Functional Principal Component Analysis for Derivatives of Multivariate Curves ⋆

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

[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] Sim Kee Boon Institute for Financial Economics, Singapore Management University, 81 Victoria Street, Singapore 188065
[3] Institute for Financial Economics and Statistics, Department of Economics, Rheinische Friedrich-Wilhelms-Universität Bonn, Adenauerallee 24-26, 53113 Bonn

Abstract

We present two methods based on functional principal component analysis (FPCA) for the estimation of smooth derivatives of a sample of random functions, which are observed in more than one-dimensional domain. We apply eigenvalue decomposition to a) the dual covariance matrix of the derivatives, and b) the dual covariance matrix of the observed curves. To handle noisy data from discrete observations, we rely on local polynomial regressions. If curves are contained in a finite-dimensional function space, the second method performs better asymptotically. We apply our 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 tail factors. We also find evidence for term structure variation.

Keywords: functional principal component, dual method, derivatives, multivariate functions, state price densities

JEL codes: C13, C14, G13

⋆ 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 random functions. Usually, discrete and noisy versions of them are observed. Oftentimes, these entities are multivariate functions, i.e., functions with more than one-dimensional domain. Examples include brain activity recordings generated during fMRI or EEG experiments (van Bömmel et al. (2014), Majer et al. (2015)). In a variety of applications though, the object of interest is not directly observable but can be recovered from the observed data by means of derivative. Typical examples of financial applications are functionals retrieved from the observed prices, such as implied risk neutral or state price density (Grith et al. (2012)), pricing kernel (Grith et al. (2013)) or market price of risk (Härdle and Lopez-Cabrera (2012)). Motivated by such data analysis situations, we address the problem of estimating derivatives of multivariate functions from existing discrete and noisy data.

Functions, which are objects of an infinite-dimensional vector space, require specific methods that allow a good approximation of their variability with a small number of components. FPCA is a convenient tool to address this task because it allows us to explain complicated data structures with only a few orthogonal principal components that fulfill the optimal basis property in terms of its L^2 accuracy. These components are given by the Karhunen-Loève theorem, see for instance Bosq (2000). In addition, the corresponding principal loadings to this basis system can be used to study the variability of the observed phenomena. An important contribution in the treatment of the finite-dimensional PCA was done by Dauxois et al. (1982), followed by subsequent studies that fostered the applicability of the method to samples of observed noisy curves. Besse and Ramsay (1986), among others, derived theoretical results for observations that are affected by additive errors. Some of the most important contributions for the extension of the PCA to functional data belong to Cardot et al. (1999), Cardot et al. (2007), Ferraty and Vieu (2006), Mas (2002) and Mas (2008). Simple, one-dimensional spatial curves are well understood from both numerical and theoretical perspectives and FPCA is easy to implement in this case. Multivariate 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, computational issues are considerable in higher-dimensional domain.

To our best knowledge, FPCA for derivatives has been tackled by Hall et al. (2009) and Liu and Müller (2009). The first study handles one-dimensional directional derivatives and gradients. The second paper analyses a particular setup in one-dimensional domain where the observations are sparse. This method is applicable to non-sparse data but can be computationally inefficient when dealing with large amounts of observations per curve. For the study of observed functions, there are a series of empirical studies for the two-dimensional domain case, see Cont and da Fonseca (2002) for an application close to our empirical study. Further proposals to implement FPCA in more than two dimensions to analyze functions, rather than their derivatives, have been done particularly in the area of brain imaging, see for instance, Zipunnikov et al. (2011) who implement multilevel FPCA (Staicu and Carroll (2010), Di et al. (2009)) to analyze brain images of different groups of individuals. However, a thorough derivation of statistical properties of the estimators is missing in these works.

In this paper, we aim to fill in the existent gaps in the literature on FPCA for the study of derivatives of multivariate functions. We present two alternative approaches to obtain the derivatives. The paper is organized as follows: the theoretical framework, estimation procedure and statistical properties are derived through Section 2. 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 multivariate functions using FPCA

In this section, we review FPCA from a technical point of view and make the reader familiar with our notation.

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 $\sigma(t,v) \overset{\text{def}}{=} \mathsf{E}\left[X(t)X(v)\right]$ 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 such that

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

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

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

for $d = (d_1,\ldots,d_g)^\top$ 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 is stated in terms of the Karhunen-Loève decomposition applied to their covariance function. We define $\sigma^{(d)}(t,v) \overset{\text{def}}{=}$

$\mathsf{E}\left[X^{(d)}(t)X^{(d)}(v)\right]$ and $\lambda_1^{(d)} \geq \lambda_2^{(d)} \geq \ldots$ be the corresponding eigenvalues. The principal components $\varphi_r^{(d)}$, $r = 1, 2, \ldots$ 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 $|d| = 0$, we introduce the following notation $\varphi_r^{(0)}(t) \equiv \gamma_r(t)$. Similarly to equation (2), the decomposition of $X^{(d)}$ in terms of principal components $\varphi_r^{(d)}(t)$ is given by

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

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 $a_{rp} = \left\langle \gamma_p^{(d)}, \varphi_r^{(d)} \right\rangle = \int_{[0,1]^g} \gamma_p^{(d)}(t)\varphi_r^{(d)}(t)\,dt$ such that $\sum_{r=1}^{\infty} a_{rp}\varphi_r^{(d)}(t) = \gamma_p^{(d)}(t)$. However, if we consider a truncation of (2) after a finite number of components this is no longer true in general. An advantage of using $\varphi_r^{(d)}$ instead of $\gamma_r^{(d)}$ is that the decomposition gives orthonormal basis that fulfill the best basis property, such that for any fixed $L \in \mathbb{N}$ and every other orthonormal basis system φ'

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

This guarantees that by using $\varphi_r^{(d)}$, $r = 1, \ldots, L$ we always achieve the best L dimensional subset selection in terms of the L^2 error function. In the next section we show that estimating the basis functions with such property comes at the cost of inferior rate of convergence. However, if the true underlying structure lies in a L-dimensional function space, which is equivalent to a factor model setup, the advantage of deriving the best L-orthogonal basis vanishes, because it is possible to derive a basis system with the same features using $\text{span}(\gamma^{(d)})$. This is achieved by performing a spectral decomposition of the finite-dimensional function space of $\gamma_r^{(d)}$, $r = 1, \ldots, L$ to get an orthonormal basis system fulfilling (8).

2.2 Sample inference

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

N curves is given by the sample counterpart

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

and for the covariance operator by

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

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

2.3 The model

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

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

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

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 Kneip and Utikal (2001) to estimate density functions and later adapted by Benko et al. (2009) for general functions. Let $\nu = (\nu_1,\ldots,\nu_g)^\top,\ \nu_j \in \mathbb{N}$, $|\nu| < \rho \le m$ 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 $l_r^{(\nu)}$ be the eigenvalues of matrix $M^{(\nu)}$ and $p_r^{(\nu)} = (p_{1r}^{(\nu)},\ldots,p_{Nr}^{(\nu)})$ be the corresponding eigenvectors. For $\nu = d$, the estimators for the quantities in the right-hand side of equations (4) and (5) are given 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 the corresponding orthonormal basis $\hat{\gamma}_r \overset{\text{def}}{=} \hat{\varphi}_r^{(0)}$ and loadings $\hat{\delta}_{ri} \overset{\text{def}}{=} \hat{\delta}_{ri}^{(0)}$. It is straightforward to derive

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

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. For any vectors $a,\ b \in \mathbb{R}^g$ and $c \in \mathbb{N}^g$, we define $|a| \overset{\text{def}}{=} \sum_{j=1}^{g} |a_j|$, $a^{-1} \overset{\text{def}}{=} (a_1^{-1}, \ldots, a_g^{-1})^\top$, $a^b \overset{\text{def}}{=} a_1^{b_1} \times \ldots \times a_g^{b_g}$, $a \circ b \overset{\text{def}}{=} (a_1 b_1, \ldots, a_g b_g)^\top$ and $c! \overset{\text{def}}{=} c_1! \times \ldots \times c_g!$.

Consider a curve Y observed at points $t = \{t_1, \ldots, t_T\}$, generated as in equation (11). Let $k = (k_1, \ldots, k_g)^\top$, $k_l \in \mathbb{N}$ and consider a multivariate local polynomial estimator $\hat{\beta}(t) \in \mathbb{R}^\rho$ that solves

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

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

As noted by Fan et al. (1997) the solution of the minimization problem (15) can also be represented using a weight function W_ν^T, see Appendix 5.2, such that

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

Local polynomial regression estimators are better suited to estimate integrals like (12) than other kernel estimators, e.g., Nadaraya-Watson or Gasser-Müller estimator, since the bias and variance are of the same order of magnitude near the boundary as well as in the interior, see for instance Fan and Gijbels (1992). We propose the following estimator for the squared integrated functions $\int_{[0,1]^g} X^{(\nu)}(t)^2 dt$

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

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

Lemma 2.1 *Under Assumptions 5.1- 5.4, X is $m \ge 2|\nu|$ times continuously differentiable, the local polynomial regression is of order ρ with $|\nu| \le \rho < m$ and $|\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2| =$*

$\mathcal{O}_P(T^{-1/2})$. *As* $T \to \infty$ *and* $\max(b)^{\rho+1} b^{-\nu} \to 0$, $\frac{\log(T)}{T b_1 \times \ldots \times b_g} \to 0$ *as* $T b_1 \times \ldots \times b_g b^{4\nu} \to \infty$

$$
\begin{aligned}
&\mathsf{E}_{t,Y}\left[\theta_{\nu,\rho}\right] - \int_{[0,1]^g} X^{(\nu)}(t)^2 \, dt = \mathcal{O}_p\left(\max(b)^{\rho+1} b^{-\nu} + \frac{1}{T^{3/2}(b^{2\nu} b_1 \times \ldots \times b_g)}\right) \\
&\mathrm{Var}_{t,Y}(\theta_{\nu,\rho}) = \mathcal{O}_p\left(\frac{1}{T^2 b_1 \times \ldots \times b_g b^{4\nu}} + \frac{1}{T}\right),
\end{aligned} \tag{18}
$$

where $\mathsf{E}_{t,Y}$ denotes the conditional expectation and $\mathrm{Var}_{t,Y}$ the conditional variance given t, Y. 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. For uniformly sampled points $t_1, \ldots, t_T \in [0,1]^g$ with $T = \min_{i \in 1,\ldots,N} T_i$, we replace the integrals in (17) with the Riemann sums, such that

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

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

There are two possible sources of error in the construction of the estimator $\hat{M}^{(\nu)}$. One is coming from smoothing noisy curves at a common grid, and has been analyzed in Lemma (2.1). The other one is from approximating the integral in (17) by a sum, see equation above. In Appendix (5.3) we show that the error of the integral approximation is of order $T^{-1/2}$.

