t_j) \right\}^2 dt_j} \right]^{1/(2p+3)}, \tag{29}$$

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

Like in the conventional local polynomial smoothing case $C_{d_j,p}(K)$ 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 **?**.

In practice, we can not use equation (29) 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 h is given by

$$h_{jr,rot}^{(\nu)} = \left(T^{-1} \frac{C_{d,p}^{2p+3} \hat{\sigma}_\varepsilon^2}{f_j \int_0^1 \left\{ \sum_{i=1}^N \hat{p}_{ir}^{(\nu)} \tilde{X}_i^{(p+1)}(t_j) \right\}^2 dt_j} \right)^{1/(g+2p+2)}. \tag{30}$$

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

$$\begin{aligned} a_i^* &= \arg\min_{a_i} \left(X_i(m,\tau) - a_{i0} + \sum_{l=1}^{5} a_{il} m^l + \sum_{l=6}^{9} a_{il} \tau^{(l-5)} \right) \\ \tilde{X}_i^{(4)}(m) &= 24 a_{i4}^* + 120 a_{i5}^* m \\ \tilde{X}_i^{(4)}(\tau) &= 24 a_{i9}^*. \end{aligned} \tag{31}$$

- To estimate the variance for each curve we use the kernel approach given in (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 (30) the average over all $\hat{\sigma}_{i,\epsilon}$ is used.

For technical reasons, we use the product of Gaussian kernel functions to construct local polynomial estimators. The reason for is that, as we can verify from Proposition 2.4, the optimal bandwidth h will decrease in N and by using a global bandwidth and a compact kernel the matrix given in equation (40) may become singular when N is large and T is small.

To derive the individual bandwidth for curve $\hat{X}_{i,h_i}$ that is used in the simulation for comparison, $\hat{p}_{ir}^{(v)}$ is replaced in (30) by 1 and by 0 for all $j \neq i$. In our simulation and application we use the mean optimal $h_{i,rot} = L^{-1}\sum_{r=1}^{L} h_{ir,rot}$ for each $\hat{\gamma}_r$ to save computation time. Since we had to demean the sample in (28), finally we add $N^{-1}\sum_{i=1}^{N} \hat{X}^{(d)}_{i,h^{(v)}_{rot}}$ to the resulting truncated decomposition to derive the final estimate.

3. Estimation of the number of components— So far, we treated the case of known number of components. 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 (63) we show that under the conditions from Proposition 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 **?** can be adapted to our case and give several criteria for finding L or L_d by generalizing **?** 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

$$PC^{(v)}(k^*) = \min_{k\in\mathbb{N}, k\leq L_{\max}} \left[\left(\sum_{r=k+1}^{N} \hat{\lambda}_r^{(v)} \right) + k \left(\sum_{r=L^{\max}}^{N} \hat{\lambda}_r^{(v)} \right) \left(\frac{\log(C_{NT}^2)}{C_{NT}^2} \right) \right], \tag{32}$$

for the constant $C_{NT} = \min(\sqrt{N}, \sqrt{T})$ and a prespecified $L^{\max} < \min(N, T)$. Alternatively, **?** propose an information criteria that do not depend on the choice of $L_{\max}$. We consider the above modified criteria

$$IC^{(v)}(k^*) = \min_{k\in\mathbb{N}, k\leq L_{\max}} \left[\log\left(\frac{1}{N} \sum_{r=k+1}^{N} \hat{\lambda}_r^{(v)} \right) + k \left(\frac{\log(C_{NT}^2)}{C_{NT}^2} \right) \right], \tag{33}$$

Here using $v = (0, \ldots, 0)^\top$ will give L while using $v = d$ 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

$$\mathrm{Var}(\hat{\delta}_{r,T}^{(d)} \hat{\gamma}_{r,T}^{(d)}) = \langle \hat{\gamma}_{r,T}^{(d)}, \hat{\gamma}_{r,T}^{(d)} \rangle \hat{\lambda}_r, \tag{34}$$

and order them and use equations (32) or (33) to select the number of components.

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 3.3. Simulated SPDs are modeled as mixtures of L components, $q(m,\tau) = \sum_{l=1}^{L} w_l q^l(m,\tau)$, 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 $\sum_{l=1}^{L} w_{il} = 1$, where $i = 1,\ldots,N$ is the index for the day. Then

$$q_i(m,\tau) = \sum_{l=1}^{L} w_{il} \frac{1}{m\sqrt{2\pi\sigma_l^2\tau}} \exp\left[-\frac{1}{2}\left\{\frac{\log(m) - \left(\mu_l - \frac{1}{2}\sigma_l^2\right)\tau}{\sigma_l\sqrt{\tau}}\right\}^2\right] \tag{35}$$

Following **?** the prices of call options for these SPDs are

$$C_i(m,\tau) = \exp(-r_{i\tau}\tau) \sum_{l=1}^{L} w_{il} \left\{\exp(\mu_l\tau)\Phi(y_1) - m\Phi(y_2)\right\} \tag{36}$$

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 simulated components q^l and their orthonormal counterparts (principal components) are shown in Figure 1. 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 (m_{ik},τ_{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 (35) are used to verify the performance of $\hat{X}_{FPCA_1}$, $\hat{X}_{FPCA_2}$ as well as for the individually estimated curves $\hat{X}_{Indiv.}$ at common 256 random points in terms of the mean squared error (MSE), which averages the errors for each curve over number of curves time observations per curve. To derive the optimal bandwidth in each case we stick to the rule-of-thumb approach presented in Section 3.1. 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,

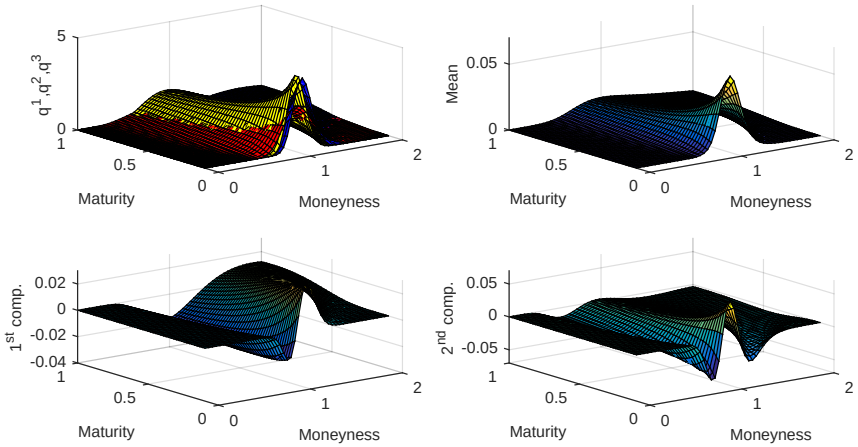

Figure 1. Simulated mixture components - log-normal densities q^1 - red; q^2 - blue; q^3 - yellow; Mean curve and orthonormal components

	T	50				250			
N	$\hat{X}_\bullet$	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 1. Results of the simulation with framework described in Section 3.2 with different preconditions on T and N. $FPCA_1$ and $FPCA_2$ are superior in sense of MSE over the 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 gets better with raising N and T. These results support our asymptotic results given by Proposition 2.2 and 2.5.

variance and the inter quartile distance based at the MSE of the repetitions are given in Table 1.

