Variational Bayesian functional PCA

Angelika van der Linde

FB03: Institute of Statistics, University of Bremen, PO Box 330440, 28334 Bremen, Germany

Abstract

A Bayesian approach to analyze the modes of variation in a set of curves is suggested. It is based on a generative model thus allowing for noisy and sparse observations of curves. A Demmler-Reinsch(-type) basis is used to enforce smoothness of the latent ('eigen'-)functions. Inference, including estimation, error assessment and model choice, particularly the choice of the number of eigenfunctions and their degree of smoothness, is derived from a variational approximation of the posterior distribution. The proposed analysis is illustrated with simulated and real data.

Key words: Variational principal components, functional data analysis, eigenfunctions, rotation, interpolation splines, Demmler-Reinsch basis, Canadian weather data.

1 Introduction

Functional Principal Component Analysis (FPCA) is Principal Component Analysis (PCA) applied to functions instead of vectors, but functions differ from vectors by continuity or smoothness. Hence some of the objectives of FPCA when applied to a bundle of curves are similar to those of classical PCA: to describe the set of curves by a characteristic average curve and patterns of variability, to find a low dimensional representation of individual curves, to relate their shape to covariates or to compare different sets of curves (see for example (Nerini and Ghattas, 2007)). If curves are only partially observed, their low dimensional representation amounts to a reconstruction of the whole curve, a problem that is specific for functions.

Email address: avdl@math.uni-bremen.de (Angelika van der Linde).
URL: www.math.uni-bremen.de/ avdl (Angelika van der Linde).

In traditional multivariate analysis principal components (PCs) are defined as linear transformations of the observed random vector, obtained by forming inner products with eigenvectors of the covariance matrix. This approach is generalized in most of frequentist FPCA, substituting the Euclidean space by an inner product function space, often L_2, covariance matrices by covariance operators and eigenvectors by eigenfunctions According to the choice of the function space and particularly the inner product different versions of FPCA result (Besse, 1988; Silverman, 1996; Ocaña, Aguilera and Valderrama, 1999; Ocaña, Aguilera and Escabias, 2007; Manté, Yao and Degiovanni, 2007).

To date there is a vast literature on functional data analysis (FDA) and in particular FPCA. The monographs by Ramsay and Silverman (2005, 2002) established FDA as a topic on its own in statistics and demonstrated its potential in many applications. The recent monograph by Ferraty and Vieu (2006) provides a complementary view on FDA emphasizing a nonparametric approach. The paper by Müller (2005) on FPCA and functional regression offers a review as well as an introduction to the ongoing discussion. Also, special issues of several journals indicate that there is growing interest in the field of FDA (Davidian et al., 2004; Valderrama, 2007; Manteiga and Vieu, 2007). Applied studies using FPCA have proven the versatility and usefulness of the methods. In their book Ramsay and Silverman (2002) present a collection of case studies from the fields of growth analysis, criminology, meteorology and others. Biomedical data have been analyzed in the papers by Grambsch et al. (1995), Zhao et al. (2004), Behseta et al. (2005), Müller and Wang (2005) or Hyndman and Ullah (2007). Economical applications of FPCA were considered for example by Aguilera et al. (1999a) or Kneip and Utikal (2001). A geophysical application was presented by Manteiga and Vieu (2007).

As technique for dimension reduction FPCA has been applied as part of a more comprehensive analysis. Particularly Principal Component Regression was extended to functional regression (with scalar, for instance binary response and functional predictors) to cope with multicollinearity among regressors (Escabias, Aguilera and Valderrama, 2004, 2005; Aguilera, Escabias and Valderrama, 2006). Functional regression with PCA applied to functional predictors was also studied by Cardot et al. (2003) and compared (based on simulations and asymptotics) to more direct (penalized likelihood) estimation of the parameter functions The technique was again used by Müller and Stadtmüller (2005) and Leng and Müller (2006). The approach was however critisized on the same grounds as finite dimensional PCA in regression (Cardot et al., 2003; Saporta et al., 2007; Escabias et al., 2007): the dimension reduction is obtained independently of the response variable, such that the resulting functional predictors are not necessarily optimal for explaining the response. In an approach related to functional principal component regression techniques for forecasting were elaborated in a series of papers (Aguilera, Ocaña and Valderrama, 1997, 1999b,c; Valderrama, Ocaña and Aguilera, 2002;

Aguilera, Escabias and Valderrama, 2008). FPCA can further be used for purposes that multivariate PCA is not suitable for. For example, FPCA has been successfully employed to model and estimate features of stochastic processes (Bouzas, Valderrama, Aguilera and Ruiz-Fuentes, 2006; Bouzas, Ruiz-Fuentes and Ocaña, 2007; Fernández-Alcalá, Navarro-Moreno and Ruiz-Molina, 2007). Chiou and Müller (2007) use FPCA scores for a residual analysis in functional regression.

FPCA like PCA is often used as an exploratory technique and functional estimates are displayed without error bounds. From a frequentist point of view statistical inference and error assessments are based on asymptotic theory or bootstrapping methods. Asymptotic theory in FPCA mainly addresses consistency of estimates for the covariance (operator), eigenfunctions and eigenvalues and the asymptotic distribution of eigenvalues and PCs. Asymptotic results are obtained under varying assumptions, in particular with respect to the choice of the function space and the sampling scheme (for example, (Dauxois et al., 1982) for fully observed curves in L_2 or (Müller and Wang, 2005) and (Hall et al., 2006) for discretized curves and curves possibly observed at irregularly spaced sparse points). Related to the specification of the function space is the choice of the smoothing method applied to fit observed functions or eigenfunctions like the use of roughness penalties (Pezzulli and Silverman, 1993; Silverman, 1995; Yao and Lee, 2006), kernel methods (Boente and Fraiman, 2000) or local weighted least squares (Hall et al., 2006). Less attention has been paid to the estimation of the mean function (average curve) as nuisance parameter in frequentist FPCA and to the reconstruction of the observed functions based on estimated eigenfunctions. Only recently, when FPCA was extended to curves observed at irregular designs, this has come into the focus of asymptotic theory as well (Müller and Wang, 2005). Alternatively bootstrap methods have been applied to obtain confidence intervals for reconstructed curves (James et al., 2000).

The Bayesian approach provides a principled alternative to asymptotic inference. Considering observations and parameters as dual, additional information is not drawn from infinitely many hypothetical observations but from a prior distribution on the parameter space. Inference can then be based on the posterior distribution of the parameters conditional on the data The Bayesian approach to FPCA elaborated in this paper yields an easy to handle approximate posterior distribution which allows for simple and fast error assessments for all quantities of interest. This holds for discretized curves on a common regular grid as well as for curves observed at different designs with irregularly spaced sparse points.

Bayesian contributions to FDA and especially FPCA are rare (Behseta et al., 2005). There are mainly two reasons why Bayesians tend to be reluctant about FDA and FPCA: one concern is about the function spaces occurring as sample

space or as parameter space in genuine FDA. There is hardly any choice for densities on function spaces to be used as likelihood or prior. A probability distribution on function space usually is conceptualized as stochastic process implying finite dimensional margins for a discretization of functions. In practice working with a discretization may be sufficient, theoretically it is less ambitious. Behseta et al. (2005) present an approach referring to a stochastic process. In this paper a finite dimensional discretization with a stochastic process in the background will be employed.
A second reason lies in the tradition of perceiving (and teaching) PCA for random vectors according to the 'representative' approach, where the PCs are (linear) transformations of the observed random vector. The stochastically independent PCs thus 'represent' the original variables. This is the prevailing approach which is generalized in most of frequentist FPCA. To Bayesians this view implies a rather inconvenient parameterization of the problem. Much more appealing is the (geometrically) 'generative' approach based on the optimality property of PCA that the (standardized) eigenvectors of the covariance matrix form (orthonormal) basis vectors of low dimensional best approximating subspaces for the observed data vectors. Related is the (stochastically) generative approach based on the optimality property of PCs as random variables to best predict the observed random variables componentwise.
From a frequentist perspective the generative approach was elaborated by a French group of statisticians in Toulouse (Besse, 1994; Besse et al., 1997; Cardot, 2000). Although it was observed very early (Reinsel and Velu, 1998, ch.7) that PCs can be obtained from a generative model performing a Factor Analysis with isotropic noise and Bayesian Factor Analysis had been around for a long time (see (Rowe, 2003) and the references therein), it was the revival of this approach by Tipping and Bishop (1999) under the name of 'probabilistic PCA' that enhanced Bayesian PCA in recent years (Bishop, 1999a, 2006; Minka, 2001; Zhang et al., 2004). Šmídl and Quinn (2007) provide a review while presenting their own version.

Given the generative model key inferential issues in Bayesian PCA are:

- A factor analytic model implies an indeterminacy of factors up to a nonsingular transformation or at least up to sign and rotation. In Factor Analysis therefore often a rotation of estimated factors is carried out to improve interpretability of the factors. In classical PCA orthonormal eigenvectors span the best approximating subspaces, hence it is natural to impose orthogonality constraints in the generative model (Minka, 2001; Šmídl and Quinn, 2007).

- Factor Analysis and thus probabilistic PCA is based on a generative latent variable model. Factors are not initially assumed to be transformations of observed variables. Generative models can be easily generalized to allow for non-Gaussian distributional assumptions for the latent variables as exemplified by the machine learning community in the development of Independent

Component Analysis, respectively Independent Factor Analysis. See (Roberts and Everson, 2001) or (Hyvärinen et al., 2001).

- Bayesian inference in Independent Component Analysis or Independent Factor Analysis is often based on variational methods or 'ensemble learning' to derive an approximate posterior distribution in closed form (Attias, 1998; Choudrey and Roberts, 2001). Variational inference was also elaborated for probabilistic PCA without orthogonality constraints by Bishop (1999b) and with orthogonality constraints by Šmídl and Quinn (2007).

- The choice of the number of PCs or factors is an important subproblem in PCA or Factor Analysis. The variational approach includes options to decide on the number of factors (Bishop (1999b); Bishop (2006), ch.10.1.4). A more sophisticated approach with a prior on the number of factors requires inference based on sampling methods (Lopes and West, 2004; Zhang et al., 2004).

