Bayesian Analysis (2011) **6**, Number 1, pp. 77–126

Reduced rank regression models with latent variables in Bayesian functional data analysis

Angelika van der Linde*

Abstract. In functional data analysis (FDA) it is of interest to generalize techniques of multivariate analysis like canonical correlation analysis or regression to functions which are often observed with noise. In the proposed Bayesian approach to FDA two tools are combined: (i) a special Demmler-Reinsch like basis of interpolation splines to represent functions parsimoniously and flexibly; (ii) latent variable models initially introduced for probabilistic principal components analysis or canonical correlation analysis of the corresponding coefficients. In this way partial curves and non-Gaussian measurement error schemes can be handled. Bayesian inference is based on a variational algorithm such that computations are straight forward and fast corresponding to an idea of FDA as a toolbox for explorative data analysis. The performance of the approach is illustrated with synthetic and real data sets.

Keywords: functional data analysis, functional canonical correlation analysis, functional regression, functional prediction, functional discriminant analysis.

1 Introduction

Frequently it is of interest to analyze the impact of a functional covariate X on a functional or non-functional response Y. An introduction to regression analysis with functions and more generally to functional data analysis (FDA) is given in the monographs by Ramsay and Silverman (2002, 2005). The aim of the analysis typically is either descriptive - to characterize the structure of dependence between X and Y - or predictive - to set up a model for predicting Y given $X = x$. Special models are discriminated according to the type of conditional distribution of Y. Important special cases with functional Y are functional canonical correlation analysis (FCCA) respectively functional regression. For a non-functional response Y generalized linear models with functional predictors constitute a popular class of models. With binary response Y functional discriminant analysis respectively functional logistic regression arise as special cases. Also, the special case of scalar prediction (with Gaussian response Y) has attracted a lot of attention. Recent approaches to regression with functional predictors are discussed in the special issues on FDA of several statistical journals (Davidian et al., 2004: Statistica Sinica; Valderrama, 2007: Computational Statistics; Manteiga and Vieu, 2007: CSDA). Nonparametric approaches are presented by Preda (2007) and in the monograph by Ferraty and Vieu (2006).

A major part of Bayesian FDA has been developed in the spirit of nonparametric Bayesian statistics (e.g. Petrone, Guindani and Gelfand, 2009) rather than Bayesian

*Institute of Statistics, University of Bremen, Germany, mailto:avdl@math.uni-bremen.de

DOI:10.1214/11-BA603

multivariate analysis. Much of the work is focused on clustering functions (Bigelow and Dunson, 2006; MacLehose and Dunson, 2009; Ray and Mallick, 2006), also applied to solve regression problems (Dunson et al., 2008; Rodriguez et al., 2009). Other Bayesian work, utilizing the expansion of functions in a suitable basis (Thompson and Rosen, 2008; Baladandayuthapani et al., 2007), is closer to the approach presented here.

An intermediate step in most approaches to FDA is regularization, often by a representation of X with respect to a (finite dimensional) basis of functions. It allows to cast functional regression back into the framework of multivariate multiple regression with basis coefficients representing the functions. In this paper interpolation splines will be used as basic functions. Often, in order to simplify the structure of dependence and to stabilize estimation and prediction, beyond the representation in a basis further dimension reduction in the predictor space is desirable. It can be achieved by variable selection or the identification of effective (linear) functional subspaces. Thus after regularization standard dimension reduction techniques of multivariate analysis may be applied and have indeed been tried in FDA: partial least squares (PLS), principal components in regression (PCA) or reduced rank regression (RRR). These techniques for dimension reduction can be related in hybrid approaches addressed as "continuum regression" such that the transition from one approach to the other is driven by a single parameter (Brooks and Stone, 1994; Merola and Abraham, 2001; Sundberg, 2002; Bougeard et al., 2008). An early review paper providing insights on implicit parameter constraints, also from a Bayesian point of view, is (Frank and Friedman, 1993).

The regression models mentioned so far are based on the conditional distribution of Y given $X = x$. In many applications with functional X the data is noisy and/or incomplete such that functions as realizations of X are not directly observable. Measurement models for X may be set up, the curves individually filtered and then used in regression (often neglecting the error due to filtering) if there is sufficient data. Thus the underlying curves are regarded as latent curves, which can be determined individually (Zhang et al., 2007; Cardot et al., 2007). In contrast, if the data is sparse, individual denoising may not be possible. Common latent variables can be introduced instead which are estimated using all partial (noisy) curves (James and Hastie, 2001; James, 2002). Such latent curves can simultaneously be related to the idea of dimension reduction like principal components in regression and one is naturally led to reduced rank regression. This paper gives special emphasis to problems arising in all models if the regressor X is functional and only partially observed with noise in a subject-specific way.

RRR models with *manifest* variables have been studied extensively (Reinsel and Velu, 1998; Yee and Hastie, 2003; Srivastava, 2007), also from a Bayesian point of view (Schmidli, 1994; Geweke 1996), but RRR models with *latent variables* (which are not assumed to be linear combinations or other explicitly modelled transformations of the observed variables) less so. Within these models the regression of Y on X can be subjected to the requirement of de-correlating X (PCA) and/or Y ("redundancy analysis"), classically under Gaussian distributional assumptions. Related ideas were pursued by Klami and Kaski (2008), and an extension to variables with distributions in exponential families was suggested by Rish et al. (2008). Reduced rank models with latent variables are promising starting points in Bayesian multivariate analysis as demonstrated by Tipping and Bishop (1999) introducing “probabilistic principal component analysis” and by Bach and Jordan (2005) recovering canonical correlation analysis probabilistically. A similar non-functional model called “supervised probabilistic PCA” was suggested by Yu et al. (2006). Probabilistic principal component analysis was extended to functions by van der Linde (2008), and also investigated under error schemes in exponential families (van der Linde, 2009).

The purpose and contribution of this paper is to extend these ideas to try a Bayesian FCCA in an analogous way and furthermore to explore the potential of the reduced rank model in Bayesian functional regression more generally, allowing for partial and noisy observations of functions. Approximate posterior inference will be based on a variational algorithm extending that of Wang (2007). It is easy to implement, fast and hence provides a pragmatic approach to quick explorative Bayesian analyses. It is sufficient to reveal the potential respectively the potential weakness of a data set and indicates if further investigation using more sophisticated sampling techniques is worthwhile.

The RRR models studied in this paper are motivated by the need to cope with partial (noisy) predictor curves and are suitable for prediction. However, they share the principled weakness of latent variable models in (Bayesian) multivariate analysis, non-identifiability, and hence interpretation is not immediate. Alternative(ly parameterized) regression models may be preferable if the aim of the analysis is descriptive and data is not sparse. Bayesian FDA is only being emerging and not yet rich in methods and experience. In a more extensive technical report (van der Linde, 2010) pointers to alternative models, brief reviews of the largely frequentist literature related to important special cases of FDA and more worked examples of the RRR models proposed in this paper are given.

In the next section 2 the model is formally introduced. Inference and model choice are discussed in section 3. In section 4 special cases are addressed and examples of RRR with latent variables are given. Section 5 concludes with a discussion of extensions and open problems. Technical details of the variational algorithm are given in an appendix.

Thus a fully general model is specified first with the impact of a heavy notation and possibly difficult reading. However, the emphasis is on the versatility of the RRR model, and the examples mainly illustrate how special applications can be embedded while any special application could still be substantially refined. A reader being interested in a particular subclass of models like FCCA or discriminant analysis might prefer to start with an example in section 4, then study the graphical (sub-)models in section 2.1 for an overview and in order to identify his/her model and continue with the relevant sections on details of model specification and inference.

2 Model specification

2.1 Overview

Assume that for each of M subjects a pair of random vectors (X_m, Y_m) is observed with values x_m in $\Re^{N_m^X}$, $\quad y_m$ in $\Re^{N_m^Y}$ and that across subjects observations are (conditionally) independent. X_m and Y_m may represent partially observed curves f_m^X, f_m^Y such that the numbers N_m^X, N_m^Y of observed noisy function values as well as their locations (designs d_m^X, d_m^Y) may vary with m. The vector of function values at a design will be denoted using the design as subscript, e.g. $f^X_{md^X_m}$. For simplicity a generic X will be called the functional regressor or functional predictor, and a generic Y the response. A key idea in model building is to represent all X-functions as interpolation splines w.r.t. a *common* design d^X, decomposing them into a mean function and individual residual functions which are spanned in bases of splines interpolating values at d^X. The corresponding "outer model" is visualized in figure 1 where the usual convention of graphical models is applied: stochastic and logical nodes are represented by ellipses, constants by rectangles, edges by single arrows if they represent a distributional specification and by double arrows if they represent a deterministic relation. A plate indicates repetitions.