In all three cases, FPCA applied to undersmoothed curves yields better derivative estimates of the call functions than the local polynomial smoothing applied to individual curves. Both the mean and the median of the MSE are smaller which is a result of the additional average over N as given by Proposition 2.5. However, the $FPCA_1$ method performs decisively better for small T than the other two alternatives both in terms of mean and standard deviation of the mean squared error. In addition $FPCA_1$ benefits more from raising N then $FPCA_2$. 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 stabilized the estimates by averaging over N. To get the same effect using $FPCA_2$ a much bigger T is needed. A possible explanation for this behavior is given by Propo-

sition 2.2, because the rates of convergence to estimate elements of the dual matrix solely rely on T.

3.3 Real Data Example

1. 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. 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 (26), and standardize both the strike and the call prices within one day by the forward stock index value to ensure that the observations are in the same range. Following the methodological framework, we apply the FPCA to the rescaled call prices 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 for the desired maturities.

r, $L_{\max}$	1	2	3	4	5	6	7	8	9	10
$\lambda_r \times 10^6$	133.29	18.90	2.69	1.62	0.49	0.34	0.26	0.09	0.08	0.05
$\lambda_r^{(d)} \times 10^2$	14.74	3.76	2.83	2.14	1.12	0.80	0.20	0.19	0.515	0.12
λ_r/λ_{r+1}	7.05	7.01	1.66	3.28	1.44	1.31	2.83	1.18	1.70	1.35
$\lambda_r^{(d)}/\lambda_{r+1}^{(d)}$	3.92	1.33	1.32	1.92	1.40	4.01	1.04	1.23	1.27	1.37
$k^*(PC^{(0)})$	1	2	3	4	5	6	**7**	**8**	**9**	9
$k^*(PC^{(d)})$	1	2	3	4	5	**6**	6	**8**	**9**	**10**

Table 2. Estimated eigenvalues and eigenvalue ratios. Number of factors by $PC^{(v)}$ criteria.

2. Estimation results — The first eigenvalue of the dual covariance matrix $\hat{M}^0$ for the call option surfaces has a dominantly strong explanatory power and the order of

magnitude of the following eigenvalues decreases by a factor of ten with every few additional components. Following **?**, we also construct the eigenvalue ratio of two consecutive eigenvalues in descending order. The first two terms in the sequence are relatively high and there are a few other increasing terms, e.g., the fourth, seventh and ninth, before the sequence decreases towards one. The choice of the maximum number of factors by the $CP^{(0)}$ criteria suggests at least seven components. This can be seen by looking at the values of k^* marked in bold in Table 2, which are independent of the constrain $k \leq L_{max}$ in the objective function (32). $IC^{(0)}$ criterion gives as well seven factors.

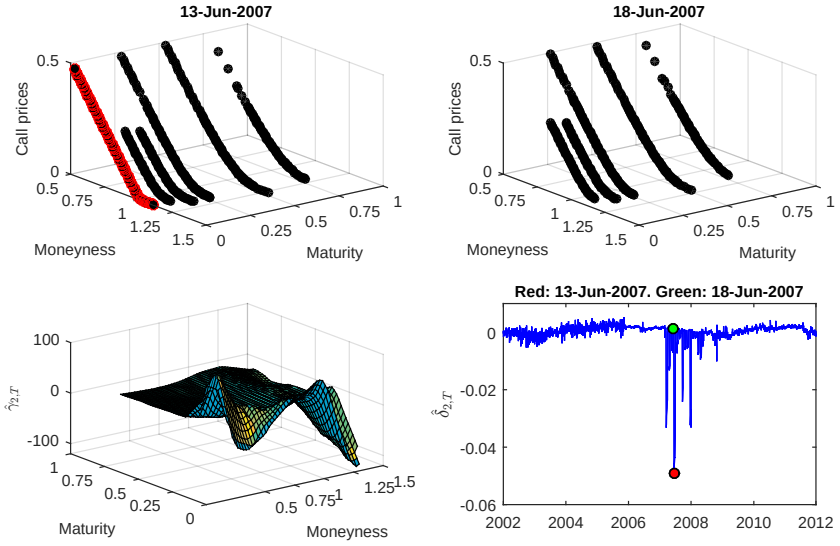

Figure 2. The effect of the expiration date on $\hat{\delta}_{2,T}$

A closer look at the dynamics of the loadings, shows an unusual behavior of some of them - $\hat{\delta}_{2,T}$, $\hat{\delta}_{4,T}$, $\hat{\delta}_{5,T}$ and $\hat{\delta}_{6,T}$ - 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 identified these jumps with the Mondays following an expiration date - which occurs on a single Friday in every month, see Figure 2 that links the dynamics of $\hat{\delta}_{2,T}$ to the expiration days. After the sudden decrease, the loadings increase day-by day and approach a 'normal' level after about two weeks.

During this period, there are not many 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 possibility that

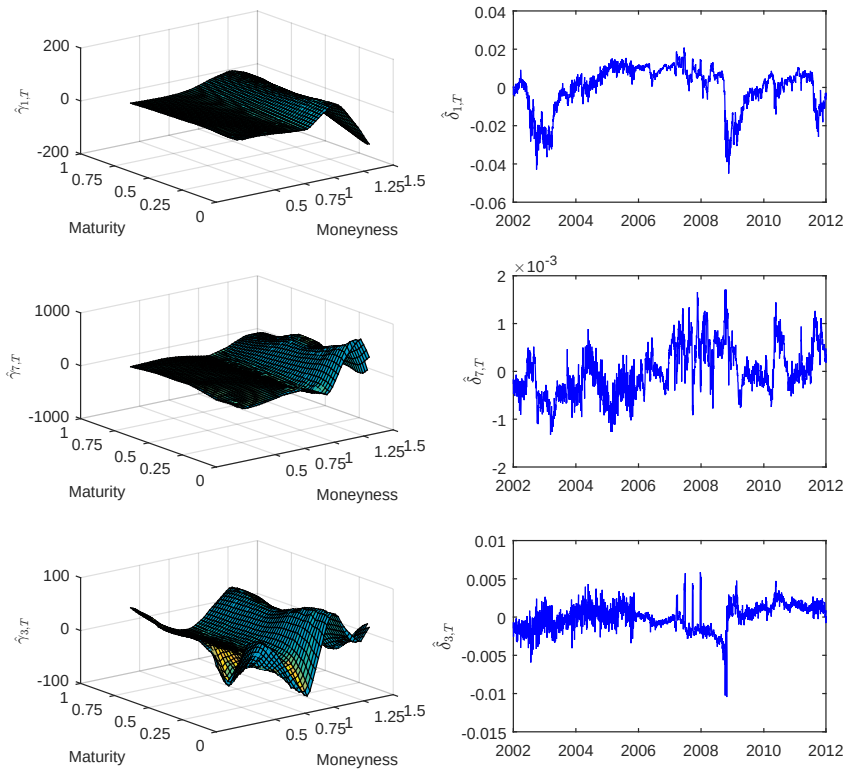

Figure 3. Estimated non-orthogonal components and their loadings obtained by the decomposition of $\hat{M}^{(0)}$, in decreasing order of explained variance from up to bottom

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 $\hat{\gamma}_{2,T}$, displayed in Figure 2, suggests that it is related to the short end of the SPD term structure effect. The other components $\hat{\gamma}_{4,T}$, $\hat{\gamma}_{5,T}$, $\hat{\gamma}_{6,T}$ and $\hat{\gamma}_{8,T}$, whose loadings have similar behavior, are similar in shape to the four components we have discussed so far, i.e., $\hat{\gamma}_{1,T}$, $\hat{\gamma}_{2,T}$, $\hat{\gamma}_{3,T}$ and $\hat{\gamma}_{7,T}$. It is yet not totally clear but it is well possible 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.