Putting all these results together it seems promising to try a Bayesian FPCA for discretized functions, possibly observed with noise, as a probabilistic PCA incorporating a smoothness constraint in order to find smooth eigenfunctions to span a low dimensional subspace for the original functions. In frequentist FDA usually smoothness is enforced introducing roughness penalties (Ramsay and Silverman, 2005; Besse, 1994). A roughness penalty is equivalent to a smoothness prior and is an obvious option in a Bayesian analysis. Results obtained from an implementation of this idea will be discussed. An alternative to a smoothness prior is the choice of a basis of smooth functions like splines in which the eigenfunctions are represented. With a common basis also observations of functions at different possibly irregularly spaced points can be handled. This advantage motivated the generative model of James et al. (2000), which is closest to the approach pursued here, though their inference is frequentist.

Most often, because of their flexibility, free knot B-splines are chosen as common basis functions. The knots then act as smoothing parameters the estimation of which in a Bayesian approach requires sampling techniques. Hence the speed and ease of implementation in closed form of the variational algorithm get lost. In order to reduce the number of smoothing parameters and to retain the applicability of a variational algorithm including its device for a quick assessment of models with various numbers of eigenfunctions a different spline basis is suggested. All functions are represented as interpolation splines on a fine grid of points (i.e. many, but known knots), and the Demmler-Reinsch basis is chosen to build up interpolation splines. This is most parsimonious as the Demmler-Reinsch basis itself results from a PCA for interpolation splines (van der Linde, 2003). Such a basis function is the more important in reconstructing an interpolation spline the rougher it is. Therefore, this basis is very effective in filtering noise, and the degree of smoothness of the eigenfunctions can be

controlled by just one smoothness parameter, the number of Demmler-Reinsch basis functions. Enforcing smoothness technically conflicts with orthogonality constraints imposed by a prior. Therefore, the selection of a rotation is not incorporated into the derivation of the (approximate) posterior distribution, but based on that distribution.

In summary, the proposed Bayesian FPCA essentially combines a generative model like that considered by James et al. (2000) with a variational algorithm like the one outlined by Bishop (1999b). To achieve this, a special basis of smooth functions is selected. The corresponding variational algorithm is partially newly derived, and finally a rotation of the resulting eigenfunctions is suggested. The approach is applied to simulated data to illustrate its performance and to real data for comparison to previous analyses in the literature emphasizing the estimation of the (smooth) mean function and the error assessment for reconstructed curves.

In the sequel the paper is organized as follows: In section 2 a motivating example with simulated curves is described illustrating the goals and problems in FPCA. Next the model for FPCA is specified and the variational algorithm for inference is outlined in section 3. In section 4 the motivating example is resumed. Also, another example with a real data set is investigated. A concluding discussion is given in section 5.

2 Description of simulated data and preliminary analyses

Consider the mixture of two Gaussian densities,

$$g(t) = 0.5p_{N(0,1)} + 0.5p_{N(6,4)}. \tag{1}$$

To obtain a single curve a sample of 30 values t_j was generated from this mixture distribution, and a kernel estimate

$$\widehat{g}(t) = \frac{1}{30}\sum_{j=1}^{30} k(\frac{t_j - t}{h}) \tag{2}$$

was formed with a standard Gaussian kernel k and bandwidth $h = 1.06\widehat{\sigma}_t * 30^{-0.2}$, where $\widehat{\sigma}_t$ denotes the empirical standard deviation in the sample. Although this bandwidth is asymptotically optimal, the kernel estimate tends to undersmooth the first peak of g at zero.

100 curves were generated in this way. A discretization, a vector of function values, is obtained for each curve on a regular grid with 50 points in $(-7, 14)$.

Each of the 10 curves displayed in fig.1 is a linear interpolation of 50 function values. Several modes of variation show up already in this small data set: (i) the (a)symmetry of the peaks, (ii) the location of the peaks (and troughs), (iii) the separation of the two peaks.

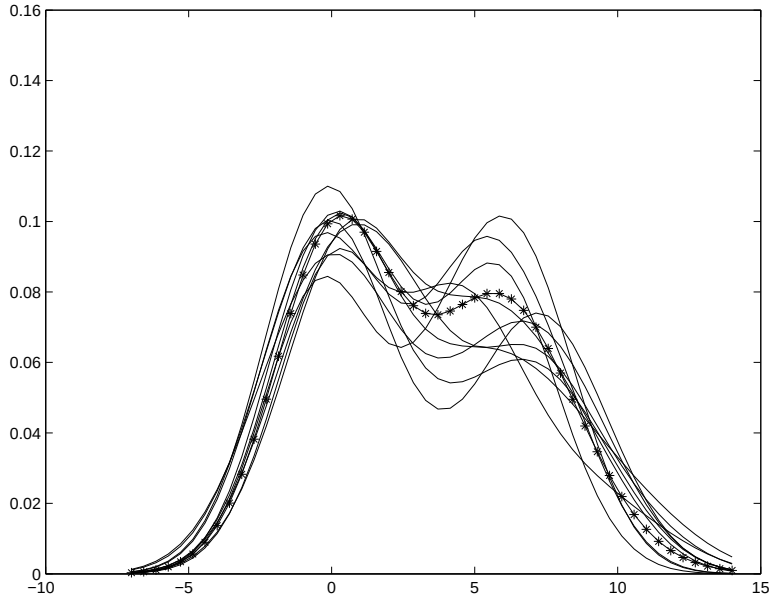

Fig. 1. Ten kernel estimates of g and mean curve (stars) based on 100 estimates.

A conventional PCA of these 100 50-dimensional vectors reveals that more than 99% of variation can be explained by six eigenvectors of the empirical covariance matrix. The corresponding vector of (cumulated) explained variances is (0.62, 0.87, 0.94, 0.98, 0.993, 0.998). The eigenvectors again can be displayed as eigenfunctions interpolating linearly the given 50 values. One way to interpret them is to look at the deviation from the mean function as shown in fig.2 (first column).

This suggests skewness of the curve as the most important mode of variation, separation of the peaks (depth of trough) as next important and the location of the peaks (shift of the whole curve) as a third feature. The next three modes of variation are induced by more oscillating eigenfunctions, and are harder to interpret. Plots based on the distribution of scores for each PC as suggested by Jones and Rice (1992) may also be helpful for the interpretation of the eigenfunctions as illustrated in the second column of fig.2.

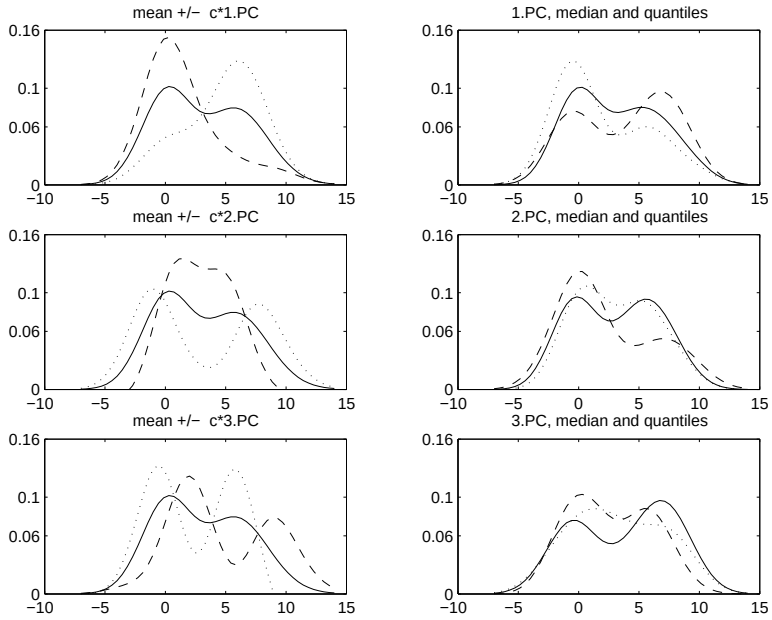

Fig. 2. First column: Empirical mean function (solid line) plus (dashed line) and minus (dotted line) $c = 0.2$ times an eigenfunction. Second column: Curves corresponding to the 10%-quantile (dashed), median (solid) and 90%-quantile (dotted) of the distribution of scores.

If the curves were observed without noise such a direct discrete PCA could be sufficient. Here, however, we are interested in a set-up where the observation of functions is corrupted by noise and each curve may only be observed on a part of the interval. In this case the PCA above describes what is to be recovered.

Noisy discretized curves were obtained adding independent random errors generated from a N(0,0.0001) distribution to the 50 function values of each curve. A conventional PCA of the 100 50-dimensional vectors of noisy function values now yields rather noisy eigenfunctions as shown in fig.3. and hence noisy reconstructions of the curves.

Also, the first six PCs now explain less than 70% of the variation in the set of noisy curves, and the increase in explanatory potential with the number of PCs is slow. A requirement of smoothness of the eigenfunctions helps to filter the noise and aims at the reconstruction of the original smooth functions. Preliminary individual smoothing of curves and application of a standard PCA to smoothed curves might be an option if the curves differ considerably in curvature. If the curves are homogeneously smooth it is more efficient to incorporate smoothing into PCA (Besse, 1994; Rice, 2004). This is part of the procedure described in the sequel.