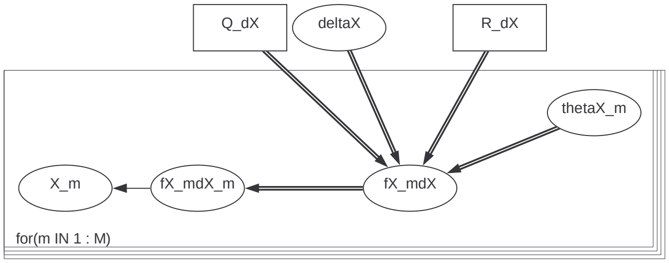

Figure 1: Graphical model for interpolation splines observed with noise.

Here the leftmost arrow represents an observational error scheme where the distribution of observations will be assumed to belong to an exponential family with unknown function values $f^X_{md^X_m}$ at the individual observational design d_m^X as canonical parameters. The horizontal double arrow symbolizes the interpolation step; given function values $f^X_{md^X}$ at a common design d^X the values $f^X_{md^X_m}$ at the individual design are obtained applying an interpolation matrix IP_m^X: $f^X_{md^X_m} = IP_m^X f^X_{md^X}$. The interpolation spline at the common design is in turn decomposed into a mean function, common to all X-curves, and a specific residual function. The basis functions of the mean function correspond to the columns of a matrix Q_{d^X}, those of the residual function to the columns of a matrix

R_{d^X} with coefficients δ^X and θ^X_m respectively: $f^X_{md^X} = Q_{d^X}\delta^X + R_{d^X}\theta^X_m$. The model is defined analogously for functional Y. If Y is univariate, functions reduce to values and interpolation is not needed.

The next step is to define reduced rank regression for the individual spline coefficients θ^X_m, θ^Y_m. In an "inner model", graphically displayed in figure 2, the coefficients are related by a common latent variable s_m.

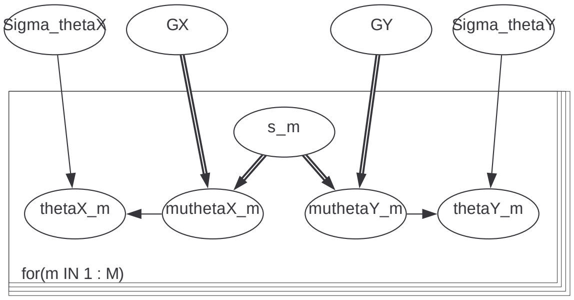

Figure 2: Graphical model for the RRR of coefficients θ^Z_m.

More formally, for $Z \in \{X, Y\}$ it is assumed that $\theta^Z_m \sim N(\mu_{\theta^Z_m}, \Sigma_{\theta^Z})$, $\mu_{\theta^Z_m} = G^Z s_m$, and conditionally θ^X_m, θ^Y_m are independent. This is essentially the probabilistic model introduced by Bach and Jordan (2005) for multivariate canonical correlation, where the columns of G^X, G^Y represent "patterns of covariation" and the values of s_m can be read as scores of transformed canonical variates, and both have to be estimated. The idea of this paper is (1) to extend the model to canonical correlation analysis for functions (possibly partially observed with noise) using the frame of the "outer model" and (2) to generalize it to a reduced rank regression model with functional covariate but arbitrary response by adapting the outer model for the response. In prediction problems the numbers of observed units are unequal, $M^X = M + m^X \geq M^Y = M$, say. In this case, for $m = 1, ..., m^X$, the latent variable s_m is determined by X_m only and θ^Y_m has to be inferred from s_m.

The full hierarchical model comprising the outer models for X and Y, the inner model for the coefficients θ^X_m, θ^Y_m and all (hyper-)priors is displayed in figure 3 for a Gaussian functional regressor X and a non-Gaussian scalar response Y.

The outward model in the lower part is parameterized by $\varphi^Y = (\delta^Y, \theta^Y_1, ..., \theta^Y_{M^Y})$ and $\varphi^X = (\delta^X, \theta^X_1, ..., \theta^X_{M^X}, \lambda_X)$, where λ_X denotes the precision of the Gaussian observed noise. The inner regression model in the upper part depends on $\psi^Z = (G^Z, \lambda_{G^Z}, \Lambda_{\theta^Z})$

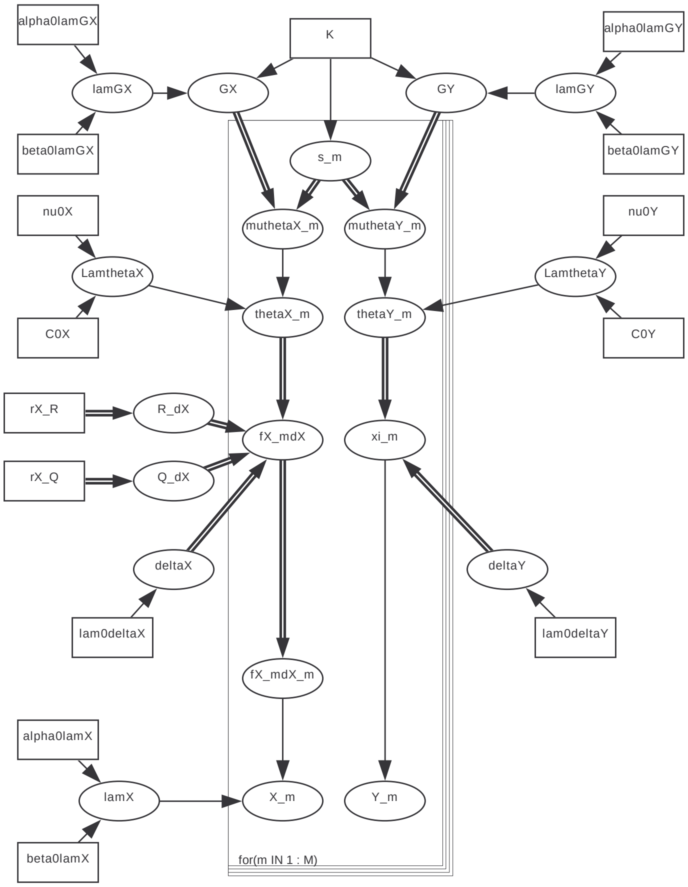

Figure 3: Full hierarchical model for a Gaussian functional X and a non-Gaussian non-functional Y with $M^X = M^Y = M$.

as hyperparameters of θ^Z_m, where $\Lambda_{\theta^Z} = \Sigma^{-1}_{\theta^Z}$ and λ_{G^Z} summarizes precisions of the columns of G^Z, and on the latent variables s_m. For the functional variable X the known matrices IP^X_m, Q_{d^X} and R_{d^X} describe interpolation. The rectangles in the outmost left and right column of the diagram represent mainly known hyperparameters, but also three model parameters occur: K, the number of latent variables ($s_m \in \Re^K$) and r^X_Q, r^X_R, the number of columns of the matrices Q_{d^X}, R_{d^X} respectively, which act as smoothing parameters. (In figure 3 the matrices Q_{d^X}, R_{d^X} are shown as logical nodes depending on the smoothing parameters while in figure 1 they are shown as constants, implicitly assuming that the smoothing parameters r^X_Q, r^X_R are given.)

In the sequel the hierarchical model is specified in detail comprising (conjugate) standard distributions like Gaussian ($N(\mu, \sigma^2)$), Gamma - ($\Gamma(\alpha, \beta)$), Wishart - ($W(\nu, C^{-1})$) and Multinomial ($Mult(J, \pi)$) distributions. Related to Gaussian distributions variances are denoted by σ^2, covariance matrices by Σ and precisions by λ respectively Λ. The parametrization of the Gamma distribution is such that the mean is given by α/β, and that of the Wishart-distribution such that the mean is given by νC^{-1}. To keep the notation interpretable the following rules are applied: parameters of distributions occurring more than once are indexed by the corresponding variable. Analogous assumptions for X and Y are often formulated only once, indexed by Z, always meaning $Z \in \{X, Y\}$. Fixed values of the parameters at the highest level of hierarchy are marked with a superscript "0", parameters in posterior distributions are marked with an asterisk (as superscript).

2.2 Model for the functional predictor

Each x_m represents a curve f^X_m recorded possibly with noise at a design $d^X_m = \{t^X_{1m}, ..., t^X_{N^X_m m}\}$. As an error model a one-parameter exponential family with canonical parameters ξ^X_{im} equal to function values, $\xi^X_{im} = f^X_m(t^X_{im})$, is assumed, that is for $x_m = (x_{1m}, ..., x_{N^X_m m})^T$

$$p^X(x_{im}|f^X_m(t^X_{im})) = a^X(x_{im}) \exp(x_{im} f^X_m(t^X_{im}) - b^X(f^X_m(t^X_{im})), \qquad (1)$$

$i = 1, ..., N^X_m$, and conditionally on $f^X_{md^X_m}$ the variables X_{im} are independent.