The remaining three estimated components are displayed in Figure 3, in order of their explained variance, see equation (34). These three components describe three types of asymmetry present in the dynamics of the SPDs. Their interpretation, is linked to the shape of the empirical mean of the SPD, which has a long tail on the left side of the peak. The first component has positive values around the peak and negative around the tails and is related to the volatility dynamics. An increase in the loadings of this component decreases the volatility of SPD. The second component 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 as well changes in the kurtosis of the density. We interpret it as the negative skewness factor. The third component has a more 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.

For the reduced model $\sum_{r\in\{1,3,7\}} \hat{\delta}_{r,T}\hat{\gamma}_{r,T}$, the functional principal components retrieved from the decomposition of its covariance operator resemble closely the three components from Figure 3. 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, the previous approach seems to provide more accurate estimates that allow for a better interpretation of the results.

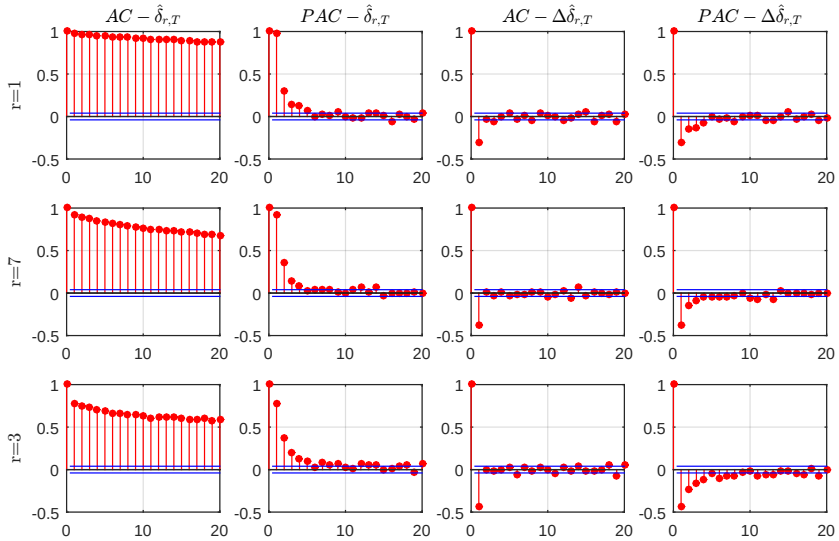

Figure 4. Autocorrelation (AC) and partial autocorrelation (PAC) of the estimated loadings and of their difference

3. Dynamic analysis of the loadings — We look at the dynamics of the SPD as summarized by the reduced model. By construction, the loadings are orthogonal, however their movements suggest strong dependencie. 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.

In the following, we would like to understand how past realizations of the loadings influence present observations. Figure 4 displays the sample autocorrelation and partial autocorrelation functions for the loadings from Figure 3 and their first difference. Autocorrelation functions of the loadings decay very slowly suggesting nonstationarity in the time series. Standard tests for the presence of unit roots are performed for the loadings provide evidence of unit-root in the level of the loadings. Engle-Granger and Johansen tests for pairwise and multiple cointegration relationship between the loadings indicate that the series are cointegrated.

After taking the first difference the high persistence vanishes and there is a single significant spike for the first lag. The partial autocorrelation functions for the first difference spike at the first lag and decreases over the next few lags. The results suggest that an ARIMA(1,1,1) model might be appropriate to model the loadings. Its continuous-time equivalent, the mean-reversion process, has been used to model the changes in the implied volatility surfaces, see **?**. In the following, we focus on the discrete time model and use a vector-autoregressive model (VAR) for the first difference of the loadings.

$$y_i = \Omega y_{i-1} + \varepsilon_i^{VAR},$$

where $y_i = (\Delta\hat{\delta}_{i1}\ \Delta\hat{\delta}_{i3}\ \Delta\hat{\delta}_{i7})^\top$, for $\Delta\hat{\delta}_{ir} = (\hat{\delta}_{ir,T} - \hat{\delta}_{i-1r,T})$, and Ω is a coefficient matrix. The results of the estimation procedure are presented in Table 3a. This model specification also allows us to employ a Granger causality test to assess whether the loadings of one component is actually useful in predicting the future levels of the other SPD components. The results of the test, summarized in Table 3b show that daily changes in the negative skewness factor are not caused by the changes in the volatility factor. The other variables Granger-cause each other. Also notice that we have tested for Granger Causality for multiple lags and the results are quite robust.

All variables react negatively to changes in their own past levels. Previous increases in the negative skewness and kurtosis via $\hat{\delta}_3$ and $\hat{\delta}_7$, lead to higher $\hat{\delta}_1$ (lower implied volatility) in the future. Past values of $\hat{\delta}_3$ and $\hat{\delta}_7$ have negative effects on each other. Increases in the volatility factor $\hat{\delta}_1$ (lower implied volatility) predicts higher $\hat{\delta}_3$ the next trading day. However, these predictions are at the level of the entire sample. The relatively low values for the R^2 in Table 3a can be partly explained by looking at the dynamic contemporaneous correlation structure for the first-difference of loadings. The graph might be an indicative that for forecasting purposes, one is advised to use a time varying VAR model.

Examining closer the dynamic relation for the loadings first difference, represented in Figure 5 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 move-

$\Omega_{i,j}$	$j=1$	$j=2$	$j=3$	a. R^2
$i=1$	-8.84^{**}	58.59^{***}	91.05^{**}	0.1328
$i=2$	3.96^{**}	-32.90^{***}	-35.58^{**}	0.2012
$i=3$	0.22	-2.51^{***}	-28.90^{***}	0.1589

(a) VAR coefficient matrix

$j \nrightarrow i$	$j=1$	$j=2$	$j=3$
$i=1$	3.78^{*}	65.58^{***}	10.63^{***}
$i=2$	4.37^{**}	89.19^{***}	8.77^{***}
$i=3$	1.10	24.67^{***}	125.41^{***}
$j \neq i \nrightarrow i$	82.84^{***}	26.30^{***}	49.35^{***}

(b) Granger test statistics

Table 3. Estimation and test results for the VAR(1) model of the first difference for loadings $\hat{\delta}_{1,T}$, $\hat{\delta}_{3,T}$, and $\hat{\delta}_{7,T}$. The reported coefficients are multiplied by 10^2. Granger causality testing that y_i is not caused by y_j (or any y_j, $j \neq i$) uses a variance-covariance matrix estimated under the assumption of heteroskedastic and correlated covariance of the errors.
$^{***}p<0.01,^{**}p<0.05,^{*}p<0.1$

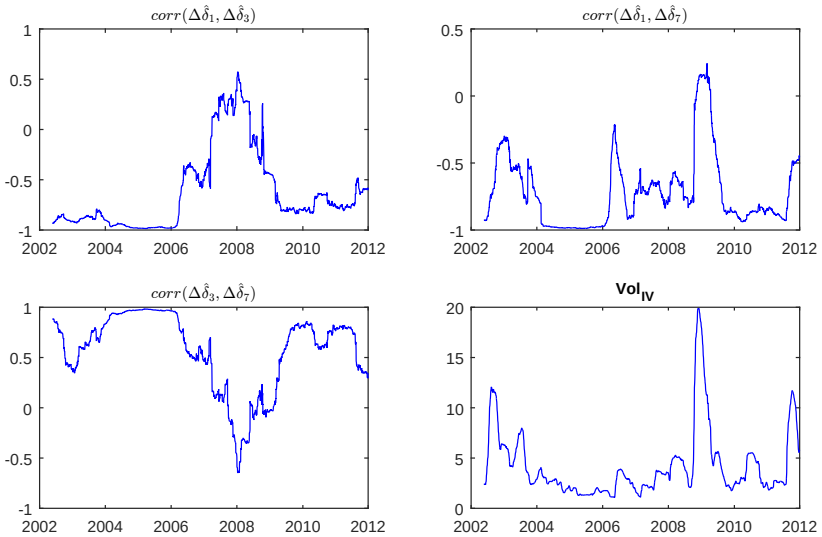

Figure 5. 100-days moving window correlation coefficient for the first-difference of the loadings and volatility of implied volatility

ments of the volatility of implied volatility (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 (**?**, **?**). 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 **?**. 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.