Preliminary smoothing of single curves is not possible if the curves are observed at different parts of the interval. In order to generate a set of partial curves a

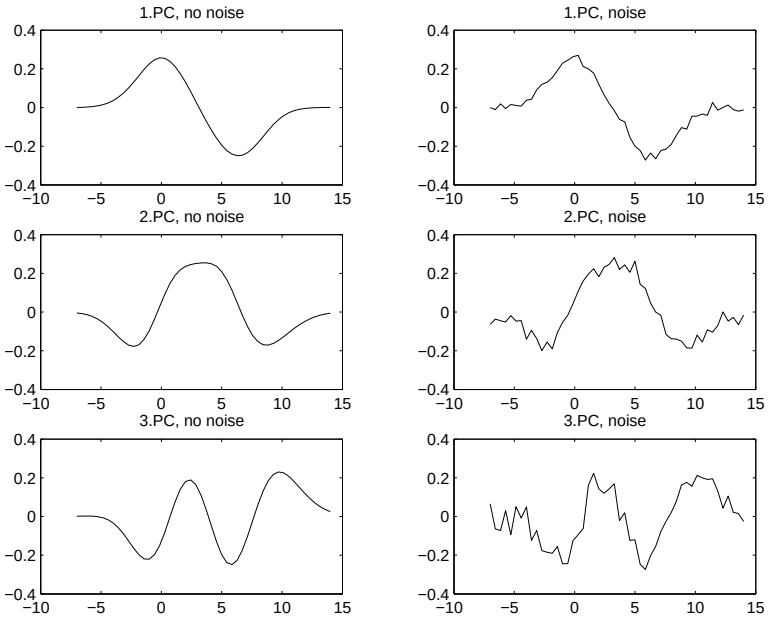

Fig. 3. First column: The first three eigenfunctions from the PCA without noise. Second column: The corresponding eigenfunctions resulting from noisy observations.

partial design for a single curve was constructed selecting a subset of points from the grid of 50 points. To this end a starting point was chosen from a uniform distribution on {1,2,...,49,50}, and a direction (forward/backward) was drawn from a Bernoulli distribution with probability 0.5. Additionally, a random length between 1 and 20 was chosen. Then a set of up to 20 points was finally determined, including the starting point and all points along the given direction with the last point either realizing the prescribed length or being a marginal point (1 or 50) of the grid. One data set of 100 noise free partial curves and another one with 100 noisy partial curves was created in this way. Conventional PCA is no longer applicable, not even to the noise free partial curves.

In section 4 Bayesian estimates of the eigenfunctions and their scores will be presented as well as reconstructed individual curves along with an error assessment in terms of the (approximate) posterior distribution.

3 Model(s) and variational algorithm

Consider real valued functions f_m, $m = 1...M$, on $\mathbb{R}$, and assume f_m to be a thin plate spline of order 2 (see Wahba, 1990, ch.2.4). To discretize, a grid d of N points is chosen, yielding vectors of function values $f_{md} \in \mathbb{R}^N$. Each f_m is approximated by the interpolation spline $h_{I(f_{md})}$, the smoothest function (in the sense of minimizing $\int (h''(t))^2 dt$) among all functions h attaining the

values f_{md} on d. Thus the discretization is equivalent to restricting ourselves to a function space $H_{I(d)}$, say, of interpolation splines on d.

3.1 The generative model

Assume the m-th curve to be observed (with or without noise) on a grid d_m with N_m points. $d_m \subset d$ may occur and does occur in the example but is not required. Denote the corresponding vector of function values by f_{md_m} and the vector of observed, possibly noisy function values by y_m. Hence, without noise, $y_m = f_{md_m}$. From given values f_{md} the values f_{md_m} can be linearly predicted (as values of the interpolation spline) by a matrix $IP_m : f_{md_m} = IP_m f_{md}$. A distribution on $H_{I(d)}$ is induced by a Gaussian distribution $h_d \sim N(c_d, \Sigma_d)$, with $c_d \in \mathbb{R}^N$, $\Sigma_d \in \mathbb{R}^{N\times N}$ positive definite. In a generative model Σ_d is not targeted, but it is postulated that the residual vectors $f_{md} - c_d$ are spanned by $K < N$ latent basic (eigen-)vectors $a_{1d}...a_{Kd} \in \mathbb{R}^N$. These basis vectors induce basis interpolation splines $h_{I(a_{kd})}$, $k = 1...K$, to be addressed as eigenfunctions. For fixed K, with $A_d = (a_{1d}...a_{Kd}) \in \mathbb{R}^{N\times K}$ and s_m denoting a vector of unknown coefficients, the generative model states

$$y_m|c_d, A_d, s_m, \sigma^2 \sim N(IP_m(c_d + A_d s_m), \sigma^2 I_{N_m}), \tag{3}$$

independently for $m = 1...M$.
The key modelling assumption then is that the mean function and the eigenfunctions are parsimoniously smooth. In order to enforce smoothness of $h_{I(c_d)}$ and $h_{I(a_{kd})}$, c_d and a_{kd} are represented in a special basis (matrix) Q_d of dimension r_Q and R_d of dimension r_R respectively, which yield the best approximating subspaces of dimension r_Q (resp. r_R) for interpolation splines (van der Linde, 2003). The more basis functions are included (the larger r_Q or r_R) the smoother the mean function or the eigenfunctions might be, and thus r_Q and r_R act as smoothing parameters. Hence $c_d = Q_d\delta$, $A_d = R_d G$, $G = (\gamma_1...\gamma_K)$, and with $Q_m = IP_m Q_d$, $R_m = IP_m R_d$ the generative model is parameterized as

$$y_m|\delta, G, s_m, \sigma^2 \sim N(Q_m\delta + R_m G s_m, \sigma^2 I_{N_m}), \tag{4}$$

where K is given, and r_Q and r_R are fixed. Technical details on the matrices IP_m, Q_d and R_d can be derived from (Wahba, 1990, ch.2.4) and (van der Linde, 2003).
Note that if the functions are observed with noise, the error $\varepsilon_m \sim N(0, \sigma^2 I_{N_m})$ comprises an approximation error due to reconstructing f_{md} with $K < N$ and an observational error e_m according to $y_m = f_{md_m} + e_m$, $e_m \sim N(0, \sigma_e^2 I_{N_m})$, say. The observational error is to be filtered effectively forcing c_d and a_{kd} to consist of evaluations of smooth functions.

The model is completed specifying prior distributions for the unknown parameters δ, G, σ^2 and the coefficients s_m, $m = 1...M$.

$$\delta \sim N(0, \beta I_{r_Q}),$$

$$\gamma_k|\sigma_k^2 \sim N(0, \sigma_k^2 I_{r_R}) \quad \text{independently for } k = 1...K,$$

$$\sigma_k^2|\alpha_{A0}, \beta_{A0} \sim IG(\alpha_{A0}, \beta_{A0}) \quad \text{independently,}$$

$$\sigma^2|\alpha_0, \beta_0 \sim IG(\alpha_0, \beta_0),$$

where IG denotes an inverse Gamma distribution. $\alpha_0 = \beta_0 = \alpha_{A0} = \beta_{A0} = \beta^{-1} = 10^{-3}$ are chosen. Setting $\sigma_A = (\sigma_1^2...\sigma_K^2)$, the unknown parameter is $\theta = (\delta, G, \sigma_A, \sigma^2)$ with prior density

$$p(\theta) = p(\delta) \prod_{k=1}^{K} p(\gamma_k|\sigma_k^2) \prod_{k=1}^{K} p(\sigma_k^2) p(\sigma^2). \tag{5}$$

The generative model here is motivated as implementing the idea of a best approximating subspace for observed interpolation splines, and the s_m from this perspective are vectors of coefficients with respect to eigenfunctions spanning that subspace. In factor analysis the interpretation of such a model emphasizes that the N random variables $F(t_n)$, $t_n \in d$, (of which the $f_m(t_n)$ are realizations) are spanned by K latent factors $S_1...S_K$ with respect to which the the entries of the matrix A_d appear to be coefficients. In this interpretation the values s_{km} are 'scores', that is m realizations of K stochastically independent normalized factors. This view motivates the prior

$$s_m \sim N(0, I_K) \quad \text{independently for } m = 1...M,$$

which from a subspace point of view may be seen as a convenience prior. The factor analytic derivation of PCs as outlined by Tipping and Bishop (1999) justifies this choice, though. Combining all s_m as columns of a $K \times M-$ matrix s it is finally assumed that

$$p(s, \theta) = \prod_{m=1}^{M} p(s_m) p(\theta).$$

These priors are close to those chosen by Bishop (1999b). The likelihood of the model though is augmented by the matrices Q_m and R_m Given K, the smoothing parameters are to be estimated from the data, and the choice of all three parameters amounts to a problem of model choice. The posterior uncertainty about s and θ is to be interpreted conditional on the selection of

K, r_Q and r_R. Thus the uncertainty about the model is not reflected in the assessment of uncertainty about the mean function, the eigenfunctions or the reconstructed functions.

3.2 Variational inference

3.2.1 General set-up

By now variational inference has become an established method of deriving approximate posterior distributions, generalizing the EM-algorithm (Bishop, 2006, ch.10.1). Here it is reviewed only in as much as necessary to introduce notation. For the problem under consideration a good starting point is (Bishop, 1999a). Let y denote the data (here: $y_1, ..., y_M$) and z all unknown quantities (here: $z = (s, \theta)$), and let z be decomposed as $z = (z_1, ..., z_J)$, say. The key idea is to decompose the log-marginal density of the data, $\log p(y)$, into the Kullback-Leibler divergence between the posterior density $p(z|y)$ and an approximate posterior density $q(z)$, $KL(q(z), p(z|y)) = E_q(\log \frac{q(z)}{p(z|y)})$, and a remainder $L_q = E_q(\log \frac{p(y,z)}{q(z)})$. As a Kullback-Leibler divergence is non-negative, L_q is a lower bound of $\log p(y)$ which is the sharper the better q approximates the posterior density,

$$\log p(y) = L_q + KL(q(z), p(z|y)) \geq L_q. \tag{6}$$

To make approximate inference tractable, $q(z)$ is factorized into parametric densities, $q(z) = \prod_{j=1}^{J} q_j(z_j)$. The optimal approximate posterior density then satisfies

$$q_j(z_j) \propto \exp(E_{q \backslash j}(\log p(y, z)), \tag{7}$$

where $q \backslash j$ denotes the current joint density of the z_i without z_j. In practice initial values of the parameters of q_j are chosen and eq.(7) is evaluated iteratively. Application of eq.(7) guarantees that the lower bound L_q increases. L_q can often be computed and is used to monitor convergence defining a threshold for a minimum increase.

3.2.2 Application to functional PCA

The vector of unknown quantities (s, θ) is grouped into $(s, \delta, G, \sigma_A, \sigma^2)$ such that

$$q(s,\theta) = q(s)q(\delta)\prod_{k=1}^{K} q(\gamma_k)q(\sigma_A)q(\sigma^2) \tag{8}$$

(omitting now the subscripts j and giving the argument of q instead). Note that the factorization differs from the one proposed by Bishop (1999b) for ordinary probabilistic PCA without smoothness constraints. Bishop starts with $q(A_d)$ unstructured, and the update then yields $q(A_d) = \prod_{n=1}^{N} q(\widetilde{a}_{nd})$ where $\widetilde{a}_{nd}$ denotes the n-th *row* of A_d. Here initially a factorization $q(G) = \prod_{k=1}^{K} q(\gamma_k)$ is introduced which results in approximate posterior independence of the *columns* of G.