The functions f^X_m, $m = 1, ..., M^X$, are represented as interpolation splines with respect to a *common* design d^X with N^X points $t^X_1, ..., t^X_{N^X}$, and hence are characterized by the function values at the design d^X, $f^X_m = h_{I(f^X_{md^X})} \in H_{I(d^X)}$, say. The vector of function values $f^X_{md^X_m}$ at the subject-specific design d^X_m is obtained from the vector of function values at the common design by multiplication with an interpolation matrix: $f^X_{md^X_m} = IP^X_m f^X_{md^X}$. The function f^X_m may be regarded as a logit-transformed probability function of a Bernoulli process or a log-intensity function of a Poisson process or a mean function of a Gaussian process to name the most popular cases.

Furthermore the interpolation splines f^X_m are decomposed into a mean function and a residual function which again are interpolation splines. Hence $f^X_{md^X} = c^X + \rho^X_m$, say, induces $f^X_m = h_{I(f^X_{md^X})} = h_{I(c^X)} + h_{I(\rho^X_m)}$, where $h_{I(c^X)}$ denotes the mean function and $h_{I(\rho_m^X)}$ denotes the residual function. Both functions are expanded in the first r_Q^X respectively r_R^X functions of a Demmler-Reinsch like basis of interpolation splines such that for the vector of function values at the common design d^X

$$c^X = Q_{d^X}\delta^X, \quad \rho_m^X = R_{d^X}\theta_m^X \tag{2}$$

holds. Here the r_Q^X columns of Q_{d^X} are the vectors of function values of the basis functions and similarly for R_{d^X}. For details of the construction of Q_{d^X} and R_{d^X} see (van der Linde, 2009). In the terminology of generalized linear models the decomposition $f_{md_m^X}^X = IP_m^X Q_{d^X}\delta^X + IP_m^X R_{d^X}\theta_m^X$ may be seen as defining a predictor using the canonical link function.

If for example the noise model is Gaussian with homogeneous errors,

$$X_m|\delta^X, \theta_m^X, \sigma_X^2 \sim N(IP_m^X(Q_{d^X}\delta^X + R_{d^X}\theta_m^X), \sigma_X^2 I_{N_m^X}), \tag{3}$$

and as in generalized linear models the Normal distribution is regarded as a belonging to a one-parameter exponential family even if σ_X^2 is unknown.

If the functions f_m^X respectively vectors of function values $f_{md_m^X}^X$ are observed without noise, the assumption (3) will still be made because the representation of f_m^X in the truncated basis is an approximation. Thus for observations with Gaussian noise σ_X^2 comprises an approximation error and a measurement error, for exact observations σ_X^2 represents an approximation error only and can be expected to take smaller values. The impact of an assumption of errors analogous to (3) even if observations are without measurement error was studied by Klami and Kaski (2008) for CCA. They demonstrated convergence with decreasing σ_X^2 in a simulation study.

2.3 Models for the response

For the response variables Y_m three options are considered.

(i) The vectors y_m, $m = 1, \ldots, M^Y$, represent functions f_m^Y and assumptions analogous to those about X_m are then made, changing only the superscript in notation. In this case a *functional response model* is set up.

(ii) The variables Y_m are univariate, distributed according to a one-parameter exponential family,

$$p(y_m|\xi_m^Y) = a^Y(y_m)\exp(y_m\xi_m^Y - b^Y(\xi_m^Y)) \tag{4}$$

where $\xi_m^Y = \delta^Y + \theta_m^Y$. This can be seen as a special simple case of (i) with $N^Y = N_m^Y = 1$, $r_Q^Y = r_R^Y = 1$, $Q_{d^Y} = R_{d^Y} = 1$. Again a special case would be

$$Y_m \sim N(\delta^Y + \theta_m^Y, \sigma_Y^2) \tag{5}$$

even if σ_Y^2 is unknown. This set-up is relevant for scalar prediction and for the classification of curves into two groups based on functional covariates using working observations.

(iii) Classification in more than two groups can be described with Y_m multinomially distributed, $Y_m \sim Mult(J, \pi_m^Y)$, $J > 2$. Equation (4) can be generalized to a multiparameter exponential family with canonical parameter ξ_m^Y with $\xi_{jm}^Y = \log(\pi_{jm}^Y/(1 - \sum_{k=1}^{J-1} \pi_{km}^Y))$ for $j = 1, ..., J-1 = N_m^Y$. In the sequel, also the alternative "softmax parameterization" in η_m^Y with $\pi_{jm}^Y = \exp(\eta_{jm}^Y)/\sum_{i=1}^J \exp(\eta_{im}^Y)$ and the decomposition $\eta_m^Y = \delta^Y + \theta_m^Y$, $\delta^Y, \theta_m^Y \in \Re^J$ will be used where $N_m^Y = J$.

Obviously other set-ups, particularly response vectors with partial vectors fitting into (i)-(iii) would be similar, but are not considered here. Also, multivariate responses distributed according to general multiparameter exponential families are beyond the scope of this paper.

A final assumption is that conditional on the canonical parameters $\xi_m^X = f_{md_m^X}^X$, ξ_m^Y (and σ_X^2, σ_Y^2 if applicable) X_m and Y_m are stochastically independent. Their marginal dependence is captured in the relation of the (centered) canonical parameters specified in an inner model.

In summary, the likelihood given data D is a function of the parameters $\varphi = (\varphi^X, \varphi^Y)$ where $\varphi^Z = (\delta^Z, \Theta^Z)$ or $\varphi^Z = (\delta^Z, \Theta^Z, \lambda_Z)$ with $\Theta^Z = (\theta_1^Z, ..., \theta_{M^Z}^Z)$, $\lambda_Z = \sigma_Z^{-2}$, $Z \in \{X, Y\}$. The likelihood is given by

$$p(D|\varphi) = \prod_{m=1}^{M^X} p(x_m|\varphi^X) \prod_{m=1}^{M^Y} p(y_m|\varphi^Y). \tag{6}$$

The prior on φ is factorized as

$$p(\varphi) = \prod_{Z \in \{X,Y\}} p(\delta^Z) p(\lambda_Z) p(\Theta^X, \Theta^Y) \tag{7}$$

and completed assuming

$$\delta^Z \sim N(0, (\lambda_{\delta^Z}^0)^{-1} I_{r_Q^Z}) \tag{8}$$

and

$$\lambda_Z \sim \Gamma(\alpha_{\lambda_Z}^0, \beta_{\lambda_Z}^0). \tag{9}$$

The prior of (Θ^X, Θ^Y) is specified by an inner regression model.

2.4 The inner regression model

The inner regression model is a Gaussian reduced rank regression model with latent variables for the coefficients of the centered canonical parameters $\xi_m^Z - Q_{d^Z}\delta^Z = R_{d^Z}\theta_m^Z$.

$$\begin{aligned} \theta_m^X | G^X, \Sigma_{\theta^X}, s_m &\sim N(G^X s_m, \Sigma_{\theta^X}), \quad G^X \in \Re^{r_R^X \times K},\ \Sigma_{\theta^X} \in \Re^{r_R^X \times r_R^X}, \quad (10) \\ \theta_m^Y | G^Y, \Sigma_{\theta^Y}, s_m &\sim N(G^Y s_m, \Sigma_{\theta^Y}), \quad G^Y \in \Re^{r_R^Y \times K},\ \Sigma_{\theta^Y} \in \Re^{r_R^Y \times r_R^Y}. \quad (11) \end{aligned}$$

Note that G^X (G^Y) and the latent variables $S = (s_1, ..., s_{M^X})$, $s_m \in \Re^K$ are determined only up to an invertible linear transformation. The problems resulting from non-identifiability and the interpretation and use of this inner model which depends on the response will be discussed in section 3 and for special examples in section 4. The key assumption is that conditionally on the latent variable S the Θ^X, Θ^Y are independent, more precisely that

$$p(\Theta^X, \Theta^Y | G^X, \Lambda_{\theta^X}, G^Y, \Lambda_{\theta^Y}, S) = \prod_{Z \in \{X,Y\}} p(\Theta^Z | G^Z, \Lambda_{\theta^Z}, S), \tag{12}$$

(where, remember, $\Lambda_{\theta^Z} = \Sigma_{\theta^Z}^{-1}$). This part of the prior specification for the parameters within the hierarchical model is detailed by

$$p(\Theta^Z | G^Z, \Lambda_{\theta^Z}, S) = \prod_{m=1}^{M^Z} p(\theta_m^Z | G^Z, \Lambda_{\theta^Z}, s_m). \tag{13}$$

In $p(\theta_m^Z | G^Z, \Lambda_{\theta^Z}, s_m)$ two hyperparameters, G^Z and Λ_{θ^Z}, occur. The hyperprior for G^Z is specified by independent Gaussians for the columns γ_k^Z of G^Z, $k = 1, ..., K$,

$$\gamma_k^Z | \lambda_{\gamma_k^Z} \sim N(0, \lambda_{\gamma_k^Z}^{-1} I_{r_R^Z}), \tag{14}$$

and

$$\lambda_{\gamma_k^Z} \sim \Gamma(\alpha^0_{\lambda_{\gamma_k^Z}}, \beta^0_{\lambda_{\gamma_k^Z}}). \tag{15}$$

All precisions $\lambda_{\gamma_k^Z}$ are collected in a vector $\lambda_{G^Z} = (\lambda_{\gamma_1^Z}, ..., \lambda_{\gamma_K^Z})$.