4. Relationship to the dynamics of VDAX index — The VDAX index expresses the implied volatility of the DAX recovered from the prices of call and put options. It is an important indicator on the market, often called "fear" index because it reflects the market expectation for the 30 day ahead volatility of the DAX index under the risk neutral measure, which is then annualized. As an indicator of the second moment of the DAX index under the SPD, we investigate how well the loadings explain its dynamics.

d.v.	$\hat{\delta}_{1,T}$	$\hat{\delta}_{2,T}$	$\hat{\delta}_{3,T}$	$\hat{\delta}_{4,T}$	$\hat{\delta}_{5,T}$	$\hat{\delta}_{6,T}$	$\hat{\delta}_{7,T}$	Adj. R^2
VDAX	-0.9221	0.1362	-0.1458	-0.0146	-0.0411	-0.0109	0.0965	0.975
VDAX	-0.9374	-	-0.1787	-	-	-	0.0935	0.972

Table 4. Contemporaneous regressions of the VDAX index on the loadings of $\hat{\gamma}_{r,T}$. All coefficients are significant at 99%. confidence level

Statistical tests for stationarity suggest a unit root process. Cointegration tests indicate that the VDAX index and the loading series are cointegrated with high probability. This means that there is a contemporaneous relationship between them and we can express VDAX in terms of the loadings. The results of the linear regressions of the VDAX on the loadings of $\hat{\gamma}_{5,T}$ components are reported in Table 4, for three and seven component loadings. The most important factor for explaining the dynamic of VDAX is $\hat{\gamma}_{1,T}$. Both changes in the skewness and kurtosis improve the fit but their impact is decidedly much smaller. Variances increases are associated with to negative movements of these components, which results in flatter densities. In the first regression, the coefficient of $\hat{\delta}_{2,T}$ seems to be relatively important but if we compare the smaller nested model, we see that the impact of the other four loadings improve the overall fitness very little. Also, notice that $\hat{\gamma}_{3,T}$ is the second most important components for explaining VDAX. This shows that, even the amount of variance explained by it is very small compared to the first two ones, its importance for characterizing the

moments of SPD is quite significant. Part of the reason is that VDAX is very sensitive to the changes in the tails of the SPD.

These results help improve our understanding of the volatility index in terms of the shape components of the SPD surface. Furthermore, if we can estimate the loadings of the factors on a particular day, we can calculate the implied VDAX quite well. We have investigated if the loadings have a better predictive power for the VDAX than a univariate random walk. Results, not reported here, show that even they predict VDAX fairly well, they are not superior to a random walk for VDAX. However, the forecasting model where the predictors are past loadings is still useful to model expected short-term changes in the SPD surfaces.

5. Forecasting the DAX index and its realized volatility — Previous studies show that SPD moments improve significantly the accuracy of returns and volatility forecasts for future realizations. Most of these studies infer the option implied SPD moments using the model-free methodology of **?**. In this section, we investigate if the option implied SPD estimated components contain information about the future DAX return and volatility, which are realizations under the real world physical density. We use the Oxford Man daily realized volatility (RVol) as a proxy for the DAX volatility under the physical measure. This is calculated from the realized high-frequency DAX index values, using the quadratic variation method. We select the realized volatility estimates that are using 5 minutes sampling frequency. **?** show that this measure have good performance relative to other candidates. RVol is different from implied volatility index because it describes the real world volatility and not a theoretical value, like the option implied risk neutral volatility. Statistical tests show that the equity index and RVol have stochastic trend in level with high probability. Their first difference is stationary and does not reveal any autocorrelation in the series and errors. This means that the current level of RV is a good predictor for it's future realization. We study if changes in the loadings have additionally predictive power for the future realizations. The forecasting equation is

$$\mathsf{E}[Z_{i+\Delta}|\Psi_i] = c + \sum_{r\in\{1,3,7\}} \alpha_r(\hat{\delta}_{ir,T} - \hat{\delta}_{i-\Delta r,T}),$$

where $Z_{i+\Delta}$ refers either to the future log-returns $\log(S_{i+\Delta}/S_i)$ or to the future realized volatility $RVol_{i+\Delta} - RVol_i$, $i+\Delta$ is the forecasting horizon date Δ days from date i and Ψ_i is the conditioning information set. In order to control for the contribution that the skewness and kurtosis factors have on the implied volatility, we reestimate the previous equation also using the difference in the lagged values of the VDAX instead of the loadings of the first component. Entries in Tables 5 and 6 report the regression results.

For log-returns, in the first regression, the third component is the only one significant in predicting returns for maturities of one day and one week. This factor remains important for increasing horizons but the impact that it has on returns reverses signs. For small horizons, an increase in the loadings of $\hat{\gamma}_{3,T}$ has a positive impact on the returns and in most of the cases a negative effect on long-term returns. The negative skewness and volatility reverse sign after one-month horizon. Positive changes in the volatility factor have negative effects on the returns under one month, i.e. short-term expected decreases in volatility is accompanied by negative returns, and positive ef-

τ	1D	1W	2W	3W	1M	2M	3M
$\Delta\hat{\delta}_{1,T}$	-2.83	1.23	-11.34***	-6.32***	-2.39	14.23***	27.11***
$\Delta\hat{\delta}_{3,T}$	4.45*	2.95*	3.79**	-0.31	-4.89***	-9.48***	7.32***
$\Delta\hat{\delta}_{7,T}$	0.43	1.69	-8.88***	-7.82***	-5.20***	5.23***	0.59
Intc.	3.08**	9.03***	14.18***	14.63***	14.97***	23.05***	23.19***
a. R^2 %	0.81	0.99	4.46	2.92	2.75	10.09	12.53
ΔIV	3.00**	-4.58***	6.89***	4.29***	1.33	-10.81***	-28.88***
$\Delta\hat{\delta}_{3,T}$	6.00***	2.52	5.46***	1.31	-4.51***	-12.46***	0.45
$\Delta\hat{\delta}_{7,T}$	0.88	2.17	-6.25***	-6.88***	-4.73***	4.66***	3.80**
Intc.	3.14**	9.07***	13.94***	14.53***	14.96***	23.33***	25.57***
a. R^2 %	0.87	1.29	3.38	2.63	2.70	9.04	15.49
n.o.	2556	2550	2543	2536	2529	2501	2473