Updating iteratively according to eq.(7) yields approximate posterior distributions for $s, \delta, \sigma_A, \sigma^2$ which generalize those given by Bishop (1999b) in a straightforward manner. Here only the update of γ_k, $k = 1...K$, is discussed in more detail.

The update of γ_k is closely related to Bayesian estimation of γ_k in the 'stacked' linear model

$$\begin{bmatrix} y_{1k} \\ \vdots \\ y_{Mk} \end{bmatrix} = \begin{bmatrix} s_{k1}R_1 \\ \vdots \\ s_{kM}R_M \end{bmatrix} \gamma_k + \widetilde{\varepsilon}_k, \quad \widetilde{\varepsilon}_k \sim N(0, \sigma^2 I_{\widetilde{N}}), \tag{9}$$

where $\widetilde{N} = \sum_{m=1}^{M} N_{d_m}$, $y_{mk} = y_m - Q_m\delta - R_m G_{\setminus k} s_{\setminus km}$, with $G_{\setminus k}$ denoting G without the k-th column and $s_{\setminus km}$ denoting s_m without the k-th row, and $\gamma_k|\sigma_k^2 \sim N(0, \sigma_k^2 I_{r_R})$ apriori. In matrix notation eq.(9) reads $\widetilde{y} = \widetilde{X}\gamma_k + \widetilde{\varepsilon}$, and aposteriori $\gamma_k \sim N(\mu_k, \Sigma_k)$ with

$$\Sigma_k = (\frac{1}{\sigma^2}\widetilde{X}^T\widetilde{X} + \frac{1}{\sigma_k^2}I_{r_R})^{-1} = (\sum_{m=1}^{M}\frac{1}{\sigma^2}s_{km}^2 R_m^T R_m + \frac{1}{\sigma_k^2}I_{r_R})^{-1}, \tag{10}$$

$$\mu_k = \frac{1}{\sigma^2}\Sigma_k\widetilde{X}^T\widetilde{y} = \frac{1}{\sigma^2}\Sigma_k(\sum_{m=1}^{M} s_{km}R_m^T y_{mk}). \tag{11}$$

This transcribes into the update of $\gamma_k \sim N(\mu_{\gamma_k}, \Sigma_{\gamma_k})$ by

$$\Sigma_{\gamma_k} = \left(E_q(\frac{1}{\sigma^2}) R_d^T (\sum_{m=1}^{M} E_q\left(s_{km}^2\right) IP_m^T IP_m) R_d + E_q(\frac{1}{\sigma_k^2}) I_{r_R} \right)^{-1} \tag{12}$$

and

$$\mu_{\gamma_k} = E_q(\frac{1}{\sigma^2}) \Sigma_{\gamma_k} E_q \left(\sum_{m=1}^{M} s_{km} R_m^T (y_m - Q_m \delta - R_m G_{\setminus k} s_{\setminus km}) \right). \tag{13}$$

Note that in the evaluation of eq.(13) the covariance between the k-th row of s and the other rows in s has to be calculated under the current distribution q. Aposteriori only the columns of s are independent.

3.3 Rotation

One way to assess the relative importance of interpolation splines $h_{I(a_{kd})},\ a_{kd} = R_d \gamma_k$, is to investigate the posterior distribution of σ_k^2. If it is concentrated on small values, the coefficients γ_k tend to be small, that is, without much effect (see Bishop, 1999a). This approach, called 'automatic relevance determination' (ARD) is sometimes used to determine the number of 'necessary' factors, K, $K < \widetilde{K}$, where $\widetilde{K}$ is chosen initially large. However, the generative model determines the best approximating subspace of dimension K only on the whole. This feature also shows up in the fact that in the generative model, given K, $A_d s_m = A_d U U^{-1} s_m$ for any nonsingular $K \times K-$matrix U, and hence the eigenfunctions are indeterminate. A transformation GU may yield eigenfunctions which are possibly easier to interpret. We chose a rotation U (with $U^T = U^{-1}$) based on a singular value decomposition (SVD) of the posterior mean of A_d. Thus the columns of U are the K orthonormal eigenvectors of $E_{post}(G^T) R_d^T R_d E_{post}(G)$. The columns u_k can then be sequenced according to the size of the eigenvalues (that is, the first column corresponds to the largest eigenvalue etc.) and we set $\widetilde{A}_d = R_d E_{post}(G) U$. Thus the columns of $\widetilde{A}_d$ yield the interpolation splines corresponding to eigenfunctions in decreasing order. Note that orthogonality of the u_k in Euclidean space $\mathbb{R}^K$ implies neither orthogonality nor standardization of eigenfunctions $h_{I(R_d E_{post}(G) u_k)}$ in $H_{I(d)}$. An adaptation of the SVD to the (semi-)norm in $H_{I(d)}$ may therefore be an alternative option (compare the discussion in (van der Linde, 2003)). The choice depends on how one decides to measure the difference between two functions, taking explicitly into account the difference between their second derivatives or not. Here the Euclidean version is applied because in the examples d is a rather fine grid and this choice allows for a direct comparison to results obtained in ordinary PCA for vectors of function values without noise.

3.4 Choice of K, r_Q and r_R

The number of eigenfunctions K and the smoothing parameters r_Q for the mean function and r_R for the eigenfunctions are model parameters which can be determined maximizing the log marginal likelihood $\log p(y|K, r_Q, r_R)$ over a range of values. It is argued (Bishop, 2006, ch.10.4.1) that maximizing instead the lower bound $L_q(K, r_Q, r_R)$ (eq.(6)) is a reasonable alternative. For a model with indeterminacies the lower bound should be adjusted if the number of possible indeterminacies depends on the model parameters (see Bishop, 2006, ch.10.4.1). In our application the indeterminacy is driven by K (the model is determined up to a $K \times K-$matrix), however, for any K there are infinitely many feasible matrices. Hence in this case model comparison is directly based on the lower bound. The proposed method is empirically Bayesian rather than fully Bayesian. It implies that model uncertainty is not taken into account in the posterior distribution of other parameters. In contrast, in a fully Bayesian approach priors would have to be specified for K, r_Q and r_R yielding posterior distributions as well. When integrated out in the joint posterior distribution of all parameters these induce marginal distributions for the other parameters, particularly δ, G and s_m in which the model uncertainty is reflected. However, this comes at the cost of more sophisticated sampling techniques (Zhang et al., 2004).

3.5 Alternative smoothing priors

In a more direct approach smoothness priors could be specified on the columns a_{kd}. This approach was also tried with a Gaussian kernel prior and the thin plate spline prior. Both did not yield satisfying results, though. Introducing a smoothness prior amounts to smoothing *residual* curves in a linear model analogous to eq.(9), and separating noise becomes a subtle problem to be solved by an appropriate smoothing parameter.
One disadvantage of the thin plate spline prior is that it is a partially improper prior (Wahba, 1990; van der Linde, 1995). Although the corresponding posterior distributions are proper and can be calculated in closed form, the lower bound in variational inference requires the computation of expectations of the log prior density of A_d, respectively a_{kd}, and cannot be determined for an improper Gaussian prior. Hence this monitoring device is not directly available to guide the selection of the smoothing parameters σ_A and σ^2. (Remember that conventionally the smoothing parameter is given by $\lambda_k = \sigma^2/\sigma_k^2$.) This problem can be overcome setting up a separate flat but proper prior for the linear part of the function, mimicking the improper prior. However, next the choice of the convenience $IG-$priors with $\alpha_0 = \beta_0 = \alpha_{A0} = \beta_{A0} = 10^{-3}$ which is not too bad when smoothing single functions with splines (van der Linde,

2001) did not work in our example resulting in undersmoothed eigenfunctions. The estimates are very sensitive to the choice of the hyperparameters α_{A0}, β_{A0}, and although appropriate smoothing parameters could be found by trial and error and visual inspection or by using empirical input to set up sharp empirical priors, no satisfying 'automatic' procedure can be recommended.
Specifying alternatively a Gaussian smoothness prior process based on the covariance kernel $C(s,t) = \exp(-(s-t)^2/l^2)$ with nonsingular marginal covariance matrices was tried. Here l^2 acts as a smoothing parameter with the resulting estimated curves being the smoother the larger l. On a fine grid d the evaluation of the kernel may however result in a numerically singular covariance matrix if l is too large. For our design this happened for $l \geq 0.95$ but a larger l is needed to achieve sufficient smoothness of the eigenfunctions. This particular kernel appears to be too inflexible for the problem at hand. A modified kernel function with more smoothing parameters might be tried, but this approach was not pursued further. Instead the use of an effective spline basis with essentially one smoothing parameter as outlined in section 3.1 was felt to be more promising.

4 Examples

4.1 Simulated data

4.1.1 Complete curves with noise

$M = 100$ curves are now given as kernel estimates $\widehat{g}$ (eq.(3)) plus Gaussian noise as described in section 2. The dimensionality r_Q of Q_d to represent the mean function and r_R of R_d to represent the K eigenfunctions was determined maximizing the lower bound L_q for each $K \in \{1, ..., 6\}$ with respect to r_Q and r_R both within a range of 5 to 15. The maximization of the lower bound was based on 15 iterations (with one iteration comprising an update of all parameters), which correspond to a threshold of 0.0005 for the relative gain in the last iteration. (More formally, if $L_q(j)$ denotes the value of the lower bound L_q at iteration step j, we have $(L_q(15) - L_q(14))/L_q(14) \leqq 0.0005$. Ten iterations correspond to a threshold for the relative gain of 0.001 for $K \geqq 2$, and $(L_q(15) - L_q(10))/L_q(10) \leqq 0.003$.) Thus a few iterations are sufficient to reach the maximum lower bound.