The hyperprior for an unrestricted precision matrix is a Wishart distribution

$$\Lambda_{\theta^Z} \sim W(\nu_Z^0, (C_Z^0)^{-1}). \tag{16}$$

If the precision matrix is assumed to be diagonal, $\Lambda_{\theta^Z} = \lambda_{\theta^Z} I_{r_R^Z}$, the hyperprior is given by

$$\lambda_{\theta^Z} \sim \Gamma(\alpha^0_{\lambda_{\theta^Z}}, \beta^0_{\lambda_{\theta^Z}}). \tag{17}$$

In summary, the parameters φ are augmented by hyperparameters $\psi = (\psi^X, \psi^Y)$, $\psi^Z = (G^Z, \lambda_{G^Z}, \Lambda_{\theta^Z})$, and the prior on all parameters factorizes with respect to Z conditionally on S,

$$p(\varphi, \psi | S) = \prod_{Z \in \{X,Y\}} p(\varphi^Z | \psi^Z, S) p(\psi^Z) \tag{18}$$

where

$$p(\psi^Z) = p(\Lambda_{\theta^Z}) \prod_{k=1}^{K} p(\gamma_k^Z | \lambda_{\gamma_k^Z}) p(\lambda_{\gamma_k^Z}). \tag{19}$$

2.5 Distribution of the latent variables

The $K \times M^X-$ matrix S comprises the latent vectors for each of the M^X observed individuals which are assumed to be independent,

$$p(S) = \prod_{m=1}^{M^X} p(s_m).$$

Furthermore, the key assumption in the latent variable model is that the components of s_m are independent,

$$p(s_m) = \prod_{k=1}^{K} p(s_{km}). \tag{20}$$

The simplest distributional assumption about s_{km} is $s_{km} \sim N(0,1)$, which amounts to

$$s_m \sim N(0, I_K). \tag{21}$$

To be more flexible particularly in functional prediction it is sometimes of interest to overcome the framework of Gaussian distributions specifying the distribution of s_{km} as mixture of Gaussians. A similar approach in independent factor analysis (Choudrey et al. 2000 ; Choudrey and Roberts, 2001) could be extended to the class of models considered here, but is not studied in this paper.

2.6 Model parameters

Some parameters characterizing the model are still unspecified: the number of latent variables K and the smoothing parameters for the mean functions r_Q^Z and the residual functions r_R^Z. The choice of these parameters is discussed in section 3.3.

2.7 Special cases

In the sequel four types of analysis will be applied, corresponding to whether assumption (16) or (17) holds for Λ_{θ^Z}. They are abbreviated as in the following table:

		Λ_{θ^X}	
		(16)	(17)
Λ_{θ^Y}	(16)	$FCCA$	$INDX$
	(17)	$INDY$	$INDXY$

If Z, often the response Y, is univariate both (16) and (17) specify a Gamma distribution. Yet the analyses FCCA and INDY usually do not coincide, because typically the hyperparameters are chosen such that different Gamma distributions result. The model INDXY has the "supervised probabilistic PCA" introduced by Yu et al. (2006) as special case.

In order to illustrate the versatility and practical relevance of the proposed model some important special cases which have been discussed in FDA are singled out.
1. *Functional canonical correlation analysis* (analyzed using FCCA, with $M^X \geq M^Y$)
2. *Functional prediction* (response type (i), analyzed using FCCA or INDXY with $M^X > M^Y$)
3. *Scalar prediction* (response type (ii), analyzed using INDY or INDXY with $M^X > M^Y$)

4. *Functional discriminant analysis* or *classification* (response type (iii), analyzed using FCCA, INDY or INDXY with $M^X > M^Y$).

In the technical report (van der Linde, 2010) for each special case more comprehensive reviews of the (mostly frequentist) literature are given pointing in more detail to methodological issues and fields of application.

3 Inference

A variational Bayesian approach, also called an ensemble learning approach in the machine learning community, will be taken to derive an approximate posterior distribution q for all quantities of interest Ξ. It has to be adapted to the proposed model and its special cases which can be quite tedious, but main ingredients are known and have been tried in multivariate (sub)models already. The main algorithms are combined with ideas applying to special cases of FDA: in particular the proposal of Archambeau et al. (2006) of how to obtain patterns of covariation in CCA based on latent variables is needed, and it has to be spelled out explicitly how to obtain predictions of the response variable. If observed variables are non-Gaussian, Gaussian approximations based on working observations as in generalized linear models (van der Linde, 2009) or based on pseudo observations as introduced for classification by Bouchard (2007) can be used. They are described in this section in order to prepare a comparison of the two approaches by example in section 4. Also, the choice of the model parameters turns out to be crucial and several options are discussed. Thus in this chapter contributions from different special sources in the literature are evaluated and integrated to develop a strategy of inference for FDA with latent variables.

The main idea of variational inference is to maximize a lower bound L_q of the marginal log-density of the data D, $\log p(D) \geq L_q$ with L_q induced by the approximate posterior density q. By assumption q is factorized into parametric densities, here

$$q(\Xi) = q(S) \prod_{Z\in\{X,Y\}} q(\delta^Z) q(\theta_1^Z, ..., \theta_{M^Z}^Z) q(\lambda_Z) q(\Lambda_{\theta^Z}) \prod_{k=1}^{K} q(\gamma_k^Z) q(\lambda_{\gamma_k^Z}), \tag{22}$$

and further factorizations over m in $q(\theta_m^Z)$ and $q(s_m)$ result from prior independence. Using conjugate priors algorithms in closed form for iterative updates of the factors increasing the lower bound L_q can be obtained. The main algorithms for functional reduced rank regression with $M^X \geq M^Y$ used in the examples are given in the appendices 1 and 2. The initialization of all parameters necessary to start the updates (which was the same for all examples) is described in appendix 3. Also, default (hyper-) parameters are specified there. Variational algorithms similar to those required for functional reduced rank regression were developed by Wang (2007) (for multivariate CCA) and by van der Linde (2008, 2009) (for functional PCA), and the reader is referred to these papers for details not repeated here.

Variational algorithms will be given for the case of Gaussian errors in observing the functional predictor and the response variable. If instead (1) or (4) hold as non-Gaussian error models $p(D|\Xi)$ can be approximated using the "working Gaussian density" for "working observations". These are defined by

$$w^Z(\xi^{0Z}_{im}) = \xi^{0Z}_{im} + [(b^Z)''(\xi^{0Z}_{im})]^{-1}(z_{im} - (b^Z)'(\xi^{0Z}_{im}))$$

$i = 1, ..., N^Z_m$, $m = 1, ..., M^Z$, where if Z is functional $\xi^{0Z}_{im} = f^{0Z}_m(t^Z_{im})$ is a point of expansion in a second order Taylor approximation of $\log p(z_{im}|f^Z_m(t^Z_{im}))$. In terms of vectors $\xi^{0Z}_m = f^{0Z}_{md^Z_m} = IP^Z_m f^{0Z}_{md^Z}$, and the working distributional assumption is

$$w^Z_m(f^{0Z}_{md^Z}) \sim N(IP^Z_m f^Z_{md^Z}, diag([(b^Z)''(f^{0Z}_m(t^Z_{im}))]^{-1}). \tag{23}$$

In this way an approximating working Gaussian distribution with known covariance matrix can be used - given the point of expansion which may be updated iteratively. Thus stepwise the variational algorithm is simplified in that the updates of λ_Z can be omitted. Overall though, it is more expensive because the model parameters have to be optimized not only once but alternatingly with each point of expansion. The approach was investigated in detail for FPCA by van der Linde (2009) and is applied analogously.