Table 5. Forecasting log-return $\log(S_{i+\Delta}/S_i)$ by the changes in the loadings $\hat{\delta}_{ir,T} - \hat{\delta}_{i-\Delta r,T}$, $r = \{1,3,7\}$ (upper table) and the changes in the implied volatility index $IV_i - IV_{i-\Delta}$ and loadings $\hat{\delta}_{ir,T} - \hat{\delta}_{i-\Delta r,T}$, $r = \{3,7\}$ at horizons (lower table)Δ. All variables are standardized. The reported coefficients are multiplied by 10^2. Regressions with Newey-West standard errors.
$^{***}\,p < 0.01, ^{**}\,p < 0.05, ^{*}\,p < 0.1$

fect of the longer term returns, i.e. long-term expected decreases in volatility is accompanied by positive returns. These findings are in line with the reversal pattern for the implied volatility of the S&P 100 options as a predictor of the future market return reported in **?**. The behavior pattern is also consistent with **?** who find that very low levels of volatility appear to be followed by significantly positive average returns for horizons larger than one month. An increase in the negative skewness, has a negative effect on the returns for horizons of up to one month and a positive effect for longer prediction horizons. The current works in the applied finance literature, see for instance **?**, **?**, **?**, report that the more negative risk neutral skewness indicates a higher probability of a negative price movement. Our findings, point to a reversal pattern on SPD skewness, which is strongly linked to the volatility behavior. When increases in the implied volatility has a positive impact on the future returns an increase in the negative skewness has a negative effect on returns. And the other way around, when volatility has a negative impact on future returns, an increase in the positive skewness has positive effects on the future returns. The quality of prediction usually increases with the prediction horizon. This is in line with the findings of **?** that in-sample predictability of aggregate returns by downside risk and skewness measures increases over the term structure of equity returns. In conformity with our results, their empirical investigation highlights the positive and significant link between the downside variance risk and the equity premium, as well as a positive and significant relation between the skewness risk premium and the equity premium.

We further asses how changes in the loadings can predict future changes in the realized volatility. Table 6 indicates a positive effect of the changes in the three main SPD components to the changes in the realized volatility, i.e. decreases in the expected risk neutral volatility, increases in skewness factors, all predict higher realized volatility in the future. Part of the results are due to the ability of implied volatility to forecast realized volatility, as reported in several studies. If we use the changes in the implied volatility index instead of the loadings of the first component, the coefficient of the skewness sign factor is no longer significant for any forecasting horizon. The direction

τ	1D	1W	2W	3W	1M	2M	3M
$\Delta\hat{\delta}_{1,T}$	-0.38	13.21***	9.11***	15.34***	20.60***	21.44***	15.90***
$\Delta\hat{\delta}_{3,T}$	2.53	3.47***	0.41	2.68**	4.59***	3.82***	3.09**
$\Delta\hat{\delta}_{7,T}$	-0.77	5.29***	1.04	4.49***	6.46***	6.48***	2.45*
Intc.	6.66**	-2.54**	-3.51***	-4.52***	-5.61***	-8.07***	-9.08***
a. R^2 %	0.06	3.58	2.04	4.88	9.18	10.16	5.76
ΔIV	-5.05***	-19.40***	-13.11***	-16.36***	-24.51***	-22.87***	-18.28***
$\Delta\hat{\delta}_{3,T}$	1.74	-0.17	-1.57	0.35	0.51	-0.68	-0.59
$\Delta\hat{\delta}_{7,T}$	0.79	3.17**	0.13	2.64*	5.66***	7.13***	3.78*
Intc.	6.30**	-3.13**	-3.89***	-4.75***	-6.11***	-7.82***	-9.38***
a. R^2 %	0.6	9.33	4.37	6.10	13.62	11.51	7.54
n.o.	2556	2550	2543	2536	2529	2501	2473

Table 6. Forecasting changes in the realized volatility $RVol_{i+\Delta} - RVol_i$ by (a) the changes in the loadings $\hat{\delta}_{ir,T} - \hat{\delta}_{i-\Delta r,T}$, $r = \{1,3,7\}$ and (b) the changes in the implied volatility index $IV_i - IV_{i-\Delta}$ and loadings $\hat{\delta}_{ir,T} - \hat{\delta}_{i-\Delta r,T}$, $r = \{3,7\}$ at horizons Δ. All variables are standardized. The reported coefficients are multiplied by 10^2. Regressions with Newey-West standard errors.
$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

of impact of implied volatility and skewness changes on the realized volatility does not change. An increase in the implied volatility index commands a decrease in the realized volatility at all horizons, while a decrease in the negative skewness as positive effects on the future realized volatility. **?** also find that the realized future volatility increases in the current negative skewness. **?** show that the skewness of the risk neutral density can explain the bias of option implied volatility to forecast its physical counterpart. **?** provide evidence that skew and variance premia are manifestations of the same underlying risk factor. **?** show that the risk-neutral skewness has the stronger anticipatory power for future stock return volatility for a long period due to the presence of uninformed noise traders in the stock and option markets.

A possible interpretation for the observed behavior is provided in the following. In general, a high negative skew reflects the market expectation that the future stock index will be increase relative to the previous period. This is true, in particularly for low levels of volatility on the long run. But there are deviations from this case on the short term. Short-term decreases in the negative skewness is an indicative for the decline in the risk aversion of some of the investors, who consider that the stock is overpriced. When they face short-selling constrains, the price of the asset will continue to rise and its volatility increases. Once that market inefficiencies are exploited we return to the 'steady state behavior'. This process is yet not instantaneous and will occur through concomitant increased demand for OTM puts for hedging which needs to be met by option supply from the market-makers side. If the later are credit-constrained, the equilibrium price of options decreases. In addition, there is an asymmetry in the type of risks insured, e.g. high risks will not be insures and deep OTM puts become less expensive.

6. 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. **?**, **?**. 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. **?** for a study of S&P 500 index implied volatility. **?** 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.

4 Conclusions

We present two methods for estimating the derivatives of high-dimensional curves using FPCA techniques. The first approach FPCA is applied to the covariance operator of the curve derivative. The second approach considers the decomposition of the covariance operator for the original curves, whereas the derivatives are applied to their functional principal components. Thus, the second approach will explain the dynamics of derivatives in terms of orthogonal loadings but the components are no longer orthonormal. 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. Two factors describe the variation of the term structure of the SPD. The empirical analysis provides some insights into the economics behind the option pricing. Further findings suggest the effectiveness of using option-implied SPD components to forecast future returns or future realized volatility of the underlying asset.

5 Appendix

5.1 Assumptions summary

Assumption 5.1 *The curves Y_i $i = 1,\dots,N$ are observed at a random grid $t_{i1},\dots,t_{iT_i}$, $t_{ij} \in [0,1]^g$ having a common density f with support $supp(f)$ and the integrand $u \in supp(f)$ and $inf_u f(u) > 0$.*

Assumption 5.2 *$E(\varepsilon_{ik}) = 0, var(\varepsilon_{ik}) = \sigma_{\varepsilon,i}$ and ε_{ik} are independent of X_i, and $\sigma_{\varepsilon,i}^{-l}\mathsf{E}\left[\varepsilon_{ik}^l\right] < \infty$, $l = 3,4\ \forall i,k$.*

Assumption 5.3 *Let $K_{\boldsymbol{B}}(u) = \frac{1}{|\boldsymbol{B}|}K(u\boldsymbol{B})$. K is a product kernel based on symmetric univariate kernels. Further the kernel K is bounded and compactly supported such that for $u \in \mathbb{R}^g$ $\int uu^T K(u)du = \mu(K)I$ with $\mu(K)$ is a constant and I is the $g \times g$ identity matrix.*