Proposition 2.2 *Under the requirements of Lemma 2.1*

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

By Proposition 2.2, estimating $M^{(d)}$ gives an asymptotic higher bias and also a higher variance than estimating $M^{(0)}$. This effect becomes more pronounced in higher dimensional domain. However, by using local polynomial regression with large polynomial order 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 estimate* $M^{(\nu)}$ *such that if* $m > \rho \geq \frac{g}{2} - 1 + 3\sum_{l=1}^{g} \nu_l$, $b = T^{-\alpha}$ *with* $\frac{1}{2(\rho+1-\sum_{l=1}^{g} \nu_l)} \leq \alpha \leq \frac{1}{g+4\sum_{l=1}^{g} \nu_l}$ *then* $|M_{ij}^{(\nu)} - \hat{M}_{ij}^{(\nu)}| = \mathcal{O}_P(1/\sqrt{T})$.

We can see that the orders of polynomial expansion and the bandwidths for estimating $M^{(\nu)}$ will differ for $\nu = (0, \ldots, 0)^\top$ and $\nu = d$. In particular, the estimator of $M^{(d)}$ requires higher smoothness assumptions via $m > \rho$, and a higher bandwidth to achieve the same parametric convergence rate as the estimator for $M^{(0)}$.

In Lemma 2.1 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 $\sigma^2_{i\varepsilon}$, which achieve the desired convergence rate, are given by successive differentiation, see von Neumann et al. (1941) and Rice (1984). However, as pointed out in Munk et al. (2005), difference estimators are no longer consistent for $g \geq 4$ due to the curse of dimensionality. A possible solution is to generalize the kernel based variance estimator proposed by Hall and Marron (1990) for the multidimensional domain

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

where $w_{ilk} = K_{r,H}(t_{il} - t_{ik}) / \sum_{k=1}^{T_i} K_{r,H}(t_{il} - t_{ik})$ and $\nu_i = T_i - 2\sum_l w_{ilk} + \sum_{l,k} w^2_{ilk}$ and $K_{r,H}$ is a g-dimensional product kernel of order r with bandwidth matrix H. Munk et al. (2005) show that if $4r > g$ and if the elements of the diagonal matrix H are of order $\mathcal{O}(T^{-2/(4r+g)})$ then the estimator $\hat{\sigma}_{\varepsilon i}$ in equation (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 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 the convergence rate of $\tilde{M}^{(0)}_{ij}$ does not depend on g.

When working with more than one spatial dimension, in practice data is often recorded using an equidistant grid with T points in each direction. For our approach, this strategy will not improve the convergence rate of $\tilde{M}^{(0)}$ due to the curse of dimensionality. If it is possible to influence how data is recorded, we recommend using a common random grid, which keeps computing time and the storage space for data to a minimum and still gives parametric convergence rates for the estimator of $M^{(0)}_{ij}$. If $T \gg N$ equation (20), gives a straightforward explanation why the dual matrix is preferable to derive the eigendecomposition of the covariance operator, because taking sums has a computational cost that is linear.

4. Estimating the basis functions — We keep notations $\nu = d$ to refer to the specification in equation (5) and $\nu = (0, \ldots, 0)^\top$ to (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, compared for example to spline or wavelet estimators, is that the bias and variance can be derived analytically. For the univariate case these results can be found

in Fan and Gijbels (1996) and for the multivariate case in Masry (1996) and Gu et al. (2015). We summarize them in terms of order of convergence below

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

Using these results, it follows that for $\max(h)^{p+1}h^{-d} \to 0$, $\left(\max(h)^{p+1}Th^{-d}\right)^{-1} \to 0$ as $T \to \infty$ and p chosen such that $p - |d|$ is odd

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

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

In the next Proposition 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 these 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 $h_{r,opt} = \mathcal{O}_p\left((NT)^{-1/(g+2p+2)}\right)$. Even if the optimal bandwidth for both approaches and each basis function 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 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 with L components is assumed, the basis representation to reconstruct $X^{(d)}$ is arbitrary, in the 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 $\gamma_r^{(d)}(t) = \sum_{s\neq r} a_{sr}\gamma_s^{(d)}(t)$.

Based on the methodology described in Section 2.4, the two estimators for derivatives are given by

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

Proposition 2.5 *Assume that a factor model with L factors holds for X. For $NT^{-1} \to 0$, together with the requirements of Proposition 2.4, the true curves can be reconstructed*

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

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

A proof of Proposition (2.5) is given in Appendix (5.5). Compared with the convergence rates of the individual curves estimators, see (21), the variance of our estimators reduces not only in T but also in N. Equations (13) and (14) can be interpreted as an average over N curves for only a finite number of L components. The intuition behind it is that only those components are truncated that are related to the error term and thus a more accurate fit is possible. If N increases at a certain rate, it is possible to get close to parametric rates. Such rates are not possible when smoothing the curves individually.

For the estimation of $\hat{X}^{(d)}_{i,FPCA_2}$, as illustrated in Remark 2.3, 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 $|d|$ it is required that the true curves become much smoother which makes the applicability of estimating $\hat{X}^{(d)}_{i,FPCA_2}$ limited for certain applications. In contrast, the estimation of $M^{(0)}$ still gives almost parametric rates if less smooth curves are assumed. In addition, if the sample size is small, using a high degree polynomial needed to estimate $M^{(d)}$ might lead to unreliable results. To learn more about these issues, we check the performance of both approaches in a simulation study in Section 3.2 using different sample sizes.

3 Application to state price densities implied from option prices

In this section we analyze the state price densities (SPDs) implied by the stock index option prices. As state dependent contingent claims, options contain information about the risk factors driving the underlying asset price process and give information about expectations and risk patterns on the market. Mathematically, SPDs are equivalent martingale measures for the stock index and their existence is guaranteed in the absence of arbitrage plus some technical conditions. In mathematical-finance terminology they are known as risk neutral densities (RNDs). A very restrictive model, with log-normal marginals for the asset price, is the Black-Scholes model. This model results in log-normal SPDs that correspond to a constant implied volatility surface

across strikes and maturity. This feature is inconsistent with the empirically documented volatility smile or skew and the term structure, see Rubinstein (1985). Therefore, richer specifications for the option dynamics have to be used. Most of earlier works adopt a static viewpoint; they estimate curves separately at different moments in time, see the methodology reviews by Bahra (1997), Jackwerth (1999) and Bliss and Panigirtzoglou (2002). In order to exploit the information content from all data available, it is reasonable to consider them as collection of curves.

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

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

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

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 by the no-arbitrage condition, the forward price for maturity τ is

$$F_\tau = \int_0^\infty s_\tau q(s_\tau,\tau) ds_\tau = s_0 \exp(r_\tau \tau). \tag{26}$$

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

$$C(k,\tau) = F_\tau C(k/F_\tau,\tau). \tag{27}$$

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

$$\frac{\partial^2 C(k,\tau)}{\partial k^2} = \frac{1}{F_\tau} \frac{\partial^2 C(m,\tau)}{\partial m^2}. \tag{28}$$

Then one can show that for $d = (2,0)^\top$, $C^{(d)}(m,\tau)|_{m=s_\tau/F_\tau} = q(s_\tau/s_0,\tau) = s_0 q(s_\tau,\tau)$. In practice, it is preferable to work with densities of returns instead of prices when analyzing them jointly because prices are not stationary. Also, notice that call price curves are not centered. This leads to an additional additive term in equations (4) and (6), which refers to the population mean. We show in the next section how to handle this in practice. For our application, X will refer to the rescaled call price $C(m,\tau)$. Therein, we also assume that the index $i = 1,\ldots,N$ refers to ordered time-points.

The code used to generate the results reported in this section is published online at www.github.com/QuantLet/FPCA and www.quantlet.de. The data used in the empirical study is available from the authors upon request.

3.1 Implementation

1. Centering the observed curves — Throughout the paper it is assumed that the curves are centered. To satisfy 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

$$\overline{M}_{ij}^{(\nu)} = \hat{M}_{ij}^{(\nu)} - \frac{1}{T}\sum_{k=1}^{T}\left(\bar{X}^{(\nu)}(t_k)\hat{X}_{i,b}^{(\nu)}(t_k) + \bar{X}^{(\nu)}(t_k)\hat{X}_{j,b}^{(\nu)}(t_k) - \bar{X}^{(\nu)}(t_k)^2\right). \tag{29}$$

It is still possible to improve the centering 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 significant amount of additional computational effort for only minor improvements.