In case of functional discriminant analysis where Y is binary or more generally multinomial another Gaussian approximation to the likelihood can be used. Bouchard (2007) suggested a lower bound inducing a pseudo Gaussian approximation which can be expected to give better results than the working observations because the latter yield an *approximate* lower bound only. His algorithm is based on the "softmax parameterization" in $\eta^Y_{jm} = \log(\pi^Y_{jm})$ (see section 2.3). His approach involves variational parameters $a_m \in \Re$ and $\zeta_m \in \Re^J$, $m = 1, ..., M^Y$ which are optimized with each iteration to make the lower bound as tight as possible.

Technically pseudo observations with a pseudo Gaussian distribution

$$v_m \sim N(IP^Y_m f^Y_{md^Y}, \Sigma_{v_m}) \tag{24}$$

are defined by

$$v_m = \Sigma^{-1}_{v_m}(y_m - b_m) \in \Re^J \tag{25}$$

where

$$\Sigma_{v_m} = 2diag(g(\zeta_{jm})), \tag{26}$$

$$g(u) = \frac{1}{2u}(\frac{1}{1+e^{-u}} - \frac{1}{2}), \tag{27}$$

and

$$b_{jm} = 0.5 - 2a_m g(\zeta_{jm}) \tag{28}$$

for $j = 1, ..., J$. The update rules are summarized in appendix 4.

Variational inference provides an approximate posterior distribution only. The examples with simulated data demonstrate that the approach yields good point estimates and pointwise credible regions covering the true function values even if their width might be underestimated. A more detailed assessment of the accuracy of the approximation, especially of posterior covariances, requires a comparison with sampled posterior distributions. A systematic comparison and in particular the investigation of MCMC based model choice (as in the paper by Lopes and West (2004) for FA) is beyond the scope of this paper. However, the model can be easily implemented in WinBUGS and the estimates obtained by variational inference used to initialize sampling. WinBUGS is not well suited to multivariate problems (nodes) and as a multipurpose software possibly not efficient for the RRR model. But it is an immediate option for practitioners, and can be applied to obtain a first impression of possible improvements over variational inference by MCMC. Some experiences with the implementation of (F)CCA in WinBUGS are reported in section 4.

3.1 Functional canonical correlation analysis (FCCA)

CCA is a standard technique, described in any textbook on multivariate analysis (e.g. Mardia et al., 1995), to analyze the (linear) dependence between two zero mean random vectors $\theta^X \in \Re^{r^X_R}$, $\theta^Y \in \Re^{r^Y_R}$ with covariance matrices Γ^X, Γ^Y and cross-covariance matrix Γ^{XY}. (Here we consider the marginal distribution rather than the conditional distribution (10),(11), that is, $\Gamma^{XY} = G^X(G^Y)^T$ and $\Gamma^Z = G^Z(G^Z)^T + \Sigma_{\theta^Z}$, $Z \in \{X, Y\}$.) A major aim of CCA can be the identification of "patterns of covariation" represented by the (pairs of) columns of a matrix A^Z, $Z \in \{X, Y\}$, which form basis vectors in the reconstruction of θ^Z. A meteorological example of nicely interpretable patterns of covariation is given by von Storch and Zwiers (1999, ch.14). Functional patterns of covariation can easily be visualized and thus help interpretation. The main ideas of dimension reduction and of extracting patterns of covariation based on a RRR are briefly outlined below.

Technically CCA is a de-correlation achieved by a singular value decomposition of $(\Gamma^X)^{-1/2}\Gamma^{XY}(\Gamma^Y)^{-1/2} = V^X P (V^Y)^T$ where V^X, V^Y are orthonormal and P is diagonal with entries κ_k. The $K-$ dimensional column vectors u^Z_k of $U^Z = (\Gamma^Z)^{-1/2}V^Z$, $Z \in \{X, Y\}$, then form canonical weight vectors, and canonical variates are obtained as $S^Z_{(k)} = (u^Z_k)^T\theta^Z$, $k = 1, ..., K$. The K pairs of canonical variates can be ordered according to the size of the canonical correlation coefficients $\kappa_k = cov(S^X_{(k)}, S^Y_{(k)})$, $k = 1, ..., K$.

Dimension reduction is based on the projection of θ^Z onto the space spanned by canonical variates $S^Z_{(k)}$, $k = 1, ..., K$. Thus, with the notation $S^Z = (S^Z_{(1)}, ..., S^Z_{(K)})^T$,

$$\theta^Z_m \approx A^Z s^Z_m, \qquad where \quad A^Z = \Gamma^Z U^Z. \tag{29}$$

Each feasible G^Z (satisfying $\Gamma^Z = G^Z(G^Z)^T + \Sigma_{\theta^Z}$ and $\Gamma^{XY} = G^X(G^Y)^T$) and A^Z are related by an invertible transformation,

$$G^Z = A^Z L^Z \tag{30}$$

with $L^X(L^Y)^T = diag(\kappa_k)$. Defining $B_Z = I_K + (G^Z)^T\Lambda_{\theta^Z}G^Z$ and the columns of W by the eigenvectors of $(I_K - B_X^{-1})(I_K - B_Y^{-1})$, Archambeau et al. (2006) demonstrate that

$$L^Z = W^T(I_K - B_Z^{-1})^{1/2}. \tag{31}$$

Hence A^Z can be reconstructed from all feasible matrices G^Z in the RRR model. Note that

$$G^Z s_m = A^Z L^Z s_m = A^Z \bar{s}_m^Z \tag{32}$$

for $\bar{s}_m^Z = L^Z s_m$ with (prior) $cov(\bar{s}^X, \bar{s}^Y) = diag(\kappa_k)$ and that in contrast to (29) with two latent variables only one common latent variable is specified in (10),(11) respectively (32). For further discussion and model variants see (Bach and Jordan, 2005).

In the latent variable approach to FCCA patterns of covariation can be inferred if the joint covariance matrix Γ is known. Thus realizations from the (approximate) posterior distributions of G^Z and Σ_{θ^Z} can be used to obtain by simulation a sample of patterns of covariation and to investigate their variability. This is computationally expensive because singular value decompositions are required for each realization. An approximating shortcut is to use an estimate $\widehat{L}^Z$ of L^Z and to refer to the approximate Gaussian distribution of the columns of A^Z, $a_k^Z = G^Z(\widehat{L}^Z)_k^{-1}$, neglecting the uncertainty in the estimated k-th column $(\widehat{L}^Z)_k^{-1}$ of $(\widehat{L}^Z)^{-1}$ In this respect the probabilistic approach to (F)CCA is merely a special approach to inference about the covariance matrix, the singular value decomposition of which is not incorporated in the parametrization. The problem similarly occurs in FPCA where principal modes of variation in a set of curves are identified only ex post. (Compare the discussion of rotations in (van der Linde, 2008) and further comments in section 5.)

3.2 Prediction

In this section the problem of predicting an unobserved $\widetilde{f}_m^Y$ (resp. $\widetilde{\theta}_m^Y$) from an observed $\widetilde{X}_m$ using the inner regression model with different specifications of $\Sigma_{\theta^X}, \Sigma_{\theta^Y}$ is investigated.

Assume that M pairs of curves have been observed and that another m^X $X-$ curves (that is vectors $\widetilde{x}_m$ with observations $\widetilde{x}_{nm}$, $n = 1, ..., \widetilde{N}_m^X$, $m = 1, ..., m^X$) are given for which for example the functions $\widetilde{f}_m^Y$ are to be predicted. Thus the numbers of observed curves are unequal: $M^X = M + m^X > M^Y = M$. The key idea in predicting an unknown $Y-$ curve $\widetilde{f}_m^Y = h_{I(c^Y)} + h_{I(\widetilde{\rho}_m^Y)}$ from noisy $X-$ values is to infer $\widetilde{s}_m$ from $\widetilde{X}_m$ via $\widetilde{\theta}_m^X$ and to reconstruct $\widetilde{\rho}_m^Y = R_{d^Y}\widetilde{\theta}_m^Y$ via $\widetilde{\theta}_m^Y$ from $\widetilde{s}_m$. Rish et al. (2008) use (maximum likelihood) point estimates of $\widetilde{s}_m$ and parameter estimates in their predictive model thus neglecting the estimation error. Attias (1999) discusses (non-functional) prediction in latent variable models from a Bayesian point of view pointing to the use of an augmented data set to derive a predictive distribution. This technique will be applied here.