Assumption 5.4 *$\rho - \sum_{l=1}^g d_l$ so that $\rho - \sum_{l=1}^g d_l$ and $p - \sum_{l=1}^g d_l$ are odd.*

Assumption 5.5 *$|\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2| = \mathcal{O}_P(T^{-1/2})$*

Assumption 5.6 *We require that for the decomposition it holds that*

$$\sup_{r\in\mathbb{N}}\sup_{t\in[0,1]^g}|\varphi_r^{(d)}(t)| < \infty, \sup_{r\in\mathbb{N}}\sup_{t\in[0,1]^g}|\gamma_r^{(d)}(t)| < \infty \tag{37}$$

$$\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\mathsf{E}\left[\left(\delta_{ri}^{(\nu)}\right)^2\left(\delta_{si}^{(\nu)}\right)^2\right] < \infty, \sum_{q=1}^{\infty}\sum_{s=1}^{\infty}\mathsf{E}\left[\left(\delta_{ri}^{(\nu)}\right)^2\delta_{si}^{(\nu)}\delta_{qi}^{(\nu)}\right] < \infty,\ \nu = (0,d) \tag{38}$$

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

Assumption 5.7 *We require that the eigenvalues are distinguishable such that for any T and N and fixed $r \in 1,\dots,L$ there exists $0 < C_{1,r} < \infty$, $0 < C_{2,r} \le C_{3,r} < \infty$ such that*

$$\begin{aligned} NC_{2,r} \le l_r^{(\nu)} \le NC_{3,r} \\ \min_{s=1,..,N;s\neq r}|l_r^{(\nu)} - l_s^{(\nu)}| \ge NC_{1,r}. \end{aligned} \tag{39}$$

5.2 Proof of Lemma 2.1

1. Univariate case g=1—

As noted by **?** equation (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. In particular let e_l be a vector of length ρ with 1 at the $l+1$ position and zero else, then $W_d^T(t)$ to evaluate the function at point u is defined as $(Tb^{d+1})^{-1}e_d^T S_T(u)^{-1}(1,t,\dots,t^\rho)^T K(t)$. $S_T(u)$ 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

$$S_T(u) = \begin{pmatrix} S_{T,0}(u) & S_{T,1}(u) & \dots & S_{T,\rho}(u) \\ S_{T,1}(u) & S_{T,2}(u) & \dots & S_{T,\rho+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{T,\rho}(u) & S_{T,\rho+1}(u) & \dots & S_{T,2\rho}(u) \end{pmatrix}. \tag{40}$$

Accordingly

$$
\begin{aligned}
\mathsf{E}(S_{T,k}(u)) =& (Tb)^{-1} \int_0^1 \sum_{l=1}^{T} K\left(\frac{x-u}{b}\right)\left(\frac{x-u}{b}\right)^k f(x)dx \\
=& b^{-1} \int_u^{1+u} K\left(\frac{x}{b}\right)\left(\frac{x}{b}\right)^k f(x)dx = \int_{ub^{-1}}^{(1+u)b^{-1}} K(t)\, t^k f(tb)dt.
\end{aligned} \tag{41}
$$

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 (40) with

$$
S(u) = \begin{cases} S_c = (S_{j+l,c})_{0\leq j,l\leq \rho} & , u \text{ is a boundary point} \\ S = (S_{j+l})_{0\leq j,l\leq \rho} & , u \text{ is an interior point} \end{cases}. \tag{42}
$$

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

$$
\hat{X}_b^{(d)}(u) = d!\beta_d(u) = \frac{d!}{Tf(u)b^{d+1}} \sum_{l=1}^{T} K_{d,\rho}^{u*}\left(\frac{t_l - u}{b}\right) Y(t_l)\{1 + o_P(1)\} \tag{43}
$$

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 **?** we can further state that with a bandwidth fulfilling $\frac{log(T)}{Tmax(b)^g} \to 0$ we have uniformly in $u \in D$ (for the univalent case $D = [0,1]$ and for the multivariate case $D = [0,1]^g$) that $sup_{u\in D}|S_T(u) - S(u)f(u)| = \mathcal{O}_P\left(\frac{log(T)}{Tmax(b)^g}\right)$ almost surely. We will thus drop the u index concerning the equivalent kernel from now on.

By construction the equivalent kernel fulfills that using the Kronecker-Delta δ_{ij}

$$
\int u^k K_{d,\rho}^*(u)\, du = \delta_{d,k} \;\; 0 \leq d, k \leq \rho \; \forall i = 1, \ldots g
$$

as mentioned by **?** the design of the kernel automatically adepts to the boundary which gives as shown in **?** the same order of convergence for interior and also for boundary points. The estimator can then be rewritten as

$$
\begin{aligned}
&\int d!^2 \sum_{j=1}^{T}\sum_{l=1}^{T} W_d^T\left(\frac{t_j - u}{b}\right) W_d^T\left(\frac{t_l - u}{b}\right) Y(t_l)Y(t_j)du \\
=& \int \frac{d!^2}{T^2 f(u)^2 b^{2d+2}} \sum_{l=1}^{T}\sum_{j=1}^{T} K_{d,\rho}^*\left(\frac{t_j - u}{b}\right) K_{d,\rho}^*\left(\frac{t_l - u}{b}\right) Y(t_l)Y(t_j)\{1 + o_P(1)\}du.
\end{aligned} \tag{44}
$$

For the expectation we get

$$
\begin{aligned}
&\mathsf{E}\left(\theta_{d,\rho}|t_1,\ldots,t_T\right)\\
&=\int_0^1 d!^2\sum_{j=1}^{T}\sum_{l=1}^{T}W_d^T\left(\frac{t_j-u}{b}\right)W_d^T\left(\frac{t_l-u}{b}\right)X(t_l)X(t_j)du\\
&\quad+d!^2(\sigma_\varepsilon^2-\hat{\sigma}_\varepsilon^2)\int_0^1\sum_{j=1}^{T}W_d^T\left(\frac{t_j-u}{b}\right)^2du.\\
&=\int_0^1 X^{(d)}(z)X^{(d)}(z)dz\\
&\quad+2\frac{d!}{(\rho+1)!}\int_0^1\frac{b^{\rho+1}}{b^d}\left(\int_0^1 u^{\rho+1}K_{d,\rho}^*(u)\,du\right)X^{(\rho+1)}(z)X^{(d)}(z)dz\\
&\quad+\frac{d!^2}{(\rho+1)!^2}\int_0^1\frac{b^{2\rho+2}}{b^{2d}}\left(\int_0^1 u^{\rho+1}K_{d,\rho}^*(u)\,du\right)^2X^{(\rho+1)}(z)X^{(\rho+1)}(z)dz\\
&\quad+o_P\left(\frac{1}{T^{1/2}}+\frac{b^{\rho+1}}{b^d}\right)\{1+o_p(1)\}
\end{aligned}
\tag{45}
$$

results where obtained by using a $\rho+1$ order Taylor expansion of $X(x-bu)$. We get $\int_{[0,1]^g}X(u)^2du-\mathsf{E}(\theta_{d,\rho})=\mathcal{O}_p(h^{\rho+1-d})$.