In the example the following initial values for the parameters of the approximate distribution were used, where μ_0 denotes a mean, Σ_0 a covariance matrix, α_0 and β_0 the parameters of an inverse Gamma distribution, and 1_n is an n-

dimensional column vector of ones.

$$\mu_0(\delta) = 0.01 * I_{r_Q} \ , \qquad \Sigma_0(\delta) = I_{r_Q},$$

$\mu_0(\gamma_k)$ is a r_R-dimensional vector of random values
from a Uniform distribution on [0,1],

$$\Sigma_0(\gamma_k) = I_{r_R} \qquad \text{for } k = 1...K.$$

$\mu_0(s_m)$ is a K-dimensional vector of random values
from a standard Normal distribution,

$$\Sigma_0(s_m) = I_K \qquad \text{for } m = 1...M.$$

$$\alpha_0(\sigma^2) = \beta_0(\sigma^2) = \alpha_0(\sigma_k^2) = \beta_0(\sigma_k^2) = 1, \qquad \text{for } k = 1...K.$$

The algorithm is fairly robust with respect to the initialization. Problems may occur though, if a variance (prior) is too restrictive to allow for the necessary correction of a initial misfitting approximation during the first cycle(s).

Table 1 displays the results from which the set-up $K = 2$, $r_Q = 10$, $r_R = 8$ is read off to be optimal.

Model Parameters

K	L_q	r_Q	r_R
1	1.9102	10	8
2	1.9429	10	8
3	1.9327	10	8
4	1.9233	10	12
5	1.9085	10	15
6	1.8883	7	14

Tab.1: Each row displays the r_Q and r_R for which the lower bound attains a maximum given K. The value of the maximum is 10^4 times the value listed in the second column.

The results discussed in the sequel are based on the variational algorithm run for the optimal values with 15 iterations. $K = 2$ is coherent with the finding that in ordinary PCA without noise and without smoothing two components explained 87% of the variation.

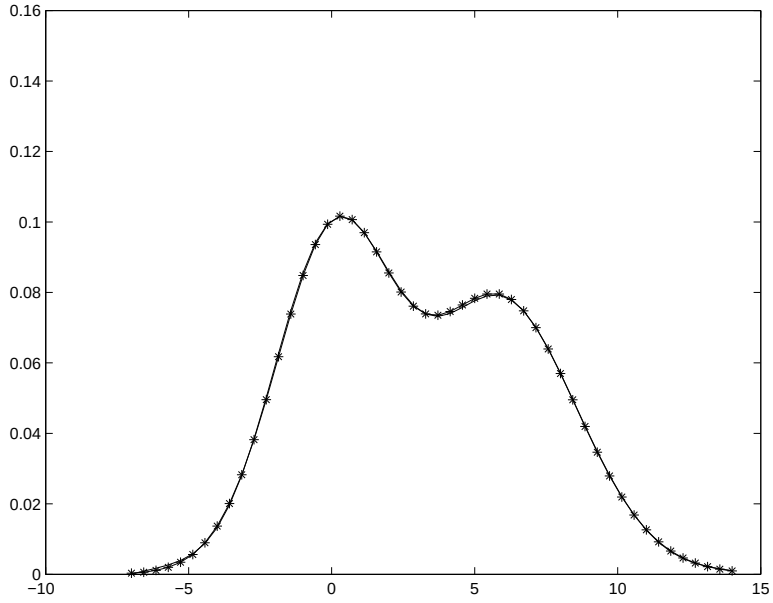

Fig. 4. The posterior mean function of the smoothed PCA (solid line) and the empirical mean function of the original curves (stars).

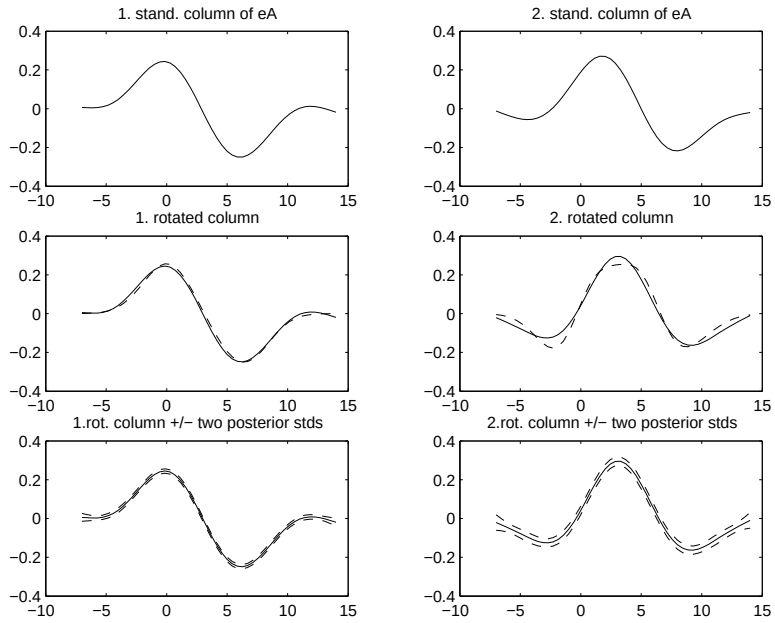

Fig. 5. First row: Posterior means of first and second column of A_d. Second row: Rotated posterior means (solid lines) and corresponding eigenfunctions from PCA of discretized curves without noise (dashed lines). Third row: Rotated posterior means (solid lines) plus/minus two posterior standard deviations (dashed lines).

In fig.4 the posterior mean function is compared to the mean function of the discretized curves, showing a perfect recovery. (In the graphical displays all functions are visualized using simple linear interpolation of 50 function values. Smooth interpolation at intermediate points would be appropriate on a coarser grid but for the fine grid in this example the difference is hardly visible.)

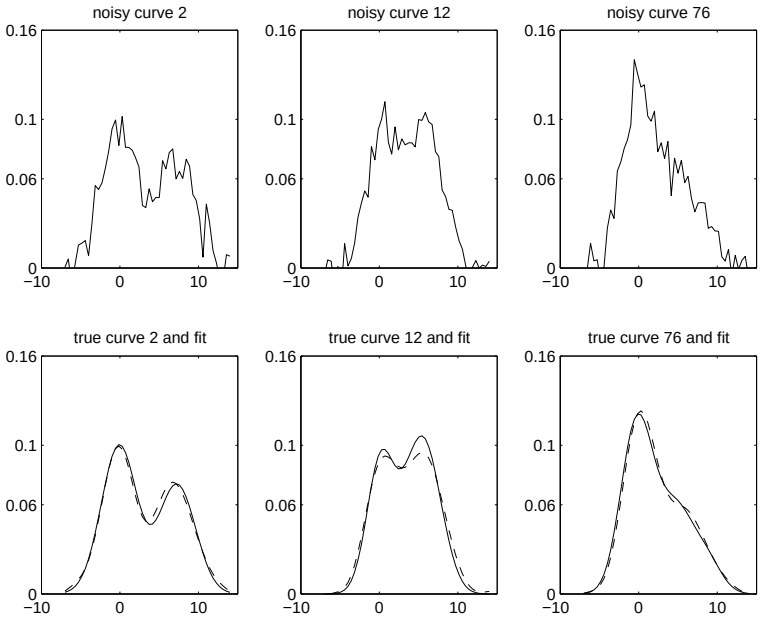

Fig. 6. First row: Three observed noisy curves. Second row: True function values (solid lines) and reconstructed function values (dashed lines).

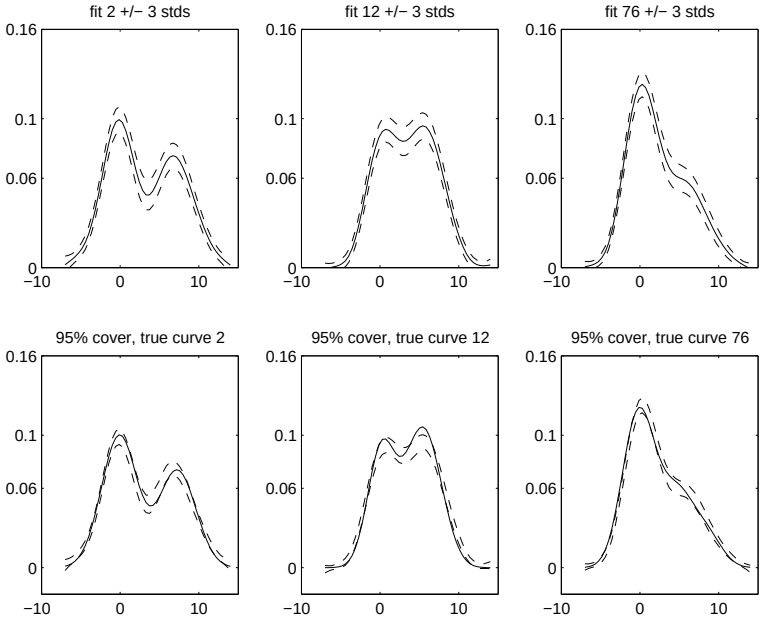

Fig. 7. First row: Fitted function values (solid lines) +/- three standard deviations (dashed lines). Second row: True functions (solid lines) in a 95% credible interval (dashed lines).

In fig.5 the two eigenfunctions are displayed: the posterior means of a_{1d} and a_{2d}, standardized to norm 1 as vectors of 50 points, appear to be similar. But after rotation the corresponding eigenfunctions resemble the first two eigenfunctions from ordinary PCA with true function values representing the same modes of variation: skewness and separation of peaks. The eigenfunctions are determined with high posterior precision. Finally, the reconstruction of three

original curves with two eigenfunctions is illustrated in fig.6. The approximate posterior distribution immediately provides an assessment of accuracy, both for the eigenfunctions as illustrated in fig.5 (third row) as for the reconstructed functions as illustrated in fig.7. While for each value of a reconstructed function the marginal standard deviation can be derived analytically (conditioning on s_m), the credible sets require (easy) simulations from the approximate factorized posterior distribution. The credible sets are based on 1000 simulations of vectors f_{md}.

A posteriori $\sigma^2 \sim IG(2.5 * 10^3, 0.3016)$, yielding $E(\sigma^2|data) = 0.00013$ as estimate of the true error variance $\sigma^2 = 0.0001$ with high accuracy, $std(\sigma^2|data) = 0.26 * 10^{-5}$. Note however, that for the curves without noise the unexplained variance given two PCs is about 0.00037.