2. Bandwidth selection — To get parametric rates of convergence for $\hat{M}^{(d)}$ related to Remark 2.3 we choose $\rho = 7$ and b between $\mathcal{O}(T^{-1/10})$ and $\mathcal{O}(T^{-1/12})$. The choice of b to estimate $\hat{M}^{(0)}$ is similar, with the difference that $\rho > 0$, we choose $\rho = 1$ and b has to lie between $\mathcal{O}(T^{-1/3})$ and $\mathcal{O}(T^{-1/5})$. We use a simple criteria to choose the bandwidth because by Proposition 2.4 the dominating error depends mainly on the choice of h. Let $t_{ik} = (t_{ik1},\dots t_{ikg})$, then the bandwidth for direction j is determined by $b_j = \left((\max_k(t_{ikj}) - \min_k(t_{ikj}))T_i\right)^{\alpha}$. When estimating state price densities $t_{ik} = (\tau_{ik}, m_{ik})$ and T_i is replaced by the cardinality of $\tau_i = \{\tau_{i1},\dots\tau_{iT_i}\}$ and m_i respectively. In the estimation of $\hat{M}^{(d)}$ we set $\alpha = -1/10$ and $\alpha = -1/3$ for $\hat{M}^{(0)}$.

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

$$h_{r,opt}^{d,\nu} = C_{d,p}(K)\left[T^{-1}\frac{\int_0^1 \sum_{i=1}^{N}(p_{ir}^{(\nu)})^2\sigma_{\varepsilon i}^2(t)f_i(t)^{-1}dt}{\int_0^1\left\{\sum_{i=1}^{N}p_{ir}^{(\nu)}X_i^{(p+1)}(t)\right\}^2 dt}\right]^{1/(2p+3)}, \tag{30}$$

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

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

For our bandwidth estimator we treat every dimension separately, similar to choosing an optimal an optimal bandwidth for derivatives in the univariate case, and correct for the asymptotic order, see Section 2.4. In practice, we can not use equation (30) to determine the optimal bandwidth because some variables are unknown and only discrete points are observed. As a rule-of-thumb, we replace these unknown variables with empirical quantities: estimates of $p_{ir}^{(0)}$ from $\hat{M}^{(0)}$ and of $p_{ir}^{(d)}$ from $\hat{M}^{(d)}$.

With these approximations, a feasible rule for computing the optimal bandwidth in direction j for the r-th basis function h_{jr} is given by

$$h_{jr,rot}^{d,\nu} = \left(T^{-1} \frac{C_{d,p}^{2p+3} \hat{\sigma}_\varepsilon^2}{f_j \int_0^1 \left\{ \sum_{i=1}^N \hat{p}_{ir}^{(\nu)} \tilde{X}_i^{(p+1)}(t_j) \right\}^2 dt_j} \right)^{1/(g+2p+2)}. \tag{31}$$

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

- The density of the observed points is approximated by a uniform distribution with $f_1 = \max_{i,j}(\tau_{ij}) - \min_{i,j}(\tau_{ij})$ and $f_2 = \max_{i,j}(m_{ij}) - \min_{i,j}(m_{ij})$.
- To get a rough estimator for $X_i^{(p+1)}$ based on X_i, we use a polynomial regression. For our application, we take $p = 3$ and are thus interested in estimates for $X_i^{(4)}(m)$ and $X_i^{(4)}(\tau)$. We expect the curves to be more complex in the moneyness direction than in the maturity direction and we adjust the degree of the polynomials to reflect this issue. The estimates are then given by

$$\begin{aligned} a_i^* &= \arg\min_{a_i} \left(X_i(m,\tau) - a_{i0} + \sum_{l=1}^{5} a_{il} m^l + \sum_{l=6}^{9} a_{il} \tau^{(l-5)} \right) \\ \tilde{X}_i^{(4)}(m) &= 24 a_{i4}^* + 120 a_{i5}^* m \\ \tilde{X}_i^{(4)}(\tau) &= 24 a_{i9}^*. \end{aligned} \tag{32}$$

- 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. In addition, these estimates are used to correct for the diagonal bias when $\hat{M}^{(0)}$ and $\hat{M}^{(d)}$ are estimated. In (31) the average over all $\hat{\sigma}_{ic}$ is used.

We use the product of Gaussian kernel functions to construct local polynomial estimators. We can verify from Proposition 2.4 that the optimal bandwidth h decreases in N. By using a global bandwidth and a compact kernel the matrix given in equation (45) may become singular when N is large and T is small.

In our simulation and application we use the mean optimal $h_{i,rot}^{d,\nu} = L^{-1} \sum_{r=1}^{L} h_{ir,rot}^{d,\nu}$ for each $\hat{\gamma}_r^{(d)}, \hat{\varphi}_r^{(d)}$ to reduce computation time. Since we demean the sample in (29), finally we add $N^{-1} \sum_{i=1}^{N} \hat{X}_{i,h_{i,rot}^{d,\nu}}^{(d)}$ to the resulting truncated decomposition to derive the final estimate.

3. Estimation of the number of components— In Section 2.5 we assumed that the number of components is given. In general, the number of basis functions needed is unknown a priori. For the case $|d| = 0$ there exists a wide range of criteria that can be adapted to our case to determine the upper bound L. The easiest way to determine the number of components is by choosing the model accuracy by an amount of variance explained by the eigenvalues. In (71) 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 Bai and Ng (2002) can be adapted to our case and give several criteria for finding L or L_d by generalizing Mallows (1973) C_p criteria for panel

data settings. These criteria imply minimizing the sum of squared residuals when k factors are estimated and penalizing the overfitting. One such formulation suggests choosing the number of factors using the criteria

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

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

$$IC^{(v)}(k^*) = \min_{k\in\mathbb{N},k\leq L} \left[\log \left(\frac{1}{N} \sum_{r=k+1}^{N} \hat{\lambda}_r^{(v)} \right) + k \left(\frac{\log(C_{NT}^2)}{C_{NT}^2} \right) \right]. \tag{34}$$

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

$$\text{Var}(\hat{\delta}_{r,T}^{(d)} \hat{\gamma}_{r,T}^{(d)}) = \langle \hat{\gamma}_{r,T}^{(d)}, \hat{\gamma}_{r,T}^{(d)} \rangle \hat{\lambda}_r. \tag{35}$$

We can sort the variances in decreasing order and use either equation (33) or (34) 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 G components, $q(m,\tau) = \sum_{l=1}^{G} w_l q^l(m,\tau)$, where q^l are fixed basis functions and w_l are random weights. For fixed τ we consider $q^l(\cdot,\tau)$ to be a log-normal density functions, with mean $\left(\mu_l - \frac{1}{2}\sigma_l^2\right)\tau$ and variance $\sigma_l^2\tau$, and simulate weights w_{il} with $\sum_{l=1}^{G} w_{il} = 1$, where $i = 1,\ldots,N$ is the index for the day, then

$$q_i(m,\tau) = \sum_{l=1}^{G} w_{il} \frac{1}{m\sqrt{2\pi\sigma_l^2\tau}} \exp\left[-\frac{1}{2} \left\{ \frac{\log(m) - \left(\mu_l - \frac{1}{2}\sigma_l^2\right)\tau}{\sigma_l\sqrt{\tau}} \right\}^2 \right]. \tag{36}$$

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

$$C_i(m,\tau) = \exp(-r_{i\tau}\tau) \sum_{l=1}^{G} w_{il} \left\{ \exp(\mu_l\tau)\Phi(y_1) - m\Phi(y_2) \right\} \tag{37}$$

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

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

Table 1. Results of the simulation described in Section 3.2 with different values for T and N. $FPCA_1$ and $FPCA_2$ are superior in sense of MSE over the individual estimation of the derivatives in each setting. $FPCA_1$ is better than $FPCA_2$ except for $N = 10$, $T = 250$. For $FPCA_1$ and $FPCA_2$ the estimation improves with raising N and T. These results support our asymptotic results given by Proposition 2.2 and 2.5.

FPCAsimulation

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

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

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

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

The true SPDs given by equation (36) are used to verify the performance of $\hat{X}^{(d)}_{FPCA_1}$, $\hat{X}^{(d)}_{FPCA_2}$ and of the individually estimated curves $\hat{X}^{(d)}_{Indiv.}$, in terms of mean integrated squared error (MSE), i.e., $T^{-1}\sum_{k=1}^{T}\left\{X^{(d)}(t_{ik}) - \hat{X}_\bullet^{(d)}(t_{ik})\right\}^2$, for $d = (2,0)$. For evaluation we generate a common grid of 256 points from a uniform distribution. To derive the optimal bandwidth in each case we stick to the rule-of-thumb approach presented in Section 3.1. The bandwidth for the individually smoothed curve i is derived by replacing $\hat{p}_{ir}^{(\nu)}$ in (31) by one and zero otherwise. The performance is recorded for sample sizes N of 10 and 25 with T observations per day of size 50 and 250. This procedure is repeated 500 times to get reliable results, mean, variance and the inter quartile distance based at the MSE of the repetitions are given in Table 1.

Both FPCA based approaches give better estimates for the derivative of the call functions than an individually applied local polynomial estimator of the individual curves. Both the mean and the median of the MSE are smaller which is a result of the additional average over N for the basis functions as given by Proposition 2.5. However, the $FPCA_1$ method performs decisively better for small T than the other two both in terms of mean and standard deviation of the mean squared error. In addition $FPCA_1$ benefits more from increasing N than $FPCA_2$. With small T for $FPCA_2$ and individual smoothing the variability of MSE is much bigger than for $FPCA_1$ while the median of $FPCA_1$ and $FPCA_2$ are comparable. This means individual smoothing and $FPCA_2$ must behave much worse than $FPCA_1$ in some instances while $FPCA_1$ was able to stabilize the estimates. To get the same effect using $FPCA_2$ a much bigger T is needed. A possible explanation for this behavior is given by Proposition 2.2. The rates of convergence for the estimators of the dual matrix entries rely on T. Thus in finite sample, when T is small, the estimated loadings might be biased.