It is sufficient to find the predictive distribution of $\widetilde{f}_{md^Y}^Y = Q_{d^Y}\delta^Y + R_{d^Y}\widetilde{\theta}_m^Y$. The posterior distribution of all parameters in the model given $\widetilde{x} = \{\widetilde{x}_m | m = 1, ..., m^X\}$ and the data $D' = \{(x_m, y_m) | m = 1, ..., M\}$ can easily be obtained running the variational algorithm. In this way the posterior distribution of $\widetilde{s}_m$ is determined by $\widetilde{x}_m$ only and is used to infer moments of the posterior predictive distribution of $\widetilde{\theta}^Y_m$. Notice that only θ^Y_m, $m = 1, ..., M$ occur as parameters in the likelihood function. And although the conditional distribution of $\widetilde{\theta}^Y_m$ is Gaussian, $\widetilde{\theta}^Y_m | G^Y, \widetilde{s}_m, \Sigma_{\theta^Y} \sim N(G^Y \widetilde{s}_m, \Sigma_{\theta^Y})$ due to the product in the mean the predictive distribution is not Gaussian. But using the approximate factorized posterior distribution we have

$$E(\widetilde{\theta}^Y_m | \widetilde{x}, D') = E(G^Y | \widetilde{x}, D') E(\widetilde{s}_m | \widetilde{x}, D'), \tag{33}$$

$$\begin{aligned} & cov(\widetilde{\theta}^Y_m | \widetilde{x}, D') \\ = \; & E(\Sigma_{\theta^Y} | \widetilde{x}, D') + cov(G^Y \widetilde{s}_m | \widetilde{x}, D') \\ = \; & E(\Sigma_{\theta^Y} | \widetilde{x}, D') + E_{\widetilde{s}_m | \widetilde{x}_m, D'}(cov(G^Y \widetilde{s}_m | \widetilde{s}_m, \widetilde{x}, D')) && (34) \\ & + cov_{\widetilde{s}_m | \widetilde{x}_m, D'}(E(G^Y \widetilde{s}_m | \widetilde{s}_m, \widetilde{x}, D')) && (35) \end{aligned}$$

and

$$cov(\widetilde{f}^Y_{md^Y} | \widetilde{x}, D') = Q_{d^Y}(cov(\delta^Y | \widetilde{x}, D')) Q^T_{d^Y} + R_{d^Y}(cov(\widetilde{\theta}^Y_m | \widetilde{x}, D')) R^T_{d^Y} \tag{36}$$

where posterior independence of δ^Y and $(G^Y \widetilde{s}_m)$ follows from the assumed factorization (22).

Classification of the m-th subject given curve X_m into one of J groups is based on estimated respectively predicted probabilities $\widehat{\pi}^Y_{jm}$. They are obtained as $\widehat{\pi}^Y_{1m} = logit^{-1}(\widehat{\xi}^Y_m)$ for $J = 2$, where $\widehat{\xi}^Y_m = \widehat{\delta}^Y + \widehat{\theta}^Y_m$ (and $\widehat{\pi}^Y_{2m} = 1 - \widehat{\pi}^Y_{1m}$) if working observations are used. For $J \geq 2$, if Bouchard's (2008) approach is employed, $\widehat{\pi}^Y_{jm} = \exp(\widehat{\eta}^Y_{jm}) / \sum^J_{i=1} \exp(\widehat{\eta}^Y_{im})$, where $\widehat{\eta}^Y_m = \widehat{\delta}^Y + \widehat{\theta}^Y_m$, and in the $J-$dimensional softmax parametrization the parameters are dependent across j. All estimates/predictions are taken as posterior (predictive) means. The m-th curve is assigned to the group with maximum $\widehat{\pi}^Y_{jm}$.

3.3 Model choice

There are five model parameters: K, r^Z_Q, r^Z_R, $Z \in \{X, Y\}$, with $K \leq \min\{r^X_R, r^Y_R\}$ in FCCA. A standard approach to model choice in variational inference is the maximization of the lower bound L_q, approximating the maximization of the marginal likelihood of the data. It can be applied in all special cases except in classification if Bouchard's approach with adaptively optimized variational parameters is used. Model search based on the lower bound is prior predictive. It might be objected that in functional prediction a prior predictive criterion for model choice (like the marginal likelihood) is less appropriate than a posterior predictive criterion (like DIC or a modification thereof). However, here the model is block-bivariate in X and Y, and given $\widetilde{X}$ one is half way between prior and posterior prediction. Hence prior predictive model choice can still be justified.

Given K *smoothing parameters* were chosen maximizing the lower bound L_q by searching over a range of candidate values. With each update the values of L_q can be calculated and the final updates yield the values to be compared. Note that thus the approach is an empirical Bayesian approach, and the uncertainty about model parameters is not taken into account when assessing the precision of parameter estimates on a lower level of model hierarchy. For fixed K the pairs of smoothing parameters (r_Q^X, r_Q^Y) respectively (r_R^X, r_R^Y) are optimized in alternating searches where the range of candidate values depends on the data set. As in functional PCA (van der Linde, 2008) an exploratory smoothing of discretized mean functions or of a subset of discretized curves (if available) is applied to initialize the smoothing parameters, and new smoothing parameters are searched in the neighborhood of old parameters.

Based on the examples of RRR tried no conclusive recommendation how to choose the *number of latent variables* K can be given. In principle there are several options.

(i) Maximization of the lower bound

This approach did not work for FCCA and INDY. In all examples the lower bound was almost decreasing with increasing K such that $K = 1$ was suggested throughout. I suppose this is due to the use of the Wishart distribution. (With growing K, Λ_{θ^Z} becomes larger and in turn the entropy term of s_m in the lower bound decreases. This is not sufficiently compensated by the entropy term of Λ_{θ^Z}.) If instead in a predictive model $\Lambda_{\theta^Z} = \lambda_{\theta^Z} I_{r_R^Z}$ was assumed and a Gamma distribution for λ_{θ^Z} was used the choice of K based on the lower bound did yield reasonable results in prediction. Thus the lower bound as a criterion to determine K can be used for INDXY but not for FCCA and INDY.

(ii) Automatic Relevance Determination (ARD)

For fixed large K columns of G^Z with small (estimated) prior variance, that is values close to zero, are discarded (cp. eq.(14), $\lambda_{\gamma_k^Z} \to \infty$). For ARD Bayesian estimates of $\lambda_{\gamma_k^Z}$ are obtainable and (in all tried examples) did indicate a jump in size at a preferable number of latent variables, but the cutpoint is to be chosen subjectively. ARD can be used in all models studied here.

(iii) Canonical correlation coefficients in FCCA

Another option in FCCA is to choose K as the number of estimated canonical correlation coefficients κ_k greater than 0.5. These estimates are obtained as a by-product of the transformation introduced by Archambeau et al. (2006), $G^Z = A^Z L^Z$, to extract patterns of covariation where $L^X(L^Y)^T = diag(\kappa_k)$.

(iv) Fit and prediction

In *scalar prediction* the mean square error of fit in the test set with M^Y observations can be evaluated, for the *classification* of functions the minimum misclassification rate in the test set provides a measure of performance. These criteria did work in the examples, but are ad hoc from a theoretical point of view. They are based on measures of fit without taking into account model complexity which protects against overfitting. This deficiency could be overcome by cross-validation which is, however, computationally demanding. In the examples only prediction errors for known (simulated) function values in the validation set with m^X observations are evaluated.

Model search approached in this way is not yet satisfying. In summary, the experience to be reported is, that in many regression contexts the lower bound could not be used to determine K and alternatives were needed. For several examples extensive tables are given to compare the performance of the different applicable criteria for the choice of K. The corroboration of the alternative strategies has to be left to further research, though.

4 Examples

The inner regression model (10),(11) specifies a structure of dependence with (10) modelling "errors in variables" θ^X. FCCA with Σ_{θ^X}, Σ_{θ^Y} unrestricted, used to infer patterns of covariation will be discussed in section 4.1. The set-up of the first example with simulated data will also be used subsequently to illustrate other types of analysis. The functional response model used for prediction of f^Y_m given X_m will be considered in section 4.2 and illustrated with Gaussian and non-Gaussian synthetic observations. The special case of scalar prediction will be the topic of section 4.3. Classification based on a functional predictor will be considered in section 4.4 and applied to a problem of speech recognition with a real data set.

4.1 Functional canonical correlation analysis (FCCA)

Two examples of FCCA are presented. In the first example data are generated according to a specified regression between θ^X and θ^Y, that is, the conditional rather than the marginal distribution of θ^Z is used in simulations. Thus the data generating process is close to the RRR model. The example illustrates the extraction of patterns of covariation with a varying amount of information in the sample and also the failure of the maximization of the lower bound as criterion for the choice of the number of latent variables K in FCCA. Alternative choices of K based on the canonical correlation coefficients and on ARD are discussed. In the second example again a synthetic data set is analyzed: a Poisson error scheme and partial designs are features that are difficult to handle in competing (frequentist) approaches but can easily be incorporated in the RRR model.