4.1.2 Partial curves with noise

Fig.8 gives an impression of the data set of noisy segments.

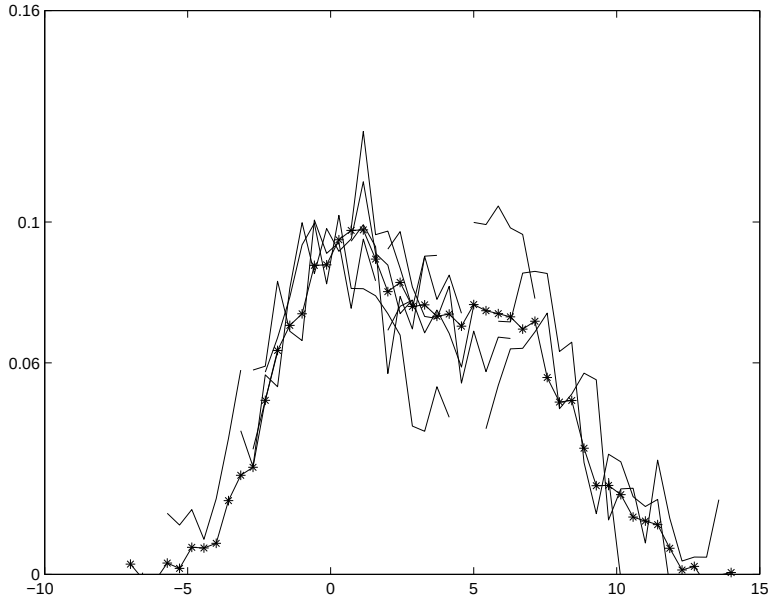

Fig. 8. 15 noisy partial curves (solid lines) and mean of all 100 noisy partial curves (stars).

Model comparison based on the lower bound suggests a model with $K = 2$, $r_Q = 9$, $r_R = 6$ which was then run with 20 iterations corresponding to a threshold of 0.0001 for the relative gain in the lower bound. The same initialization was used as for the complete curves. The two resulting eigenfunctions shown in fig.9 differ little from those obtained with complete noisy curves The reconstruction of the original curves, exemplified in fig.10, is slightly worse than with complete noisy observations.

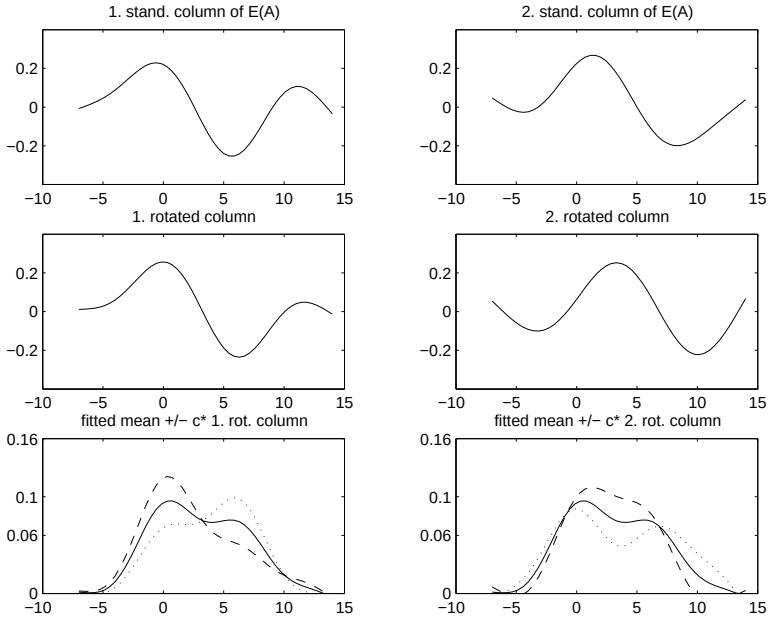

Fig. 9. First row: Linearly interpolated posterior means of first and second column of A_d. Second row: Linearly interpolated rotated posterior means (solid lines). Third row: Fitted mean function (solid line) plus (dashed line) and minus (dotted line) $c = 0.1$ times a rotated eigenfunction.

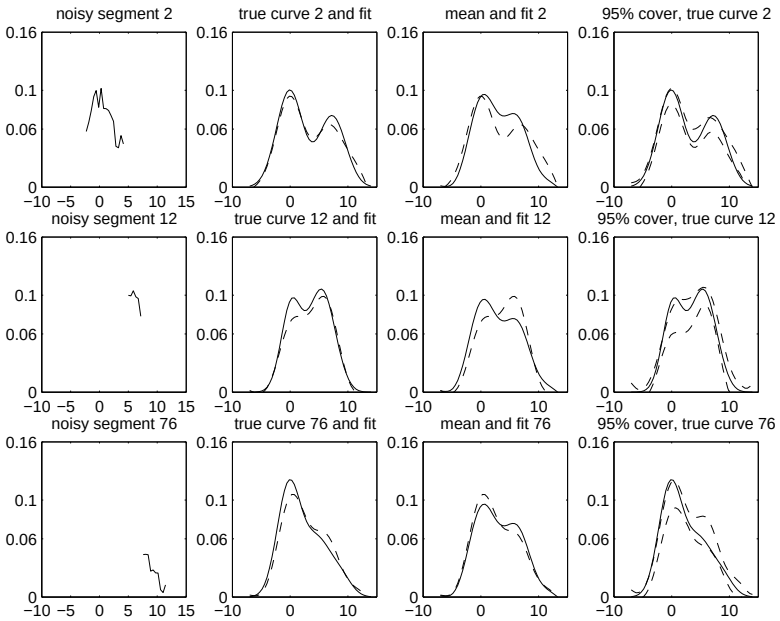

Fig. 10. First column: Noisy segments of selected curves. Second column: True curves (solid lines) and reconstructed curves (dashed lines). Third column: Fitted mean curve (solid line) and reconstructed curve (dashed line). Fourth column: True curves (solid lines) and 95% credible sets (dashed lines).

Remember that in this reconstruction the basic fit to all curves is the smoothed mean function to which all partial curves contribute. Additionally a partial (noisy) curve, which may be only a point, indicates a deviation from the mean which is incorporated using the eigenfunctions. Thus a local deviation from the mean curve induced by (noisy) observations can yield another deviation of the fitted function from the mean function in parts of the domain where no observations of that function are available. This is illustrated in the third column of fig.10. In the first row the observations mainly induce a deeper trough between the two modes, but also a heavier tail on the right hand side occurs. In the second row the modes are switched due to the few observations, and in the third row observations in the tail even yield a modification of the modes such that the resulting curve is closer to the true function than to the mean function.

The posterior distribution of σ^2 is $IG(422.5, 0.0477)$, hence $E(\sigma^2|data) = 0.000113$, a reasonable estimate of the error variance $\sigma^2 = 0.0001$.

4.2 Real data: Precipitation data

The data set comprises the average of daily precipitation for 35 Canadian weather stations recorded 1960-1994. The data set is available from Ramsay's website http://ego.psych.mcgill.ca/misc/fda. Related temperature records are analyzed in the books by Ramsay and Silverman (2005, 2002). The precipitation data were subjected to a frequentist FPCA using free-knot B-splines by Gervini (2006). The 365 observations of each curve exhibit considerable noise, showing up as large oscillations in a plot with linear interpolation of function values displayed in the left column of fig.12. The values are also rounded as becomes obvious in scatter plots shown in the right column of fig.12. Applying our approach (with 15 iterations and the same initialization as for the simulated data sets) a model with $K = 4$ modes of variation, $r_Q = 4$, that is a rather smooth mean function, and $r_R = 16$ is proposed. Indeed the Bayesian mean function is smoother than Gervini's mean function estimated independently of the eigenfunctions. An important issue of the model in eq.(3) respectively in eq.(4) is that both the mean curve and the residual curves relative to the mean curve are smoothed. In order to obtain competitive, interpretable eigenfunctions the mean function has to be estimated adequately, in contrast to the reconstruction of the individual curves where eigenfunctions can compensate for an insufficient mean function. Here, differing features (bimodality in summer, a 'shoulder' in late autumn) of Gervini's mean function are compensated by the second and third Bayesian eigenfunction. The 95% credible set based on 1000 simulations for the fitted mean function is narrow (top right panel of fig.11). The first three eigenfunctions (after rotation) and their deviations from the mean function are displayed in fig.11. The eigenfunctions are sorted

from top to bottom according to decreasing corresponding singular values.

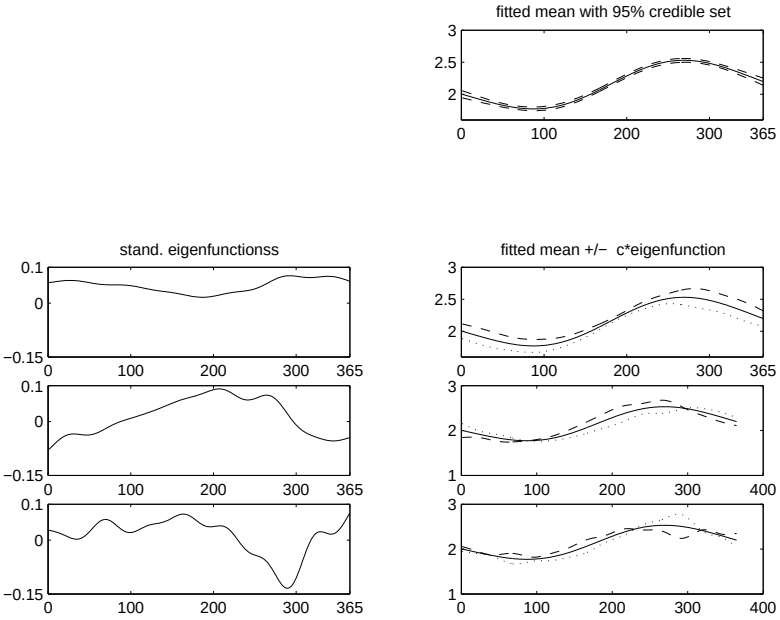

Fig. 11. Top right: Fitted smooth mean function with 95% credible set. Second to fourth row: First column: Eigenfunctions for precipitation data. Second column: Fitted mean function (solid line) plus (dashed lines) and minus (dotted lines) $c = 2$ times the eigenfunction displayed in the first column.