3.3 Real Data Example

1. Data description —We use settlement European call option prices written on the underlying DAX 30 stock index. These prices are computed by EUREX at the close of each trading day as an average of the intraday transaction prices. The data range is ten years, between January 2, 2002 and December 3, 2011, and includes 2557 days. The expiration dates of the options are set on the third Friday of the month. Therefore, on a particular day, option prices with only a few maturities are available, see Figure 1. The distance between two consecutive observed maturities is higher for more distant expiration dates, while the distance between two consecutive strike prices is relatively constant. Methods that analyze curves jointly are generally better tailored to this type of data, because they provide better estimates at grid points with only a few observations available of the individual curves. We include call options with maturity between one day and one year. The sample contains prices of options with an average of six maturities and sixty-five strikes per day.

By assuming 'sticky' coordinates for the daily observations, in accordance with equation (27), we divide the strike and the call prices within one day by the stock index forward price to ensure that the observation points are in the same range. Afterwards, we apply the estimation methodology described in Section 2 to the rescaled call prices as a function of moneyness and maturity. Our proxy for the risk-free interest rates are the EURIBOR rates, which are listed daily for several maturities. We apply a linear interpolation to calculate the rate values for desired maturities.

2. Estimation results — We report the results for the loadings estimated by spectral decomposition of dual covariance matrix for option price functions, and the estimates of the second partial derivative of the functional principal components. The first eigenvalue of the dual covariance matrix $\hat{M}^{(0)}$ for the call option surfaces has a dominant explanatory power. The order of magnitude of the following eigenvalues decreases by a factor of ten for every few additional components. To detect the relative contribution of consecutive components, we construct the ratio of two adjacent estimated eigenvalues in descending order, see Ahn and Horenstein (2013). The first

two terms are dominating the sequence and there are spikes at the fourth and seventh component ratio. $PC^{(0)}$ criterion suggests at least seven components, see values of k^* for $L_{max} \geq 7$ in Table 2. $IC^{(0)}$ criterion, which does not depend on the truncation parameter L_{max}, suggests seven components.

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

Table 2. Estimated eigenvalues and eigenvalue ratios. Number of factors by $PC^{(0)}$ criterion

FPCArealdata

A closer look at the dynamics of the loadings $\hat{\delta}_{2,T}$ shows an unusual behavior between mid-February 2007 through mid-June 2008. This interval spans the financial crisis and extends until 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 particular time regularity of jumps. We identify these jumps with the Mondays following an expiration date (options expire at a monthly frequency, always on a Friday). Figure 1 highlights the dynamics of $\hat{\delta}_{2,T}$ on and following an expiration day. After roughly two weeks, the loadings revert to a 'normal' level. During this period, for small maturities, there are only few observations available for call prices with strikes larger than the current stock index. In addition, the absence of a call string with close enough time to maturity on the following trading Monday, introduces bias in the estimated smooth call surface for grid values outside the observation points range. The shape of the second estimated component $\hat{\gamma}^{(d)}_{2,T}$, displayed in Figure 1, suggests that it is related to variations of the short end of the SPD term structure. A similar behavior of the loadings are observed for a few other components we investigated: $\hat{\delta}_{4,T}$, $\hat{\delta}_{5,T}$ and $\hat{\delta}_{6,T}$. Their variances remain important even if we exclude the interval with large jumps from the sample. The corresponding components have similar shape features to the three components we discuss below. We argue that they impact the option prices and SPDs when jumps in the underlying occur, and that they related to the asymmetric behavior of the option prices along the maturity direction.

The estimated components $\hat{\gamma}^{(d)}_{1,T}$, $\hat{\gamma}^{(d)}_{3,T}$ and $\hat{\gamma}^{(d)}_{7,T}$ together with their loadings are displayed in Figure 2. They describe three types of asymmetry present in the dynamics of the SPDs. The first component, is similar in shape to the empirical mean of the SPD. It has a long left tail, specific to the negatively skewed densities and a peak located at a value of moneyness slightly above one. For positive levels of the loadings, this component increases the mass of SPD around the mode and decreases the values in the tails. We find that this component is related to the time-varying volatility of the index re-

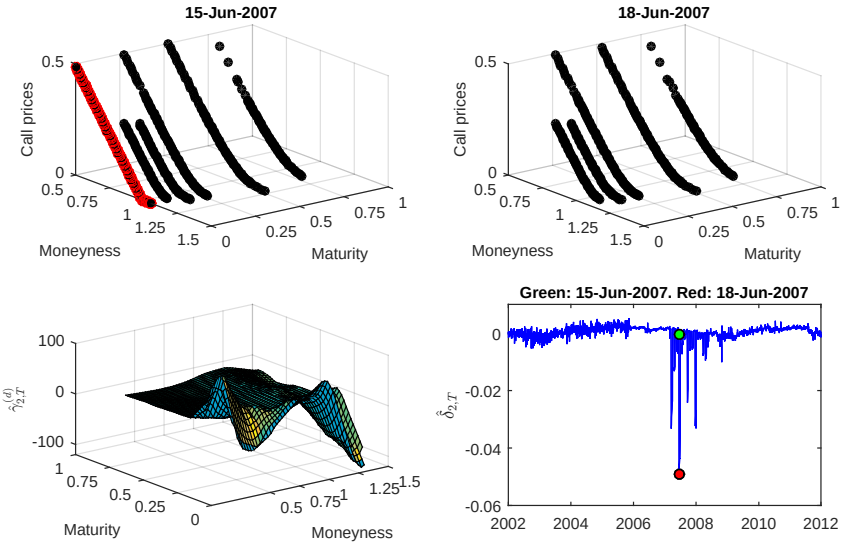

Figure 1. The effect of expiration date on the level of estimated loadings $\hat{\delta}_{2,T}$

FPCAexpiration

turns. The next component, $\hat{\gamma}^{(d)}_{3,T}$ has a 'valley-hill' pattern, which shifts mass around the central region of the density. A positive shock in the direction of this components increases the negative skewness, and a large negative shock can reverse the sign of the SPD skewness. This component is interpreted as a skewness factor. The last component, $\hat{\gamma}^{(d)}_{7,T}$ takes negative values in the left tail and displays a prominent positive valued peak at the right of the mode of the empirical SPD mean. This component represents a tail factor, and we show that its loading can be interpreted as the volatility of volatility index.

The functional principal components of the reduced model $\sum_{r\in\{1,3,7\}}\hat{\delta}_{r,T}\hat{\gamma}^{(d)}_{r,T}$ resemble closely the three components displayed in Figure 2. Further analysis shows that when including any additional component to the reduced representation, the shape of the orthogonal basis changes to some extent. The loadings of all orthogonalized components become 'contaminated' with jumps. Moreover, all the loadings estimated by decomposing $\hat{M}^{(d)}$, for $d=(2,0)^\top$ feature the jump-behavior outlined above, between mid-February 2007 and mid-September 2008. For those reasons, $\hat{M}^{(0)}$ decomposition enables a better interpretation of the components, by separating the continuous and discontinuous sources of variation in the SPDs.

We show next that the first estimated component $\hat{\gamma}^{(d)}_{1,T}$ is related to the expected variance under a risk neutral measure Q, which admits the density q. Under this measure, the prices are martingales. Equations (6) and (26) yield

$$\frac{\mathsf{E}^Q_i\left(s_{i+\tau}/F_i\right)}{\exp(r_{i\tau}\tau)}=\int_0^\infty m\tilde{q}(m,\tau)dm+\sum_{r=1}^{\infty}\delta_{ir}\int_0^\infty m\gamma^{(d)}_r(m,\tau)dm=1, \tag{38}$$

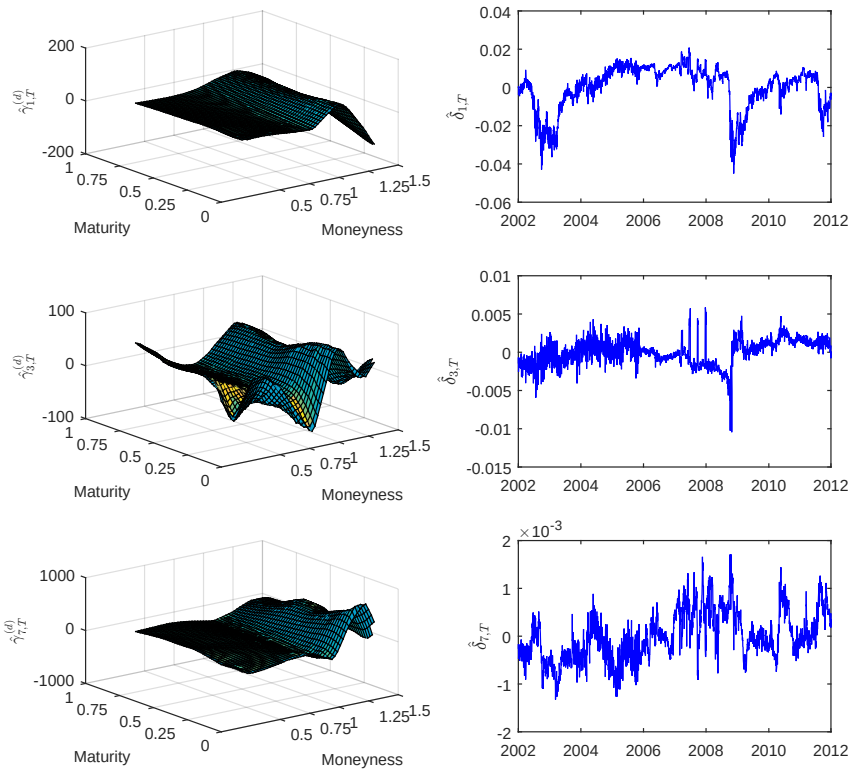

Figure 2. Estimated components $\hat{\gamma}^{(d)}_{1,T}$, $\hat{\gamma}^{(d)}_{3,T}$ and $\hat{\gamma}^{(d)}_{7,T}$ and their loadings obtained by the decomposition of the dual covariance matrix $\hat{M}^{(0)}$