Example 1

The example is constructed with a *strong linear relationship between θ^X and θ^Y*. A synthetic data set is obtained as follows: $K = 2$ is set and values $s_m \sim N(0, I_2)$ are generated for $m = 1, ..., M = 100$. G^X with $r^X_R = 5$ is chosen as

$$(G^X)^T = \begin{bmatrix} 0.3 & 0.1 & 2 & -1.2 & -1.1 \\ 0.1 & 0.1 & 2 & -1.1 & 1.5 \end{bmatrix}$$

and yields $\theta^X_m = G^X s_m + e^X_m$, where $e^X_m \sim N(0, \Sigma_{\theta^X})$ with

$\Sigma_{\theta^X} = diag(0.1, 0.12, 0.14, 0.16, 0.18)$. Then θ^Y_m is generated according to $\theta^Y_m = W\theta^X_m + e^Y_m$, $e^Y_m \sim N(0, F)$, F tri-diagonal with 0.25 as diagonal entries and 0.09 as entries in the first sub- and super-diagonal. W is set to be

$$W = \begin{bmatrix} 1 & 1.5 & 2 & 1.5 & 1 \\ -0.5 & 1 & -1 & 1 & -0.1 \\ 0.7 & 0.8 & 0.3 & 0.4 & 0.6 \\ 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 2 & 1 & 0 \\ 0.4 & 0.3 & 0.1 & 0.1 & 0.1 \\ 0.4 & -0.3 & -0.1 & -0.1 & 0.1 \end{bmatrix}$$

and thus implies $r_R^Y = 7$. R_{d^X} is defined with $N^X = 35$, R_{d^Y} with $N^Y = 40$ equidistant points in (0,1). (Discretized) mean functions are obtained from $r_Q^X = 3$, $\delta^X = 0.4 * \mathbf{J}_{3\times1}$ and $r_Q^Y = 5$, $\delta^Y = -0.5 * \mathbf{J}_{5\times1}$. Gaussian random noise with $\sigma_X = 0.05$ resp. $\sigma_Y = 0.06$ is added to $f^Z_{md^Z} = Q_{d^Z}\delta^Z + R_{d^Z}\theta^Z_m$. The resulting data set will be referred to as data 1-1. The last twenty pairs of curves along with the mean functions are displayed in figure 4.

Then the data are thinned leaving out randomly 3 subsequent observations in each $X-$ vector (such that $N_m^X = 32$) and 5 randomly chosen single points in each $Y-$ vector (such that $N_m^Y = 35$). Thus the partial designs d_m^X and d_m^Y vary over $m = 1, ..., 100$. Three modifications of this basic data set with partial designs are analyzed. The intention is to try a reduction of sample size as the inner relation in the regression model is strong and to vary the precisions in the model.

1-1/96: The first 96 pairs of curves are retained.

(In section 4.2 below the last 4 will be used to illustrate prediction.)

1-1/16: Only the last 20 pairs of curves are considered, subjecting 16 to

FCCA (and keeping the last 4 for prediction).

1-2/96: The precision is increased multiplying Σ_{θ^X} and F by 0.1. Thus the representation of $X-$ and $Y-$ curves by the coefficients θ^X and θ^Y is improved, and the linear relation between θ^X and $\theta^Y$is strengthened.

The canonical correlation coefficients obtained with $K = 5$ are listed in table 1. The true multivariate CCA refers to the SVD of $cov(\theta^X, \theta^Y) = \Sigma_{\theta^X} W^T$, the sample multivariate CCA to generated vectors θ^Z. FCCA is based on the true smoothing parameters ($r_Q^X = 3, r_Q^Y = 5, r_R^X = 5, r_R^Y = 7$), which are actually found in model search for each K. Maximizing the lower bound w.r.t. K given the optimal smoothing parameters yielded $K = 1$ for each of the three data sets.

While true and multivariately estimated canonical correlation coefficients suggest $K = 4$, FCCA focuses on $K = 3$ if 96 pairs of observations are available and on $K = 2$ if only 16 pairs are in the data set. In table 1 also estimates of the hyperparameters $\lambda_{\gamma_k^Z}$ are given, indicating the (prior) size of the coefficients in G^Z. Interpreting these values in the spirit of automatic relevance determination ("large" values of $\lambda_{\gamma_k^Z}$, that is "small" variances of entries of γ_k^Z, centered at zero, point to irrelevant columns of G^Z) underpins

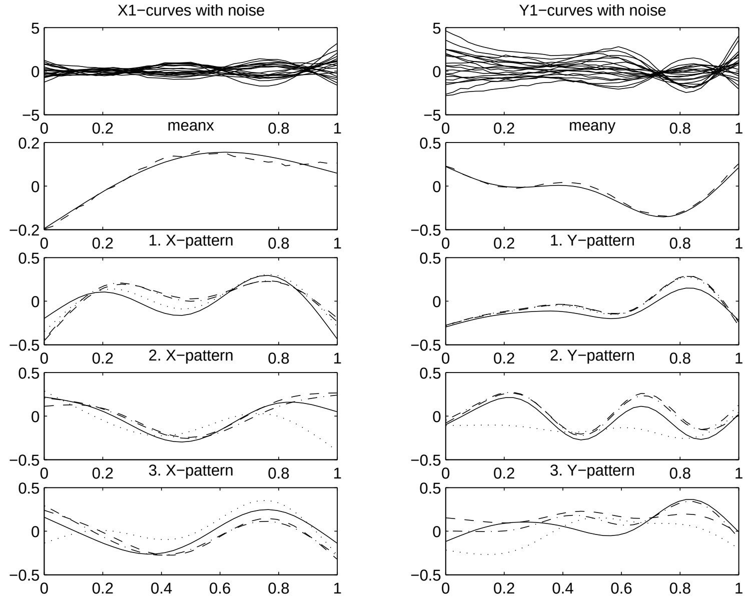

Figure 4: Ex.1: Data 1-1. First row: 20 pairs of noisy curves. Second row: true mean function (solid line) and empirical mean function (dashed line) based on all 100 generated noisy curves. Third to fifth row: first three patterns of covariation; true multivariate (solid line), 1-1/96 (dashed-dotted line), 1-2/96 (dashed line), 1-1/16 (dotted line).

Table 1: Ex.1. True and estimated coefficients of canonical correlation and estimates (posterior means) of $\lambda_{\gamma_k^Z}$.

1-1		κ_1	κ_2	κ_3	κ_4	κ_5
true		0.9986	0.9766	0.8836	0.6103	0.2424
1/96 multiv.		0.9987	0.9772	0.8979	0.6524	0.2268
1/96 FCCA		0.9976	0.9897	0.8123	0.4421	0.0368
1/16 multiv.		0.9998	0.9922	0.8968	0.7452	0.3211
1/16 FCCA		0.9954	0.9909	0.0651	0.0001	0.0000
ARD		λ_{γ_1}	λ_{γ_2}	λ_{γ_3}	λ_{γ_4}	λ_{γ_5}
1/96	X	0.47	0.96	2.31	7.31	7.28
	Y	0.10	0.58	0.72	2.65	4.99
1/16	X	1.14	0.94	131.21	328.30	364.02
	Y	0.14	0.44	107.34	252.01	288.11
1-2		κ_1	κ_2	κ_3	κ_4	κ_5
true		0.9999	0.9964	0.8866	0.6691	0.2446
2/96 multiv.		0.9999	0.9976	0.9006	0.7034	0.2282
2/96 FCCA		0.9992	0.9968	0.9349	0.0081	0.0003
ARD		λ_{γ_1}	λ_{γ_2}	λ_{γ_3}	λ_{γ_4}	λ_{γ_5}
2/96	X	0.40	0.72	2.53	5.28	7.13
	Y	0.11	0.60	0.77	3.08	4.79

the choice of $K = 3$ (with a cutpoint of 1.0 for all three data sets). Although this version of ARD gives the right feeling, the choice of the cutpoint at 1.0 is merely intuitive. The first three patterns of covariation are displayed in figure 4. While in all analyses the first pattern was unambiguously identified, the second and third were sometimes swapped. There is however considerable agreement in the identification of the three dimensional subspaces representing the main joint features of the residual curves.