The first eigenfunction accounts for the general level of precipitation between September and March which differs between continental and marine stations.. The second eigenfunction represents a shift in seasons. The third eigenfunction models the maximum difference between spring and autumn ('dry' spring and 'wet' autumn or more balanced), indicating also on a finer temporal scale variations in March and October. The forth eigenfunction (not shown) introduces additional temporal rhythm with periods of six weeks. The first three eigenfunctions are qualitatively comparable to those found by Gervini (2006, fig.10) which are, however, smoother corresponding to his different trade off between mean function and eigenfunctions discussed before. The more pronounced local minima in Gervini's third eigenfunction are accounted for by our fourth eigenfunction. The amount of smoothness required might be judged looking at the fitted individual curves (fig.12). The oscillations support the impression that eigenfunctions which are wiggly at a medium scale are more appropriate, while local peaks and troughs are less pronounced in the scatter plots. Though the rhythm of precipitation may be of meteorological interest, it hardly effects the reconstruction of curves, because the fourth eigenfunction contributes only 1% in spanning the subspace of columns of A_d. The trace criterion applied to the singular values yields 85% for the first PC, another 10% for the second PC and additional 4% for the third one. The 95%-credible sets displayed in fig.12 (left column) indicate only little uncertainty about the reconstructed functions.

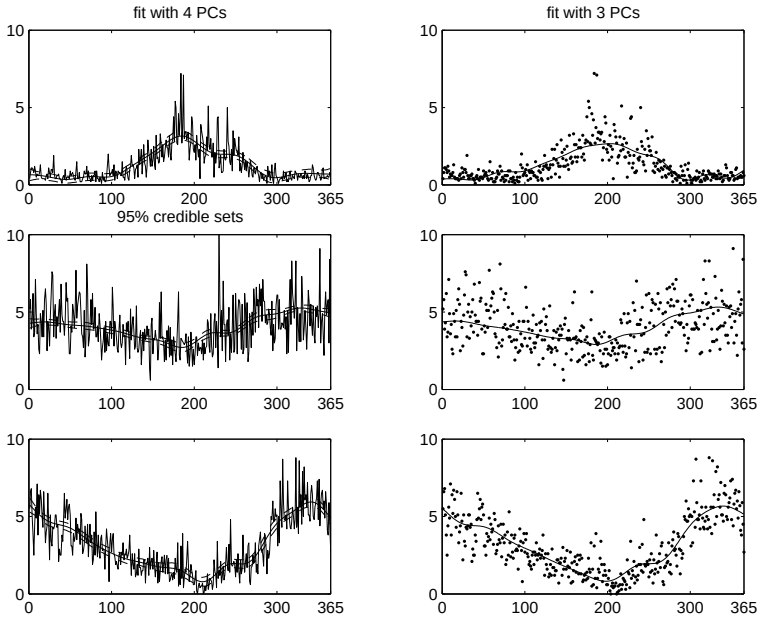

Fig. 12. Fit of precipitation curves for the Canadian weather data. First column: Observed function values linearly interpolated and fit using all four eigenfunctions together with 95% credible sets. Second column: Scatterplots of the same data and fit using the first three eigenfunctions.

5 Discussion

5.1 Smooth mean functions

A difference between the (frequentist) representative and the (Bayesian) generative approach lies in the treatment of the mean function. If the focus is on estimating the covariance, the mean function is regarded as nuisance parameter and estimated separately. In the generative approach the mean function and the eigenfunctions as smooth functions are estimated simultaneously (respectively alternatingly in the variational algorithm), and a crucial issue for the performance of the model is whether they are adequately distinguished. Experience up to now indicates that the mean function obtained from a generative model tends to be slightly smoother than the empirical mean functions given in the literature.

5.2 Error assessments

As emphasized before Bayesian inference refers to the posterior distribution based on only the data (and the prior) rather than asymptotic distributions of estimators. Yet there is also an interest in Bayesian asymptotics, that is, in the

asymptotic posterior distribution. With increasing sample size the prior tends to become less influential. A major advantage of the Bayesian approach is the coherent and integrated handling of errors, simplified in variational inference by assumptions of posterior independence. Hence the accuracy of the variational approximation to the posterior distribution is crucial for the validity of the whole approach. Wang and Titterington (2004) established (asymptotically) efficiency of the Bayesian variational estimators and Normality of the variational asymptotic posterior distribution for a broad class of models. In the approach proposed in this paper model uncertainty is not incorporated into the error assessment, and in this respect errors may be underestimated. Alternative sampling techniques (MCMC) share common features with variational inference (Cemgil et al., 2007), and hybrid computational techniques may be developed in the future.

5.3 Spiky curves

The example in section 4.2 takes us to the question whether the proposed approach is appropriate for very spiky curves like those considered by Behseta et al. (2005). Both Behseta et al. (2005) and Gervini (2006) argue that presmoothing of single curves with free-knot B-splines is necessary in order not to oversmooth in FPCA. Moreover, Gervini (2006) demonstrates that in this respect B-splines are more adequate than smoothing splines based on generalized cross-validation (GCV). The basis functions used in the generative model are related to interpolation splines in a reproducing kernel Hilbert space which do not depend on any data driven smoothing parameter. They form a very flexible class of functions: given sufficiently many knots and function values at these knots any however spiky function can be smoothly interpolated. Smoothing splines in turn are interpolation splines in estimated function values. If oversmoothing occurs with smoothing splines it is due to (insufficiently many knots or) the method of estimation which usually fits an interpolation spline subject to a *global* roughness constraint determined by one smoothing parameter. Fitting the generative model amounts to *multiple* curve fitting (without using (GCV)) and works the better the more homogeneous the curves are. This is illustrated by the good performance for the seemingly spiky, but as a group homogeneous precipitation functions. In contrast, local undersmoothing occurred in some reconstructions of mixture density estimates which are quite heterogeneous with respect to bi-modality.

5.4 Gaussian distributions

The prior assumption of Gaussianity $s_m \sim N(0, I_K)$, $m = 1...M$, does not seem to be crucial for functional data. If it is doubted the variational algorithm could be extended to latent factors assumed to be distributed like a mixture of Gaussians (Attias, 1998; Choudrey and Roberts, 2001). Non-Gaussianity of 'signals' has been a major concern in the development of Independent Component Analysis and Independent Factor Analysis, where generative models are frequently set up. See for example (Févotte and Godsill, 2006). The assumption of Gaussianity has an impact on the applicability and the properties of the variational algorithm, but it does not affect the approach of enforcing smoothness of the eigenfunctions. The Normal distributional assumption for the observed data (given the parameters) can also be doubted. Again, this assumption can be relaxed as shown by Tipping and Lawrence (2005).

5.5 Model choice

The selection of the number of eigenfunctions K and their degree of smoothness (r_R) as well as that of the mean function (r_Q) is a problem of model choice, which might be solved in many different ways (Lopes and West, 2004). Here only the effectiveness of maximizing the lower bound L_q of the marginal log-likelihood was demonstrated. The marginal log-likelihood belongs to the group of 'prior predictive' criteria, which target the explanation of the data given the prior, in contrast to 'posterior predictive' criteria like AIC or DIC which aim at the prediction of new observations following the same data generating process given the posterior. We feel that (F)PCA is used as an exploratory technique mainly, and hence a prior predictive criterion is more appropriate than a posterior predictive criterion.

5.6 Extensions

Although the approach to Bayesian FPCA proposed in this article was applied only to univariate thin plate splines it can immediately be extended to any function space that is a reproducing kernel Hilbert space, in particular to multivariate thin plate splines (images), higher orders of differentiability or to any feature, not necessarily smoothness, incorporated in a reproducing kernel. The optimality of the Demmler-Reinsch like basis functions in spanning interpolation splines holds generally (van der Linde, 2003).

The function space of interpolation splines can also be useful in other fields of FDA like functional regression. It is a finite dimensional function space in

which observed functions or parameter functions may be expanded parsimoniously. The derivation of the variational algorithm for other fields of FDA requires further investigation, though.

5.7 Conclusions

A pragmatic exploratory Bayesian approach was suggested to study modes of variation in a bundle of curves.
Advantages of the generative model are the variability of observational designs for individual curves and the parsimony and versatility of the basis functions. The main advantage of variational inference is that the algorithm is fast, in closed form and fairly robust with respect to the initialization, providing estimation, error assessment and model choice simultaneously.
The accuracy of the approximation to the posterior distribution and its impact on substantive conclusions is still under investigation and requires further research. The applicability of the proposed method may be limited to sets of curves which are fairly homogeneously smooth (or spiky) across curves and in particular not too spiky individually at different locations.

Acknowledgement

The author would like to thank two anonymous referees for valuable hints and comments which helped to considerably improve an earlier draft of this paper.

References

Aguilera, A.M., Escabias, M. and Valderrama, M.J., 2008. Forecasting binary longitudinal data by a functional PC-ARIMA model. Comp. Statist. Data Anal. 52, 3187-3197.

Aguilera, A.M., Escabias, M. and Valderrama, M.J., 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comp. Statist. Data Anal. 50, 1905-1924.

Aguilera, A.M., Ocaña, F.A. and Valderrama, M.J., 1999a. Stochastic Modelling for Evolution of Stock Prices by Means of Functional Principal Component Analysis. Appl. Stochastic Models Bus. Ind. 15, 227-234.

Aguilera, A.M., Ocaña, F.A. and Valderrama, M.J., 1999b. Forecasting time series by functional PCA. Discussion of several weighted approaches. Comp. Statist. 14, 442-467.

Aguilera, A.M., Ocaña, F.A. and Valderrama, M.J., 1999c. Forecasting with unequally spaced data by a functional principal component approach. Test 8, 233-253.

Aguilera, A.M., Ocaña, F.A. and Valderrama, 1997. An approximated principal component prediction model for continuous-time stochastic processes. Applied Stochastic Models and Data Analysis 13, 61-72.

Attias, H., 1998. Independent factor analysis. Neural Computation 11, 803-851.

Behseta, S., Kass, R.E., Wallstrom, G.L., 2005. Hierarchical Models for Assessing Variability among Functions. Biometrika 92, 419-434.