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

$$
\int f(z)^{-2}K_{d,\rho}^*\left(\frac{y-z}{b}\right)K_{d,\rho}^*\left(\frac{x-z}{b}\right)dz=f(c)^{-2}\int K_{d,\rho}^*\left(\frac{y-z}{b}\right)K_{d,\rho}^*\left(\frac{x-z}{b}\right)dz. \tag{46}
$$

We introduce a kernel convolution with

$$
K_{d,\rho}^C(y-x):=\int K_{d,\rho}^*(y-z)\,K_{d,\rho}^*(x-z)\,dz \tag{47}
$$

and thus using $z=ub^{-1}$

$$
K_{d,\rho}^C\left(\frac{y-x}{b}\right)=\int K_{d,\rho}^*\left(\frac{y}{b}-z\right)K_{d,\rho}^*\left(\frac{x}{b}-z\right)dz=\int b^{-1}K_{d,\rho}^*\left(\frac{y-u}{b}\right)K_{d,\rho}^*\left(\frac{x-u}{b}\right)du. \tag{48}
$$

Note that the integral 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 $x=\frac{v+u}{2}b,\ y=\frac{u-v}{2}b$

$$
\int\int_D K_{d,\rho}^C\left(\frac{y-x}{b}\right)dydx=\frac{b}{2}\int_0^2\int_{-1}^1 K_{d,\rho}^C\left(\frac{v+u-u+v}{2}\right)dvdu=b\int K_{d,\rho}^C(v)\,dv. \tag{49}
$$

Note that the variance can be decomposed

$$
\begin{aligned}
(50)\quad & Var\left(\theta_{d,\rho}|t_1,\ldots,t_T\right)\\
(51)\quad =&\frac{d!^4}{T^4(b^{4d+2})f(c)^4}\left(\sum_{l=1}^{T}K_{d,\rho}^{C}(0)^2\,Var(Y(t_l)^2)\right.\\
(52)\quad &+2\sum_{l=1}^{T}\sum_{k\neq l}^{T}Var(K_{d,\rho}^{C}\left(\frac{t_l-t_k}{b}\right)Y(t_l)Y(t_k))\\
(53)\quad &+4\sum_{l=1}^{T}\sum_{k\neq l}^{T}\sum_{k'\neq k}^{T}Cov(K_{d,\rho}^{C}\left(\frac{t_k-t_l}{b}\right)Y(t_k)Y(t_l),K_{d,\rho}^{C}\left(\frac{t_l-t_{k'}}{b}\right)Y(t_l)Y(t_{k'}))\\
(54)\quad &+24\sum_{l=1}^{T}\sum_{k\neq l}^{T}\sum_{k'\neq k}^{T}\sum_{l'\neq k'}^{T}Cov(K_{d,\rho}^{C}\left(\frac{t_l-t_k}{b}\right)Y(t_l)Y(t_k),K_{d,\rho}^{C}\left(\frac{t_{l'}-t_{k'}}{b}\right)Y(t_{l'})Y(t_{k'}))\quad head\\
(55)\quad &\left.+\mathcal{O}_P\left(\frac{1}{T}\right)\right)
\end{aligned}
$$

where (51) expression equals $\frac{d!^4}{T^3(b^{4d+2})f(c)^4}\int K_{d,\rho}^{C}(0)^2\,Var(Y(y)^2)f(y)dy\{1+\mathcal{O}_P(T^{-1})\}$ and is dominated by (52) while (54) vanishes. Further note that

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

Before looking at the covariance term note that with $m\geq 2d$

$$
\begin{aligned}
&\int\int\frac{d!^2}{b^{2d+1}}K_{d,\rho}^{C}\left(x-yb\right)X(x)dxdy\\
=&\int\int\frac{d!^2}{b^{2d}}K_{d,\rho}^{C}(u)\,X(y-ub)dudy\\
=&\frac{d!^2}{b^{2d}}\int\int\int K_{d,\rho}^{*}(m)\,K_{d,\rho}^{*}(z)\,X(y-(m-z)b)dzdmdy\\
=&(-1)^d\int_0^1X^{(2d)}(y)dy+o_P(1)
\end{aligned}
$$

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

We can thus derive for the covariance term that

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

2. multivariate case— The same strategy also works in the multivariate case, the bias is the same as in the multivariate case when replacing the univariate Taylor series using the multi-index notation introduces in section 2.4 and $a=(a_1,...,a_g),\ a_l\in\mathbb{N}^+$ by a multivariate taylor series of degree $k<\rho$ given by

$$X(x-u\circ b)=\sum_{0\leq|a|\leq k}\frac{X^{(a)}(x)}{a!}(u\circ b)^a+o_P\left(u^{k+1}max(b)^{k+1}\right)$$

using **?** extended with results from **?** the multivariate equivalent kernel has the properties that with $v=(v_1,\ldots,v_g),\ v_l\in\mathbb{N}^+$

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

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

$$
\begin{aligned}
&\mathsf{E}\left(\theta_{d,\rho}|t_1,\ldots,t_T\right)\\
=&\int_{[0,1]^g}X^{(d)}(z)X^{(d)}(z)dz\\
&+2\frac{d!}{(\rho+1)!}\int_{[0,1]^g}\frac{max(b)^{\rho+1}}{b^d}\left(\int u^{\tilde{\rho}}K_{d,\rho}^{*}(u)\,du\right)X^{(\tilde{\rho})}(z)X^{(d)}(z)dz\\
&+\mathcal{O}_P\left(\frac{1}{Tb_1\times\ldots\times b_gb^{2d}}\right)+o_P\left(\frac{max(b)^{\rho+1}}{b^d}+\frac{1}{Tb_1\times\ldots\times b_gb^{2d}}\right).
\end{aligned}
\tag{56}
$$

Further note that for the convoluted kernel we get

$$
\begin{aligned}
&K_{d,\rho}^{C}\left((y-x)\circ b^{-1}\right)\\
=&\int(b_1\times\ldots\times b_g)^{-1}K_{d,\rho}^{*}\left((y-u)\circ b^{-1}\right)K_{d,\rho}^{*}\left((x-u)\circ b^{-1}\right)du.
\end{aligned}
\tag{57}
$$

Accordingly, we get for the variance term that

$$
\begin{aligned}
&\frac{2d!^4}{T^4 f(c)^4 (b_1^2 \times \ldots \times b_g^2 b^{4d})} \sum_{l=1}^{T} \sum_{k \neq l}^{T} K_{d,\rho}^{C} \left((t_l - t_k) \circ b^{-1} \right)^2 Var(Y(t_l) Y(t_k)) \\
&= \frac{2d!^4 \int (\sigma^4 + 2\sigma^2 X(x)^2) f(x)^2 dx}{T^2 f(c)^4 b_1 \times \ldots \times b_g b^{4d}} \int \left(K_{d,\rho}^{C}(u) \right)^2 du \{1 + o_P(1)\}
\end{aligned} \tag{58}
$$

and because we assume that if $m \geq 2max(d_i)$ $i = 1, \ldots, g$ we get for the covariance part

$$
\begin{aligned}
&A \sum_{l=1}^{T} \sum_{k \neq l}^{T} \sum_{k' \neq k}^{T} Cov(K_{d,\rho}^{C} \left((t_k - t_l) \circ b^{-1} \right) Y(t_k) Y(t_l), K_{d,\rho}^{C} \left((t_l - t_{k'}) \circ b^{-1} \right) Y(t_l) Y(t_{k'})) \\
&= A \sum_{l=1}^{T} \sum_{k \neq l}^{T} \sum_{k' \neq k}^{T} K_{d,\rho}^{C} \left((t_k - t_l) \circ b^{-1} \right) K_{d,\rho}^{C} \left((t_l - t_{k'}) \circ b^{-1} \right) X(t_k) \sigma^2 X(t_{k'}) \\
&= \frac{4\sigma^2}{T f(c)} \int X^{(2d)}(y) X^{(2d)}(y) dy + \mathcal{O}_P \left(\frac{1}{T^2 (b^{4d} b_1 \times \ldots \times b_g)} \right)
\end{aligned} \tag{59}
$$

where $A := \frac{4d!^4}{T^4 (b^{4d} b_1^2 \times \ldots \times b_g^2) f(c)^4}$.