FPCAcomponents

where $\tilde{q}$ is the population mean. The computation of the second moment gives

$$\frac{\mathrm{Var}_i^Q\left(s_{i+\tau}/s_i\right)}{\exp(r_{i\tau}\tau)^2} = \int_0^\infty m^2\tilde{q}(m,\tau)dm + \sum_{r=1}^{\infty}\delta_{ir}\int_0^\infty m^2\gamma_r^{(d)}(m,\tau)dm - 1. \tag{39}$$

We consider the empirical version of Equation (39), for $\tau = 1$ month. Instead of computing the integrals, based on our estimates of $\tilde{q}$ and $\gamma^{(d)}$, we assume them to be unknown coefficients in a linear regression, in which the empirical loadings are used as explanatory variables of the real-data proxy for the standardized variance. In the numerator, we use the squared VDAX index multiplied by τ. This index is computed by Deutsche Börse AG from the prices of call and put options and reflects market expectation under the risk neutral measure of the 30 day ahead square root implied variance for the DAX 30 log-returns, which is then annualized. Duan and Yeh (2010) show that squared volatility index is a good approximation of the expected risk-neutral volatility when the jumps are small. While the volatility index refers to the standard deviation of the log-returns under the risk neutral measure, it can still be used in the regression because the transformation $q(\log m,\tau) = mq(m,\tau)$ maintains the linear-relationship

between the dependent and explanatory variables. We find that the most important component in the regression is $\hat{\delta}_{1,T}$ (adjusted R-squared in the univariate regression is 93.97%). When including $\hat{\delta}_{3,T}$ as an additional regressor, it increases the adjusted R-squared to 94.06%, while $\hat{\delta}_{7,T}$ has a negative marginal contribution to the goodness of fit of multivariate regression.

No skewness index is readily available, and we take a simple measure instead, Pearson's skewness coefficient. In terms of equations (38) and (39), for a fixed maturity τ, it is equal to

$$\frac{1-\arg\max_m \{q(m,\tau)\}}{\sqrt{\mathrm{Var}_i^Q\left(s_{i+\tau}/s_i\right)}/\exp(r_{i\tau}\tau)}. \tag{40}$$

Since the first component $\hat{\gamma}^{(d)}_{1,T}$ is unimodal (as it is also $\hat{\gamma}^{(d)}_{2,T}$), the SPD mode is mostly affected by the loadings of the third component $\hat{\gamma}^{(d)}_{3,T}$ (and to some extend by those of the seventh component $\hat{\gamma}^{(d)}_{7,T}$).

3. Dynamic analysis of the loadings — In this section, we investigate the dynamics of the loadings in the reduced model. The partial autocorrelation function of all three time series display a salient spike at the first lag. This suggests that an autoregressive or perhaps an integrated model of order one might be appropriate to represent their dynamics. Their serial autocorrelations decay slowly, similarly to the integrated processes that feature a stochastic trend. Unit root and stationarity test results (not reported here) are ambiguous. When the null hypothesis assumes the existence of a unit root (augmented Dickey-Fuller unit-root test, Phillips-Perron test, variance-ratio test for random walk) the tests reject the null, while stationary tests that have the unit root hypothesis as an alternative (KPSS test, Leybourne-McCabe stationarity test) favor the alternative. Based on these results, we further investigate if the loadings are fractionally integrated of order $\alpha \in (0,1)$, which is typical to long-memory processes. We employ Lo (1991)'s modified $R/\tilde{S}$ (range over standard deviation) rescaled statistic $\tilde{V}_N$, for a time series sample of N observations. The denominator of the statistic is computed as the square root of Newey and West (1987) estimator of the long run variance of the time series. For a maximum lag $q = [N^{1/4}] = 9$, we obtain $\tilde{V}_{N,1}(9) = 5.1582$, $\tilde{V}_{N,3}(9) = 4.5248$ and $\tilde{V}_{N,7}(9) = 4.9893$, with 95% confidence interval $(0{,}809,\ 1{,}862)$. The tests reject the hypothesis that loadings have short-memory. We also apply Geweke and Porter-Hudak (1983) log-periodogram regression model to estimate the Hurst exponent. The estimates are $H_1^{GPH} = 1.3736$, $H_3^{GPH} = 1.1761$ and $H_7^{GPH} = 1.1433$ for the cutoff $[N^{1/2}] = 50$. The 95% confidence interval (0.2981, 0.7019) for the the GPH estimator is calculated using a bootstrapping procedure proposed by Weron (2002). These estimates imply an order of integration $\hat{\alpha}_r^{GPH} = H_r^{GPH} - 0.5$, $r = 1,3,7$. It is known that in the presence of large autoregressive or moving average terms, $\hat{\alpha}_r^{GPH}$ is biased upwards. In general, these models are nontrivial to estimate by other methods. Furthermore, fractionally integrated processes lack a clear economic interpretation. Therefore, instead of including a large number of autoregressive terms we use a parsimonious AR(1) model with time varying coefficients to approximate the long memory process. This is appropriate also for $\alpha \in (1/2, 1)$, when the loadings are not stationary, see Comte and Renault (1998).

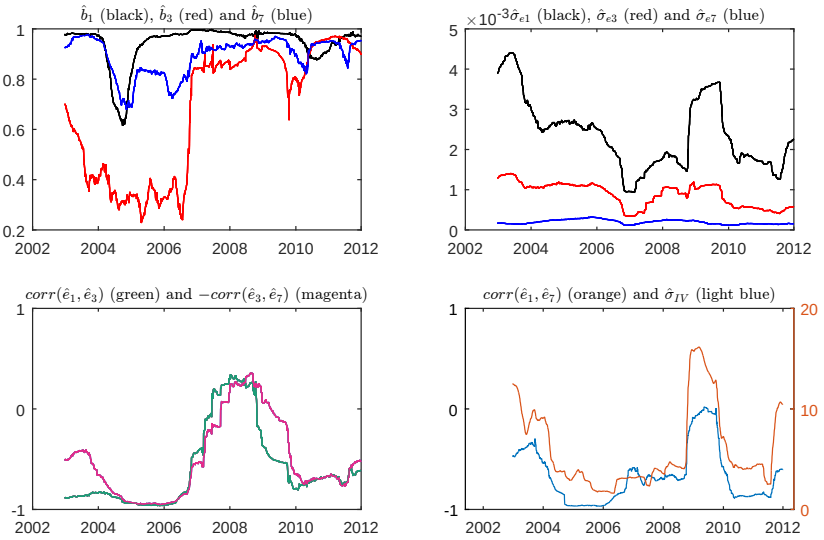

Figure 3. Time varying autoregressive coefficients, standard deviations of the error, pairwise error correlations from fitting a univariate AR(1) model to each loading; and time varying standard deviation of the VDAX volatility index estimated daily with a moving window of 250 observations

FPCAloadings

We consider the following law of motion for the loadings

$$\hat{\delta}_{ir,T} = b_r \hat{\delta}_{i-1r,T} + e_{ir}, \quad \text{Var}(e_{ir}) = \sigma^2_{er}, \tag{41}$$

where b_r is the autoregressive coefficient, $r = 1, 3, 7$. We reestimate equation (41) daily on a moving window of 250 observation using OLS. This adaptive estimation procedure helps detect the possible sources of non-stationarity in the estimated loadings, by allowing the autoregressive coefficient and the error variance to vary over time.

The upper left panel of Figure 3 displays the estimated autoregressive coefficients. $\hat{\delta}_{1,T}$ is very persistent ($\hat{b}_1$ is close to one) for most of the time, except 2004. Interestingly, $\hat{b}_3$ is relatively small between 2003-2006 and increases significantly thereafter, suggesting a possible regime shift. $\hat{b}_7$ is relatively high and its variation seems sensitive to the changes in the other two parameters.

In addition, we compute the time-varying correlations between the error terms in each loading equation. These correspond to the non-diagonal entries of the empirical error covariance matrix for a vector autoregressive model of order one VAR(1) fitted to the loadings, with diagonal autoregressive coefficient matrix. The two lower panels of Figure 3 illustrate the results. The error correlation of the skewness with the volatility and the tail factor loadings move closely together, suggesting a strong relationship between the volatility and the tail factors. We focus on $corr(\hat{e}_1, \hat{e}_3)$, which describes the dynamic relationship between the changes in SDP volatility and skewness. Most of the time, the plotted correlation is negative, meaning that positive changes in the SPD variance are associated on average with increases in the negative skewness. The negative correlation between an asset return and its changes of volatility is generally

known as the leverage effect. The correlation reverses sign and becomes positive between 2007-2009. This means that when volatility increases, there is a change in the concentration mass in the left side of the density, in the area of medium-ranged negative returns. We identify this behavior with the implied volatility skew puzzle, as documented by Constantinides and Lian (2015). The authors rationalize this behavior through the reduction in put option supply from credit-constrained market makers together with an increase in the demand for OTM puts required for hedging purposed, see net buying pressure in Bollen and Whaley (2004), Gârleanu et al. (2009).

Typically, the error correlation $corr(\hat{e}_1, \hat{e}_7)$ is negative. Its magnitude decreases and reaches values close to zero in 2009. In the lower right panel of Figure 3, we also plot the 250-observation standard deviation $\hat{\sigma}_{IV}$ of the VDAX implied volatility index. The two time-series are strongly correlated (the correlation coefficient is 90.78%). This suggests that the tail component can be interpreted as the volatility of volatility risk factor. Similar interpretations were proposed in Du and Kapadia (2012), Huang and Shaliastovich (2014) or Park (2015), who use different measures of the volatility-of-volatility implied by VIX (the implied volatility index of the S&P 500) as a tail risk indicator. The tail factor takes highest positive values during the financial crisis, consistent with fat tail and high risk hypothesis.

To verify the stability of the results reported, we repeat the regression analysis by including a constant in equation (41). The root mean square error does not decrease significantly. We also estimate the model by including the lagged values of the other two loadings as additional explanatory variables. Some of the estimated autoregressive coefficients take value above one. Independently of these modeling choices, the estimated error covariances are very similar to those shown in Figure 3. These suggest that changes in the correlation sign for the levels is due to the error term correlation structure and not to changes in the lagged cross-term interactions.