Besse, P.C., 1994. Models for Multivariate Data Analysis. In: R. Dutter and W. Grossmann (Eds.), COMPSTAT, Proceedings in Computational Statistics, Physica, Heidelberg.

Besse, P.C., 1988. Spline functions and optimal metric in linear principal component analysis. In: J.A. van Rijckevorsel and J. de Leeuw (Eds.), Component and Correspondence Analysis, Wiley, New York, 81-101.

Besse, P., Ferraty, F. and Cardot, H., 1997. Simultaneous nonparametric regressions of unbalanced longitudinal data. Comp. Statist. Data Anal. 24, 255-270.

Bishop, C.M., 1999a. Bayesian PCA. In: M.S. Kearns, S.A. Solla and D.A. Cohn (Eds.), Advances in Neural Information Processing Systems 11, 382-388, MIT Press.

Bishop, C.M., 1999b. Variational principal components. In: Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99, IEE, 509-514.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Boente, G. and Fraiman, R., 2000. Kernel-based functional principal components. Statist. Prob. Lett. 48, 335-345.

Bouzas, P.R., Ruiz-Fuentes, N. and Ocaña, F.M., 2007. Functional approach to the random mean of a compound Cox process. Comp. Statist. 22, 467-479.

Bouzas, P.R., Valderrama, M.J., Aguilera, A.M. and Ruiz-Fuentes, N., 2006. Modelling the mean of a doubly stochastic Poisson process by functional data analysis. Comp. Statist. Data Anal. 50, 2655-2667.

Cardot, H., 2000. Nonparametric Estimation of Smoothed Principal Components Analysis of Sampled Noisy Functions. Nonparametric Statistics 12, 503-533.

Cardot, H., Ferraty, F. and Sarda, P., 2003. Spline estimators for the functional linear model. Statistica Sinica 13, 571-591.

Cemgil, A.T., Févotte, C. and Godsill, S.J., 2007. Variational and Stochastic Inference for Bayesian Source Separation. Preprint, Engineering Dept., Univ. of Cambridge

Chiou, J.-M. and Müller, H.-G., 2007. Diagnostics for functional regression via residual processes. Comp. Stat. Data Anal. 51, 4849-4863.

Choudrey, R.A. and Roberts, S.J., 2001. Flexible Bayesian independent component analysis for blind source separation In: 3rd International Conference on Independent Component Analysis and Blind Signal Separation, San Diego, Dec. 9-12, 2001, 90-95.

Dauxois, J., Pousse, A. and Romain, Y., 1982. Asymptotic theory for the principal component analysis of a random vector function: some application to statistical inference. J. Multiv. Anal. 12, 136-154.

Davidian, M., Lin, X. and Wang, J.-L., 2004. Emerging Issues in Longitudinal and Functional Data Analysis. Introduction., Statistica Sinica 14, 613-614.

Escabias, M., Aguilera, A.M. and Valderrama, M.J., 2007. Functional PLS logit regression model. Comp. Statist. Data Anal. 51, 4891-4902.

Escabias, M., Aguilera, A.M. and Valderrama, M.J., 2005. Modeling environmental data by functional principal component logistic regression., Environmetrics 16, 95-107.

Escabias, M., Aguilera, A.M. and Valderrama, M.J., 2004. Principal component estimation of functional logistic regression: discussion of two different approaches. Nonparametric Statistics 16, 365-384.

Fernández-Alcalá, R.M., Navarro-Moreno, J. and Ruiz-Molina, J.C., 2007. Functional estimation incorporating prior correlation information. Comp. Statist. 22, 439-447.

Ferraty, F. and Vieu, P., 2006. Nonparametric Functional Data Analysis. Springer, New York.

Févotte, C. and Godsill, S.J., 2006. Blind separation of sparse sources using Jeffreys' inverse prior and the EM algorithm. In: J. Rosca et al. (Eds.), ICA and blind signal separation. Lect. Notes in Comp. Sci. 3889, Springer, Berlin, 593-600.

Gervini, D., 2006. Free-knot spline smoothing for functional data. J. Roy. Statist. Soc. B68, 671-687.

Grambsch, P.M., Randall, B.L., Bostick, R.M., Potter, J.D. and Louis, T.A., 1995. Modeling the Labeling Index Distribution: An Application of Functional Data Analysis. J. Amer. Statist. Ass. 90, 813-821.

Hall, P., Müller, H.-G. and Wang, J.-L., 2006. Properties of Principal Component Methods for Functional and Longitudinal Data Analysis. Ann. Statist. 34, 1493-1517.

Hyndman, R.J. and Ullah, M.S., 2007. Robust forecasting of mortality and fertility rates: A functional data approach., Comp. Staist. Data Anal. 51, 4942-4956.

Hyvärinen, A., Karhunen, J. and Oja, E., 2001. Independent Component Analysis. Wiley, New York.

James, G.M., Hastie, T.J. and Sugar, C.A., 2000. Principal component models for sparse functional data. Biometrika 87, 587-602.

Jones, M.C. and Rice, J., 1992. Displaying the important features of large collections of similar curves. American Statistician. 46, 140-145.

Kneip, A. and Utikal, K.J., 2001. Inference for Density Families Using Functional Principal Component Analysis, J. Amer. Statist. Ass. 96, 519-532.

Leng, X. and Müller, H.-G., 2006. Classification using functional data analysis for temporal gene expression data. Bioinformatics 22, 68-76.

van der Linde, A., 2003. PCA-based Dimension Reduction for Splines. Nonparametric Statistics 15, 77-92.

van der Linde, A., 2001. Estimating the Smoothing Parameter in Generalized Spline-Based Regression II: Empirical and Fully Bayesian Approaches. Experiments with Gibbs Sampling in Simulated Examples. Comp. Statist. 16, 73-95.

van der Linde, A., 1995. Splines from a Bayesian Point of View. Test 4, 63-81.

Lopes, H.F. and West, M., 2004. Bayesian Model Assessment in Factor Analysis. Statistica Sinica 14, 41-67.

Manté, C., Yao, A.-F. and Degiovanni, C., 2007. Principal component analysis of measures, with emphasis on grain size curves. Comp. Statist. Data Anal. 51, 4969-4983.

Manteiga, W.G. and Vieu, P., 2007. Statistics for Functional Data. Comp. Statist. Data Anal. 51, 4788-4792.

Minka, T.P., 2001. Automatic choice of dimensionality for PCA. In: T.K. Leen, T.G. Dieterich and V. Tresp (Eds.), Advances in Neural Information Processing Systems 13, 598-604, MIT Press.

Müller, H.-G., 2005. Functional Modelling and Classification of Longitudinal Data. Scand. J. Statist. 32, 223-240.

Müller, H.-G. and Stadtmüller, U., 2005. Generalized functional linear models. Ann. Statist. 33, 774-805.

Müller, H.-G. and Wang, J.-L., 2005. Functional Data Analysis for Sparse Longitudinal Data. J. Amer. Statist. Ass. 100, 577-590.

Nerini, D. and Ghattas, B., 2007. Classifying densities using functional regression trees. Applications in oceanology. Comp. Statist. Data Anal. 51, 4984-4993.

Ocaña, F.A., Aguilera, A.M. and Escabias, M., 2007. Computational considerations in functional principal component analysis. Comp. Statist. 22, 449-465.

Ocaña, F.A., Aguilera, A.M. and Valderrama, M.J., 1999. Functional principal component analysis by choice of norm. J. Multiv. Anal. 71, 262-276.

Pezzulli, S.D. and Silverman, B.W., 1993. Some properties of smoothed principal components analysis for functional data. Comp. Statist. 8, 1-16.

Ramsay, J.O. and Silverman, B.W., 2005. Functional Data Analysis. Springer, New York, 2nd ed.

Ramsay, J.O. and Silverman, B.W., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer, New York .

Reinsel, G.C. and Velu, R.P., 1998. Multivariate reduced-rank regression. Springer, New York.

Rice, J.A., 2004. Functional and longitudinal data analysis: perspectives on smoothing. Statistica Sinica 14, 631-647.

Roberts, S.J. and Everson, R.M., 2001. Independent Component Analysis. Cambridge University Press.

Rowe, D., 2003. Multivariate Bayesian Statistics. Chapman and Hall, London.

Saporta, G., Costanzo, G.D. and Preda, C., 2007. Linear methods for regression and classification with functional data. IASC-ARS special conference, Seoul, June 7-8, 2007.

Silverman, B.W., 1996. Smoothed functional principal component analysis by choice of norm. Ann. Statist. 24, 1-24.

Silverman, B.W., 1995. Incorporating parametric effects into functional principal components analysis. J. Roy. Statist. Soc. B57, 673-689.

Šmídl, V. and Quinn, A., 2007. On Bayesian principal component analysis. Comput. Stat. Data Anal. 51, 4101-4123.

Tipping, M.E. and Bishop, C.M., 1999. Probabilistic principal component analysis. J. Roy. Stat. Soc. B21, 611-622.

Tipping, M.E. and Lawrence, N.D., 2005. Variational inference for Student-t models: robust Bayesian interpolation and generalized component analysis. Neurocomputing 69, 123-141.

Valderrama, M.J., 2007. An overview to modelling functional data. Comp. Statist. 22, 331-334.

Valderrama, M.J., Ocaña, F.A. and Aguilera, A.M., 2002. Forecasting PC-ARIMA models for functional data. In: W. Haerdle and B. Roenz (Eds.) Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, 25-36.

Wahba, G., 1990. Spline Models for Observational Data. SIAM, Philadelphia, Pennsylvania.

Wang, B. and Titterington, D.M., 2004. Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In: M. Chickering and J. Halpern (Eds.), Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence. AUAI Press.

Yao, F. and Lee, T.C.M., 2006. Penalized spline models for functional principal component analysis. J. Roy. Statist. Soc B68, 3-25.

Zhang, Z., Chan, K.L., Kwok, J.T. and Yeung, D.-Y., 2004. Bayesian Inference on Principal Component Analysis using Reversible Jump MCMC. In: Proceedings of the 19th National Conference on Artificial Intelligence (AAAI'04), San Jose, California, USA, 25-29 July 2004.

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