5.3 Proof of Proposition 2.2

1. Asymptotic results for of $\tilde{M}^{(0)}$ — 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 (12) with a sum.

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

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., $\int_{[0,1]^g} X_i(t) X_j(t) dt = \mathsf{E}\left[X_i X_j\right]$. 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 smaller or equal than $T^{-1/2}$. Besides the error for the off diagonal elements is smaller then for the diagonal because one additional point is used. Using Lemma 2.1 we can further state that under the proposed conditions $\rho \geq \frac{g}{2} - 1$ and $b^* = T_i^{-\alpha}\ \forall i = 1,\ldots,N$ with $\frac{1}{2\rho+2} \leq \alpha \leq \frac{1}{g}$ the error from interpolation and smoothing is also of order $T_i^{-1/2} \leq \min_i(T_i)^{-1/2} = T^{-1/2}$.

2. Asymptotic results for $\hat{M}^{(d)}$ — Asymptotic of $\hat{M}^{(d)}$ are similar to those of $\hat{M}^{(0)}$, but to get $T_i^{-1/2}$ using Lemma 2.1 additionally need that each Y_i is at least $2|d|$-times differentiable and $b = T_i^{-\alpha}$ for each direction and $\frac{1}{2(\rho+1-\sum_{l=1}^{g} d_l)} \leq \alpha \leq \frac{1}{g+4\sum_{l=1}^{g} d_l}$ to hold.

5.4 Proof of Proposition 2.4

Under the assumptions of Proposition 2.4 together with the requirements of Lemma 2.2 for $\nu = (0,d)$ by the calculations of Appendix 5.2

$$||\hat{M}^{(\nu)} - M^{(\nu)}|| \leq \operatorname{tr}\left\{\left(\hat{M}^{(\nu)} - M^{(\nu)}\right)^\top \left(\hat{M}^{(\nu)} - M^{(\nu)}\right)\right\}^{1/2} = \mathcal{O}_p\left(NT^{-1/2}\right). \tag{60}$$

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 **?** leads to

$$\mathsf{E}\left[\left(p_r^{(\nu)}\right)^\top (\hat{M}^{(\nu)} - M^{(\nu)}) p_r^{(\nu)}\right]^2 = \mathcal{O}_p\left(\frac{N}{T}\right) \tag{61}$$

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

$$\begin{aligned} &\left|\frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N} p_{ir}^{(\nu)} X_i^{(d)}(t) - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\sum_{i=1}^{N}\hat{p}_{ir}^{(\nu)}\hat{X}_{i,h}^{(d)}(t)\right| \\ &\leq \left|\frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N} p_{ir}^{(\nu)}\left[X_i^{(d)}(t) - \hat{X}_{i,h}^{(d)}(t)\right]\right| + \left|\sum_{i=1}^{N}\left(\frac{1}{\sqrt{l_r^{(\nu)}}}p_{ir}^{(\nu)} - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\hat{p}_{ir}^{(\nu)}\right)\hat{X}_{i,h}^{(d)}(t)\right|. \end{aligned} \tag{62}$$

The first term is discussed in equation (2.4). Therefore we take a look at the second term here. As a consequence of Assumption (5.7), Lemma A (a) from **?** together with equation (61) gives

$$l_r^{(\nu)} - \hat{l}_r^{(\nu)} = (p_r^{(\nu)})^T(\hat{M}^{(\nu)} - M^{(\nu)})p_r^{(\nu)}) + \mathcal{O}_p(NT^{-1}) = \mathcal{O}_p(N^{1/2}T^{-1/2} + NT^{-1}) \tag{63}$$

and

$$\frac{1}{\sqrt{\hat{l}_r^{(\nu)}}} - \frac{1}{\sqrt{l_r^{(\nu)}}} = \frac{l_r^{(\nu)} - \hat{l}_r^{(\nu)}}{\sqrt{\hat{l}_r^{(\nu)}}\sqrt{l_r^{(\nu)}}(\sqrt{\hat{l}_r^{(\nu)}} + \sqrt{l_r^{(\nu)}})} = \mathcal{O}_p\left(T^{-1/2}N^{-1} + T^{-1}N^{-1/2}\right). \tag{64}$$

Using Lemma A (b) from **?** we further get

$$|\hat{p}_{ir}^{(\nu)} - p_{ir}^{(\nu)}| = \mathcal{O}_p\left((NT)^{-1/2}\right) \text{ and } ||\hat{p}_r^{(\nu)} - p_r^{(\nu)}|| = \mathcal{O}_p\left(T^{-1/2}\right). \tag{65}$$

Putting all results together for the second term gives

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

Using Cauchy-Schwarz and(64) we see that first term is of order $(NT)^{-1/2}$. For the second term remember that $l_r^{(v)}$ is of order N together with (65) this also leads to order $(NT)^{-1/2}$. Inserting the right hand side, equation (62) becomes

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

5.5 Proof of Proposition 2.5

Note that

$$
\sqrt{l_r^{(v)}}-\sqrt{\hat{l}_r^{(v)}}=(l_r^{(v)}-\hat{l}_r^{(v)})(\sqrt{l_r^{(v)}}+\sqrt{\hat{l}_r^{(v)}})^{-1}=\mathcal{O}_p(T^{-1/2}+N^{1/2}T^{-1}) \tag{66}
$$

together with (65)

$$
\begin{aligned}
&\hat{\delta}_{ir}-\hat{\delta}_{ir,T}=\sqrt{l_r^{(v)}}p_{ir}^{(v)}-\sqrt{\hat{l}_r^{(v)}}\hat{p}_{ir}^{(v)} \\
&=\left(\sqrt{l_r^{(v)}}-\sqrt{\hat{l}_r^{(v)}}\right)p_{ir}^{(v)}-\sqrt{\hat{l}_r^{(v)}}\left(\hat{p}_{ir}^{(v)}-p_{ir}^{(v)}\right)=\mathcal{O}_p(T^{-1/2}+N^{1/2}T^{-1})
\end{aligned} \tag{67}
$$

using Proposition 2.4 it follows that

$$
\begin{aligned}
&|Y_i(t)-\hat{Y}_i(t)|=|\sum_{r=1}^{K}\hat{\delta}_{ir}\hat{\gamma}_r^{(v)}(t)-\sum_{r=1}^{K}\hat{\delta}_{ir,T}\hat{\gamma}_{r,T}^{(v)}(t)| \\
&=|\sum_{r=1}^{K}(\hat{\delta}_{ir}-\hat{\delta}_{ir,T})\hat{\gamma}_r+\hat{\delta}_{ir,T}(\hat{\gamma}_r-\hat{\gamma}_{r,T})| \\
&=\mathcal{O}_p\left(T^{-1/2}+N^{1/2}T^{-1}+max(h)^{p+1}h^{-d}+(NTh_1\times\ldots\times h_gh^{2d})^{-1/2}\right).
\end{aligned} \tag{68}
$$