Several stylized facts emerge from the moving window estimation of autoregressive models for the loadings that summarize the dynamics of SPDs. When volatility is small, the innovations to the volatility, skewness and volatility of volatility loading equations are very strongly correlated. When volatility increases, the correlation structure changes. In particular, the leverage parameter changes sign during the financial crises. By including volatility of volatility as an additional factor, see also Huang and Shaliastovich (2014), our study distinguishes between the volatility induced skewness through the leverage effect and by the volatility of volatility induced skewness, see also Feunou and Tédongap (2012). These findings may have important consequences for the formulation of stochastic volatility models for option pricing. The empirical evidence suggests that the option markets include sources of variation that may not be present in the underlying's dynamics, such as frictions between option demand for hedging purposes and credit constrains for refinancing. It may be possible to formulate a model in which the changes in error correlations modify endogenously, possibly by controlling for the credit availability on the market.

4 Conclusions

In this paper, we describe two methods for representing derivatives of multivariate curves using FPCA techniques. First, spectral decomposition is applied to the dual

covariance matrix of derivatives. Secondly, the dual covariance of the original curves is considered, and derivatives are obtained by differentiation of the functional principal components of the covariance operator. Thus, the second approach expresses the dynamics of derivatives in terms of uncorrelated loadings but the basis functions are no longer orthonormal. We demonstrate that when an underlying factor model is assumed, estimating curve derivatives from observed discrete and noisy data using a low-dimensional representation, the second method performs better both asymptotically and in finite sample.

In the empirical study, we show that the second method provides accurate results for understanding the time variability of implied SPDs. We apply this analysis to DAX 30 index option data observed at daily frequency. Three main factors are identified, which could be linked to the diffusion processes of the stochastic volatility models. The first factor is strongly connected to the risk-neutral variance, conditional upon that no jumps occur. The second factor is related to the time varying conditional skewness induced by the leverage effect. We find that this component of negative skewness declines during the financial crises, possibly as a result of the credit crunch. Time-varying volatility of volatility constitutes an additional risk channel, which manifests as negative tail risk. Further factors are related to the term structure and jump component risk.

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 bounded and continuously differentiable density f with support* $\operatorname{supp}(f) = [0,1]^g$ *and the integrand* $u \in \operatorname{supp}(f)$ *and* $\inf_u f(u) > 0$.

Assumption 5.2 $\mathsf{E}(\varepsilon_{ik}) = 0$, $\operatorname{Var}(\varepsilon_{ik}) = \sigma^2_{i\varepsilon} > 0$ *and ε_{ik} are independent of X_i, and* $\mathsf{E}\left[\varepsilon^4_{ik}\right] < \infty, \forall i,k$.

Assumption 5.3 *Let $K_B(u) = \frac{1}{b_1 \times \dots \times b_g} K(u \circ b)$. K is a product kernel based on symmetric univariate kernels. B is a diagonal matrix with $b = (b_1,\dots,b_g)^\top$ at the diagonal. The kernel K is bounded and has compact support on $[-1,1]^g$ such that for $u \in \mathbb{R}^g$ $\int uu^T K(u)du = \mu(K)I$ where $\mu(K) \neq 0$ is a scalar and I is the $g \times g$ identity matrix. Conditions 2 and 3 from Masry (1996) are fulfilled.*

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

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

Assumption 5.6 *We require that it holds*

$$\sup_{r\in\mathbb{N}} \sup_{t\in[0,1]^g} |\varphi^{(d)}_r(t)| < \infty, \sup_{r\in\mathbb{N}} \sup_{t\in[0,1]^g} |\gamma^{(d)}_r(t)| < \infty \tag{42}$$

$$\sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \mathsf{E}\left[\left(\delta^{(\nu)}_{ri}\right)^2\left(\delta^{(\nu)}_{si}\right)^2\right] < \infty, \sum_{q=1}^{\infty}\sum_{s=1}^{\infty} \mathsf{E}\left[\left(\delta^{(\nu)}_{ri}\right)^2 \delta^{(\nu)}_{si}\delta^{(\nu)}_{qi}\right] < \infty, \ \nu = (0,d) \tag{43}$$

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 $0 < C_{1,r} < \infty$, $0 < C_{2,r} \leq C_{3,r} < \infty$ such that*

$$\begin{aligned} NC_{2,r} \leq l^{(\nu)}_r \leq NC_{3,r} \\ \min_{s=1,..,N;s\neq r} |l^{(\nu)}_r - l^{(\nu)}_s| \geq NC_{1,r}. \end{aligned} \tag{44}$$

5.2 Proof of Lemma 2.1

1. Univariate case (g=1) — In the following proof we use d instead of ν. As noted by Ruppert and Wand (1994) 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. Let e_l be a vector of length ρ with 1 at the $l+1$ position and zero else. Then $W_d^T(t)$ evaluates the function at point u and is defined as $(Tb^{d+1})^{-1} e_d^T S_T(u)^{-1}(1, t, \ldots, t^\rho)^T K(t)$. $S_T(u)$ is a $\rho \times \rho$ matrix with entries $S_{T,k}(u) = (Tb)^{-1}\sum_{l=1}^T K\left(\frac{t_l-u}{b}\right)(\frac{t_l-u}{b})^k$ such that

$$S_T(u) = \begin{pmatrix} S_{T,0}(u) & S_{T,1}(u) & \ldots & S_{T,\rho}(u) \\ S_{T,1}(u) & S_{T,2}(u) & \ldots & S_{T,\rho+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{T,\rho}(u) & S_{T,\rho+1}(u) & \ldots & S_{T,2\rho}(u) \end{pmatrix}. \tag{45}$$

Accordingly

$$\begin{aligned} \mathsf{E}(S_{T,k}(u)) =& (Tb)^{-1}\int_0^1 \sum_{l=1}^T K\left(\frac{x-u}{b}\right)\left(\frac{x-u}{b}\right)^k f(x)dx \\ =& b^{-1}\int_u^{1+u} K\left(\frac{x}{b}\right)\left(\frac{x}{b}\right)^k f(x)dx = \int_{ub^{-1}}^{(1+u)b^{-1}} K(t)\, t^k f(tb)dt. \end{aligned} \tag{46}$$

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 (45) with

$$S(u) = \begin{cases} S_c = (S_{j+l,c})_{0\leq j,l\leq\rho} & , u \text{ is a boundary point} \\ S = (S_{j+l})_{0\leq j,l\leq\rho} & , u \text{ is an interior point} \end{cases}. \tag{47}$$

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

$$\hat{X}_b^{(d)}(u) = d!\beta_d(u) = \frac{d!}{Tf(u)b^{d+1}}\sum_{l=1}^T K_{d,\rho}^{u*}\left(\frac{t_l-u}{b}\right) Y(t_l)\{1 + \mathcal{O}_P(1)\}. \tag{48}$$

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

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

$$\int u^k K_{d,\rho}^*(u)\, du = \delta_{d,k} \;\; 0 \leq d, k \leq \rho. \tag{49}$$

As mentioned by Fan et al. (1997), the design of the kernel automatically adapts to the boundary which gives the same order of convergence for the interior and boundary

points, see Ruppert and Wand (1994). The estimator can be rewritten as

$$
\begin{aligned}
&\int d!^2 \sum_{j=1}^{T}\sum_{l=1}^{T} W_d^T\left(\frac{t_j-u}{b}\right) W_d^T\left(\frac{t_l-u}{b}\right) Y(t_l)Y(t_j)\,du \\
&= \int \frac{d!^2}{T^2 f(u)^2 b^{2d+2}} \sum_{l=1}^{T}\sum_{j=1}^{T} K_{d,\rho}^*\left(\frac{t_j-u}{b}\right) K_{d,\rho}^*\left(\frac{t_l-u}{b}\right) Y(t_l)Y(t_j)\{1+\mathcal{O}_P(1)\}\,du.
\end{aligned} \tag{50}
$$

For the expectation we get

$$
\begin{aligned}
&\mathsf{E}\left(\theta_{d,\rho}|t_1,\ldots,t_T\right) \\
&= \int_0^1 d!^2 \sum_{j=1}^{T}\sum_{l=1}^{T} W_d^T\left(\frac{t_j-u}{b}\right) W_d^T\left(\frac{t_l-u}{b}\right) X(t_l)X(t_j)\,du \\
&\quad + d!^2\left(\sigma_\varepsilon^2-\hat{\sigma}_\varepsilon^2\right)\int_0^1 \sum_{j=1}^{T} W_d^T\left(\frac{t_j-u}{b}\right)^2 du \\
&= \left\{ d!^2 \int_0^1\int_0^1\int_0^1 \frac{f(x)f(y)}{b^{2(d+1)}f(z)^2} K_{d,\rho}^*\left(\frac{x-z}{b}\right) K_{d,\rho}^*\left(\frac{y-z}{b}\right) X(x)X(y)\,dx\,dy\,dz \right. \\
&\quad \left. + \mathcal{O}_P\left(\frac{1}{T^{3/2}b^{2d+1}}\right)\right\}\{1+\mathcal{O}_P(1)\} \\
&= \left\{ \int_0^1 X^{(d)}(z)X^{(d)}(z)\,dz \right. \\
&\quad + 2\frac{d!}{(\rho+1)!}\int_0^1 \frac{b^{\rho+1}}{b^d}\left(\int_0^1 u^{\rho+1}K_{d,\rho}^*(u)\,du\right) X^{(\rho+1)}(z)X^{(d)}(z)\,dz \\
&\quad + \frac{d!^2}{(\rho+1)!^2}\int_0^1 \frac{b^{2\rho+2}}{b^{2d}}\left(\int_0^1 u^{\rho+1}K_{d,\rho}^*(u)\,du\right)^2 X^{(\rho+1)}(z)X^{(\rho+1)}(z)\,dz \\
&\quad \left. + \mathcal{O}_P\left(\frac{1}{T^{3/2}b^{2d+1}}\right)\right\}\{1+\mathcal{O}_P(1)\}
\end{aligned} \tag{51}
$$

These results where obtained by substitution with $x = z+ub$, $y = z+vb$ and using a $\rho+1$ order Taylor expansion of $X(z+ub)$ and $X(z+vb)$ together with (49). We get $\int_{[0,1]^g} X(u)^2\,du - \mathsf{E}(\theta_{d,\rho}|t_1,\ldots,t_T) = \mathcal{O}_p\left(b^{\rho+1-d} + \left(T^{3/2}b^{2d+1}\right)^{-1}\right)$.

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

$$
\int f(z)^{-2} K_{d,\rho}^*\left(\frac{y-z}{b}\right) K_{d,\rho}^*\left(\frac{x-z}{b}\right) dz = f(c)^{-2}\int K_{d,\rho}^*\left(\frac{y-z}{b}\right) K_{d,\rho}^*\left(\frac{x-z}{b}\right) dz. \tag{52}
$$

We introduce a kernel convolution with

$$
K_{d,\rho}^C(y-x) \stackrel{\text{def}}{=} \int K_{d,\rho}^*(y-z)\,K_{d,\rho}^*(x-z)\,dz \tag{53}
$$

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

$$
K_{d,\rho}^C\left(\frac{y-x}{b}\right) = \int K_{d,\rho}^*\left(\frac{y}{b}-z\right) K_{d,\rho}^*\left(\frac{x}{b}-z\right) dz = \int b^{-1} K_{d,\rho}^*\left(\frac{y-u}{b}\right) K_{d,\rho}^*\left(\frac{x-u}{b}\right) du. \tag{54}
$$

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

$$\int\int_D K^C_{d,\rho}\left(\frac{y-x}{b}\right)dydx=\frac{b}{2}\int_0^2\int_{-1}^1 K^C_{d,\rho}\left(\frac{v+u-u+v}{2}\right)dvdu=b\int K^C_{d,\rho}(v)\,dv. \tag{55}$$

Note that the variance can be decomposed

$$\mathrm{Var}\left(\theta_{d,\rho}|t_1,\ldots,t_T\right) \tag{56}$$

$$\begin{aligned}
&=\frac{d!^4}{T^4(b^{4d+2})f(c)^4}\left\{\sum_{l=1}^T K^C_{d,\rho}(0)^2\mathrm{Var}(Y(t_l)^2)\right. && (57)\\
&+2\sum_{l=1}^T\sum_{k\neq l}^T \mathrm{Var}(K^C_{d,\rho}\left(\frac{t_l-t_k}{b}\right)Y(t_l)Y(t_k)) && (58)\\
&+4\sum_{l=1}^T\sum_{k\neq l}^T\sum_{k'\neq k}^T \mathrm{Cov}(K^C_{d,\rho}\left(\frac{t_k-t_l}{b}\right)Y(t_k)Y(t_l),K^C_{d,\rho}\left(\frac{t_l-t_{k'}}{b}\right)Y(t_l)Y(t_{k'})) && (59)\\
&\left.+24\sum_{l=1}^T\sum_{k\neq l}^T\sum_{k'\neq k}^T\sum_{l'\neq k'}^T \mathrm{Cov}(K^C_{d,\rho}\left(\frac{t_l-t_k}{b}\right)Y(t_l)Y(t_k),K^C_{d,\rho}\left(\frac{t_{l'}-t_{k'}}{b}\right)Y(t_{l'})Y(t_{k'}))\right\} && (60)\\
&+\mathcal{O}_P\left(\frac{1}{T}\right). && (61)
\end{aligned}$$

Expression (60) vanishes and (57) given by $\frac{d!^4}{T^3(b^{4d+2})f(c)^4}\int K^C_{d,\rho}(0)^2\mathrm{Var}(Y(y)^2)f(y)dy\{1+\mathcal{O}_P(T^{-1})\}$ is dominated by (58) because

$$\begin{aligned}
&\frac{2d!^4}{T^4(b^{4d+2})f(c)^4}\sum_{l=1}^T\sum_{k\neq l}^T K^C_{d,\rho}\left(\frac{t_l-t_k}{b}\right)^2\mathrm{Var}(Y(t_l)Y(t_k))\\
&=\frac{2d!^4}{T^4(b^{4d+2})f(c)^4}\sum_{l=1}^T\sum_{k\neq l}^T K^C_{d,\rho}\left(\frac{t_l-t_k}{b}\right)^2\left\{\mathsf{E}(Y(t_l)^2Y(t_k)^2)-\mathsf{E}(Y(t_l)Y(t_k))^2\right\} && (62)\\
&=\frac{2d!^4\int(\sigma_\epsilon^4+2\sigma_\epsilon^2X(x)^2)f(x)^2dx}{T^2b^{4d+1}f(c)^4}\int\left(K^C_{d,\rho}(u)\right)^2du+\mathcal{O}_P\left(\frac{1}{T^2b^{4d+1}}\right).
\end{aligned}$$

Before looking at expression (59), note that with $m\geq 2d$

$$\begin{aligned}
&\int\int\frac{d!^2}{b^{2d+1}}K^C_{d,\rho}\left(\frac{x-y}{b}\right)X(x)dxdy\\
&=\frac{d!^2}{b^{2d}}\int\int\int K^*_{d,\rho}(m)\,K^*_{d,\rho}(z)\,X\{y+(m-z)b\}\,dzdmdy && (63)\\
&=(-1)^d\int_0^1 X^{(2d)}(y)dy+\mathcal{O}_P(1)
\end{aligned}$$

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

We can thus derive for expression (59) that

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

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

2. Multivariate case ($g>1$) — The same strategy also works in the multivariate case by using multivariate Taylor series. Using the multi-index notation introduces in section 2.4 and $a=(a_1,...,a_g)$, $a_l\in\mathbb{N}^+$ a multivariate taylor series of degree $k<\rho$ is given by

$$
X(x-u\circ b)=\sum_{0\leq|a|\leq k}\frac{X^{(a)}(x)}{a!}(u\circ b)^a+\mathscr{O}_P\left(u^{k+1}\max(b)^{k+1}\right). \tag{64}
$$

Using the equivalent kernel by Ruppert and Wand (1994) extended to the case and using Masry (1996) we can further state that with a bandwidth fulfilling $\frac{log(T)}{Tb_1\times\ldots\times b_g}\to 0$ we have uniformly in $u\in[0,1]^g$ that $S_T(u)^{-1}\to\frac{S(u)^{-1}}{f(u)}$ almost surely as $T\to\infty$. Furthermore, the multivariate equivalent kernel has the properties that with $\nu=(\nu_1,\ldots,\nu_g)$, $\nu_l\in\mathbb{N}^+$

$$
\int u^\nu K_{d,\rho}^*(u)\,du=\delta_{d,\nu},\ |\nu|\leq\rho,\ 0\leq d_i\ \forall i=1,\ldots g. \tag{65}
$$

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

$$
\begin{aligned}
&\mathsf{E}\Big(\theta_{d,\rho}|t_1,\ldots,t_T\Big)\\
&=\Bigg\{\int_{[0,1]^g}X^{(d)}(z)X^{(d)}(z)dz\\
&\quad+2\frac{d!}{(\rho+1)!}\int_{[0,1]^g}\frac{\max(b)^{\rho+1}}{b^d}\left(\int u^{\tilde{\rho}}K_{d,\rho}^*(u)\,du\right)X^{(\tilde{\rho})}(z)X^{(d)}(z)dz\\
&\quad+\mathscr{O}_P\left(\frac{\max(b)^{\rho+1}}{b^d}+\frac{1}{T^{3/2}(b^{2d}b_1\times\ldots\times b_g)}\right)\Bigg\}\{1+\mathscr{O}_P(1)\}
\end{aligned} \tag{66}
$$

Further note that for the convoluted kernel we get

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

Accordingly, we get for the multivariate equivalent of expression (58) that

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

and because we assume that $m\geq 2|d|$ we get for the multivariate equivalent of expression (59) that

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

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

5.3 Proof of Proposition 2.2

1. Asymptotic results— 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.

$$
\begin{aligned}
M_{ij}^{(0)}-\tilde{M}_{ij}^{(0)}&=\int_{[0,1]^g}X_i(t)X_j(t)dt-\frac{1}{T}\sum_{l=1}^{T}Y_i(t_{il})Y_j(t_{jl})+I(i=j)\hat{\sigma}_{i\varepsilon}^2\\
&=\int_{[0,1]^g}X_i(t)X_j(t)dt-\frac{1}{T}\sum_{l=1}^{T}\left(X_i(t_l)+\varepsilon_{il}\right)\left(X_j(t_l)+\varepsilon_{jl}\right)+I(i=j)\hat{\sigma}_{i\varepsilon}^2\\
&=\int_{[0,1]^g}X_i(t)X_j(t)dt-\frac{1}{T}\sum_{l=1}^{T}X_i(t_l)X_j(t_l)\\
&\quad-\frac{1}{T}\sum_{l=1}^{T}X_i(t_l)\varepsilon_{jl}-\frac{1}{T}\sum_{l=1}^{T}X_j(t_l)\varepsilon_{il}-\frac{1}{T}\sum_{l=1}^{T_i}\varepsilon_{il}\varepsilon_{jl}+I(i=j)\hat{\sigma}_{i\varepsilon}^2.
\end{aligned}
\tag{67}
$$

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

normal, which gives an error of order $T^{-1/2}$ regardless of dimension g. By requiring that $\hat{\sigma}_{i\varepsilon}$ is also $T^{-1/2}$ consistent we get $T^{-1/2}$ for all elements.

To understand $\hat{M}^{(0)}$ we investigate two possible sources of error in the construction of the estimator. One coming from interpolation and smoothing at a common grid and the other from approximating the integral with a sum. First note that by the same arguments as for $\tilde{M}^{(0)}$ the error of the integral approximation is of order $T^{-1/2}$. Besides the error for the off diagonal elements is smaller than for the diagonal, thus the leading error source is given by Lemma 2.1. The same arguments also work to derive asymptotic results for $\hat{M}^{(d)}$.

5.4 Proof of Proposition 2.4

Under the assumptions of Proposition 2.4 together with the requirements of Lemma 2.2 for $\nu = (0,d)$ and the setup of Remark 2.3

$$||\hat{M}^{(\nu)} - M^{(\nu)}|| \leq \operatorname{tr}\left\{\left(\hat{M}^{(\nu)} - M^{(\nu)}\right)^{\top}\left(\hat{M}^{(\nu)} - M^{(\nu)}\right)\right\}^{1/2} = \mathcal{O}_p\left(NT^{-1/2}\right). \tag{68}$$

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

$$\mathsf{E}\left[\left(p_r^{(\nu)}\right)^{\top}(\hat{M}^{(\nu)} - M^{(\nu)})p_r^{(\nu)}\right]^2 = \mathcal{O}_p\left(\frac{N}{T}\right) \tag{69}$$

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

$$\begin{aligned}
&\left|\frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N} p_{ir}^{(\nu)} X_i^{(d)}(t) - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\sum_{i=1}^{N}\hat{p}_{ir}^{(\nu)}\hat{X}_{i,h}^{(d)}(t)\right| \\
&\leq \left|\frac{1}{\sqrt{l_r^{(\nu)}}}\sum_{i=1}^{N} p_{ir}^{(\nu)}\left[X_i^{(d)}(t) - \hat{X}_{i,h}^{(d)}(t)\right]\right| + \left|\sum_{i=1}^{N}\left(\frac{1}{\sqrt{l_r^{(\nu)}}}p_{ir}^{(\nu)} - \frac{1}{\sqrt{\hat{l}_r^{(\nu)}}}\hat{p}_{ir}^{(\nu)}\right)\hat{X}_{i,h}^{(d)}(t)\right|.
\end{aligned} \tag{70}$$

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 Kneip and Utikal (2001) together with equation (69) gives

$$l_r^{(\nu)} - \hat{l}_r^{(\nu)} = (p_r^{(\nu)})^T(\hat{M}^{(\nu)} - M^{(\nu)})p_r^{(\nu)}) + \mathcal{O}_p(NT^{-1}) = \mathcal{O}_p(N^{1/2}T^{-1/2} + NT^{-1}), \tag{71}$$

where

$$\frac{1}{\sqrt{\hat{l}_r^{(\nu)}}} - \frac{1}{\sqrt{l_r^{(\nu)}}} = \frac{l_r^{(\nu)} - \hat{l}_r^{(\nu)}}{\sqrt{\hat{l}_r^{(\nu)}}\sqrt{l_r^{(\nu)}}(\sqrt{\hat{l}_r^{(\nu)}} + \sqrt{l_r^{(\nu)}})} = \mathcal{O}_p\left(T^{-1/2}N^{-1} + T^{-1}N^{-1/2}\right). \tag{72}$$

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

$$|\hat{p}_{ir}^{(\nu)} - p_{ir}^{(\nu)}| = \mathcal{O}_p\left((NT)^{-1/2}\right) \text{ and } ||\hat{p}_r^{(\nu)} - p_r^{(\nu)}|| = \mathcal{O}_p\left(T^{-1/2}\right). \tag{73}$$

Putting all results together for the second term gives

$$
\begin{aligned}
&\left|\sum_{i=1}^{N}\left(\frac{1}{\sqrt{l_r^{(v)}}}p_{ir}^{(v)}-\frac{1}{\sqrt{\hat{l}_r^{(v)}}}\hat{p}_{ir}^{(v)}\right)\hat{X}_{i,h}^{(d)}(t)\right| = \\
&=\left|\sum_{i=1}^{N}\left(\frac{1}{\sqrt{l_r^{(v)}}}-\frac{1}{\sqrt{\hat{l}_r^{(v)}}}\right)\hat{p}_{ir}^{(v)}\hat{X}_{i,h}^{(d)}(t)+\frac{1}{\sqrt{l_r^{(v)}}}\sum_{i=1}^{N}\left(\hat{p}_{ir}^{(v)}-p_{ir}^{(v)}\right)\hat{X}_{i,h}^{(d)}(t)\right| \\
&\leq\left|\left(\frac{1}{\sqrt{l_r^{(v)}}}-\frac{1}{\sqrt{\hat{l}_r^{(v)}}}\right)\right|\sum_{i=1}^{N}|p_{ir}^{(v)}|\left|\left(\hat{X}_{i,h}^{(d)}(t)\right)\right| \\
&\quad+\left|\left(\frac{1}{\sqrt{l_r^{(v)}}}-\frac{1}{\sqrt{\hat{l}_r^{(v)}}}\right)\right||\hat{p}_r^{(v)}-p_r^{(v)}||\left|\hat{X}_{i,h}^{(d)}(t)\right|+\frac{1}{\sqrt{l_r^{(v)}}}||\hat{p}_r^{(v)}-p_r^{(v)}||\left|\hat{X}_{i,h}^{(d)}(t)\right| \\
&=\mathcal{O}_p\left((NT)^{-1/2}\right)\left|\hat{X}_{i,h}^{(d)}(t)-X_{i,h}^{(d)}(t)+X_{i,h}^{(d)}(t)\right| \\
&\leq\mathcal{O}_p\left((NT)^{-1/2}\right)(\text{Bias}\left(\hat{X}_{j,h}^{(d)}(t)\right)+\sqrt{\text{Var}\left(\hat{X}_{j,h}^{(d)}(t)\right)}+\left|X_{i,h}^{(d)}(t)\right|).
\end{aligned}
\tag{74}
$$

Using Cauchy-Schwarz and equation (72) 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 (73) this also leads to order $(NT)^{-1/2}$. Inserting the right hand side, equation (70) becomes

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

5.5 Proof of Proposition 2.5

Note that

$$
\sqrt{l_r^{(v)}}-\sqrt{\hat{l}_r^{(v)}}=(l_r^{(v)}-\hat{l}_r^{(v)})(\sqrt{l_r^{(v)}}+\sqrt{\hat{l}_r^{(v)}})^{-1}=\mathcal{O}_p(T^{-1/2}+N^{1/2}T^{-1}),
\tag{75}
$$

together with (73) gives

$$
\begin{aligned}
&\hat{\delta}_{ir}-\hat{\delta}_{ir,T}=\sqrt{l_r^{(v)}}p_{ir}^{(v)}-\sqrt{\hat{l}_r^{(v)}}\hat{p}_{ir}^{(v)} \\
&=\left(\sqrt{l_r^{(v)}}-\sqrt{\hat{l}_r^{(v)}}\right)p_{ir}^{(v)}-\sqrt{\hat{l}_r^{(v)}}\left(\hat{p}_{ir}^{(v)}-p_{ir}^{(v)}\right)=\mathcal{O}_p(T^{-1/2}+N^{1/2}T^{-1}).
\end{aligned}
\tag{76}
$$

Using Proposition 2.4 it follows that

$$
\begin{aligned}
&|Y_i(t)-\hat{Y}_i(t)|=|\sum_{r=1}^{K}\hat{\delta}_{ir}\hat{\gamma}_r^{(v)}(t)-\sum_{r=1}^{K}\hat{\delta}_{ir,T}\hat{\gamma}_{r,T}^{(v)}(t)| \\
&=|\sum_{r=1}^{K}(\hat{\delta}_{ir}-\hat{\delta}_{ir,T})\hat{\gamma}_r+\hat{\delta}_{ir,T}(\hat{\gamma}_r-\hat{\gamma}_{r,T})| \\
&=\mathcal{O}_p\left(T^{-1/2}+N^{1/2}T^{-1}+\max(h)^{p+1}h^{-d}+(NTh_1\times\ldots\times h_gh^{2d})^{-1/2}\right).
\end{aligned}
\tag{77}
$$

References

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

B. Bahra. Implied risk-neutral probability density functions from option prices : theory and application. 1997.

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

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

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

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

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

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

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

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

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

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

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

F. Comte and E. Renault. Long memory in continuous-time stochastic volatility models. *Mathematical Finance*, 8(4):291–323, 1998.

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

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

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

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

J. Du and N. Kapadia. Tail and volatility indices from option prices. Working paper, 2012.

J. C. Duan and C. Y. Yeh. Jump and volatility risk premiums implied by VIX. *Journal of Economic Dynamics and Control*, 34(11):2232–2244, 2010.

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

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Chapman and Hall, London, 1996.

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

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

B. Feunou and R. Tédongap. A Stochastic Volatility Model With Conditional Skewness. *Journal of Business & Economic Statistics*, 30(4):576–591, 2012.

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

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

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

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

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

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

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

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

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

D. Huang and I. Shaliastovich. Volatility-of-Volatility Risk. *University of Pennsylvania Preprint*, 2014.

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

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

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

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

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

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

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

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

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

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

W. K. Newey and K. D. West. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica (1986-1998)*, 55(3): 703–708, 1987.

Y. H. Park. Volatility-of-volatility and tail risk hedging returns. *Journal of Financial Markets*, 26:38–63, 2015.

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

M. Rubinstein. Nonparametric Tests of Alternative Option Pricing Models Using All Reported Trades and Quotes on the 30 Most Active CBOE Option Classes from August 23, 1976 through August 31, 1978. *The Journal of Finance*, 40(2):455–480, 1985.

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

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

A. van Bömmel, S. Song, P. Majer, P. N. C. Mohr, H. R. Heekeren, and W. K. Härdle. Risk Patterns and Correlated Brain Activities. Multidimensional Statistical Analysis of fMRI Data in Economic Decision Making Study. *Psychometrika*, 79(3):489–514, 2014.

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

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

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