The model for FCCA was implemented in WinBUGS and run for data 1-2/96 with $K = 3$ and the true smoothing parameters. The model was identified setting $s_{M-K+k} = e_k$ for $k = 1, ..., K$, with e_k denoting the k-th unit vector of dimension K. The posterior means obtained by variational inference were used to initialize the sampler. For comparison, also the initial values of variational inference (in appendix A3.2) were tried. Convergence was reached first for the outer model (filtering of functions), then gradually for the precision parameters $(\Lambda_{\theta^Z}, \lambda_{\gamma_k})$ of the inner model and eventually - if at all - for elements of the matrix G^Z and the values of the latent variables s_m. Mixing for at least some entries of G^Z is poor, and convergence of the values of the latent variables can be doubted even for long chains. In figure 5 estimated functions based on 20000 updates after a burn-in of 10000 samples are displayed. (10000 updates too about 2.5 hours.) Although convergence diagnostics like the MC error or the trace plots are not yet fully convincing, the estimates hardly change if the chain is continued.

While all methods of inference perform equally well in de-noising the functions, the estimation of the mean function by variational inference is slightly improved and the extraction of the first two patterns of covariation is considerably improved by further sampling: The performance of MCMC with initial values as in A3.2 is not yet as satisfying.

In order to examine more closely the inner regression model, variational inference is compared to MCMC in WinBUGS for CCA with simulated coefficients θ_m^Z used as data. In the multivariate model fewer parameters result in reduced computing times: 10000 updates took about 25 minutes. The sampler was run with a burn-in of 120000 samples and another 40000 samples. In figure 6 the resulting patterns of covariance are displayed.

For this example sampling on top of variational inference did not improve the multivariate estimates of the patterns of covariation. Although many more samples were generated than in FCCA convergence diagnostics did not look better and convergence of the values of the latent variables continued to look crucial up to eventually 400000 updates. The estimates remained stable, though.

Example 2

The example illustrates the performance of the proposed approach with simulated non-Gaussian data. The vectors of function values $f^Z_{md^Z}$ are regarded as log-intensities according to which Poisson data were generated.

The mean functions are determined by $r_Q^X = 3$, $\delta^X = (-9.21, 8.28, 3.22)^T$ and $r_Q^Y = 4$, $\delta^Y = (-7.11, 12.08, 0, -4.94)^T$. The coefficients θ_m^Z with $r_R^X = r_R^Y = 4$ were

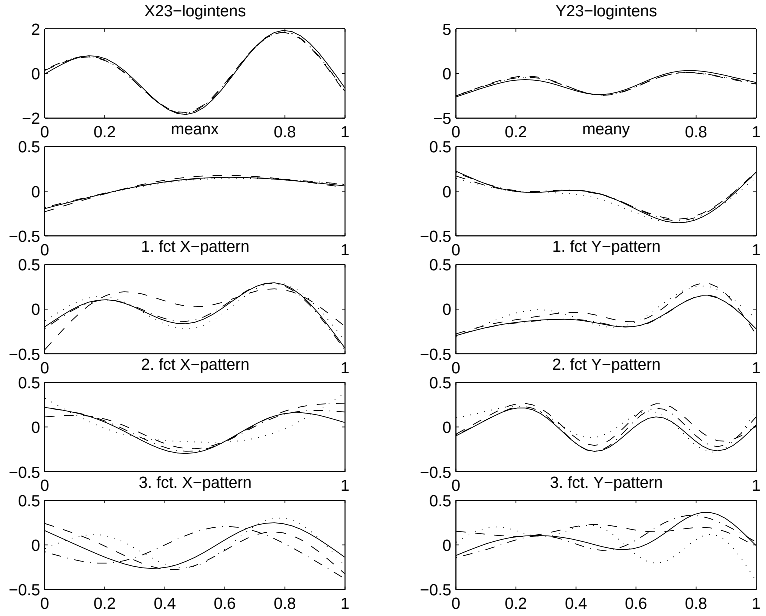

Figure 5: Ex.1. Data set 1-2/96. First row: true curves f_{23}^{Z} (solid lines) and estimates obtained by variational inference in FCCA (dashed lines), MCMC initialized by variational inference in FCCA (dashed-dotted lines), MCMC initialized as in A3.2 (dotted lines). Second row: true and estimated mean functions with line types as in the first row. Third row: first three patterns of covariation; true multivariate (solid line), estimates with line types as in the first row.

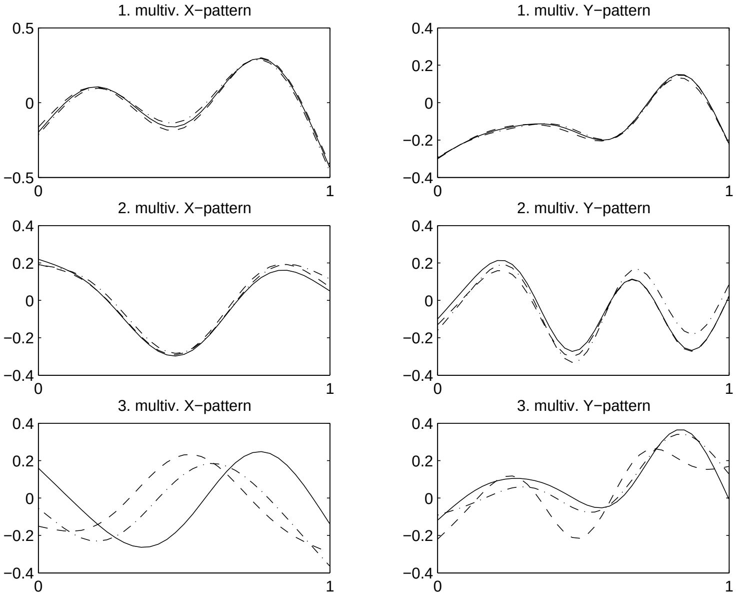

Figure 6: Ex.1. Data set 1-2/96, non-functional analysis: first three patterns of covariation. True patterns (solid lines) and estimates obtained by variational inference in FCCA (dashed lines), MCMC initialized by variational inference in FCCA (dashed-dotted lines).

simulated from a zero mean Gaussian distribution with joint covariance matrix Γ such that $\Gamma(k,k) = 1.5$ for $k = 1, ..., 8,\ \Gamma(k, k+1) = 1$ for $k = 1, ..., 7,\ \Gamma(k, k+2) = 0.7$ for $k = 1, ..., 6$. Thus, by symmetry, Γ has two sub- and super-diagonals, and the remaining entries take the value 0.5. 50 of the resulting $M = 100$ log-intensity functions are displayed in (the first row of) figure 7. The first two canonical correlation coefficients are 0.75 and 0.265, and hence there is one major pattern of covariation. A complete data set ("poi") with 50 observations of each curve and two partial designs were tried: "poipar1" with 10 (X) resp. 15 (Y) randomly selected points omitted and "poipar2" with 10 (X) resp. 15 (Y) connected points left out for each curve.

The (Gaussian) variational algorithm was applied to working observations depending on points of expansion ξ_{im}^{0Z}, $i = 1, ..., N_m = 50$, $m = 1, ..., 100$ in a second order Taylor approximation of the log-likelihood which also yields an approximate lower bound to be used for choosing the smoothing parameters. The points of expansion were obtained from randomly disturbed initial mean functions providing also the initial $r_Q^Z(0)$. Spline smoothing was used to fit the mean functions. Given a Taylor expansion a search for optimal smoothing parameters was carried out fixing (a maximum) $K = 4$, and based on this choice new points of expansion were determined. This was repeated until the model parameters did not change any more, and the number of expansions ("runs") is indicated along with the final results. For more technical details and discussion see (van der Linde, 2009). Alternating Taylor expansions and model choice the computational burden of model choice increases. However, the variational algorithm converges fast, and to illustrate its power, model choice for this example is based on only 5 iterations per analysis.

Table 2: Ex.2. Initial values for model search $r_Q^Z(0)$, estimated first canonical correlation coefficient and relative error in recovering the first pattern of covariation.

data	runs	$r_Q^X(0)$	$r_Q^Y(0)$	can.corr.	relerrX	relerrY
poi	3	7	6	0.7813	0.0259	0.0781
poipar1	3	7	7	0.7737	0.0258	0.0595
poipar2	1	11	11	0.7327	0.0831	0.1010

The results of the three final analyses are summarized in table 2 and figure 7. All analyses suggested one pattern of covariation ($\kappa_1 > 0.5$) and the same smoothing parameters $r_Q^X = 4$, $r_Q^Y = 3$ (true values swapped) and $r_R^X = r_R^Y = 4$ (correctly identified). Also, the canonical correlation coefficient of 0.75 is fairly well estimated by values between 0.73 and 0.78. For all sampling schemes the estimated mean functions deviate from the true one in the same way, obviously due to the simulated data (2. row of figure 7). The analysis with single observations left out (poipar1) hardly differs from the one with complete data (poi), but omitting segments of functions (poipar2) deteriorates the reconstruction of the pattern of covariation.

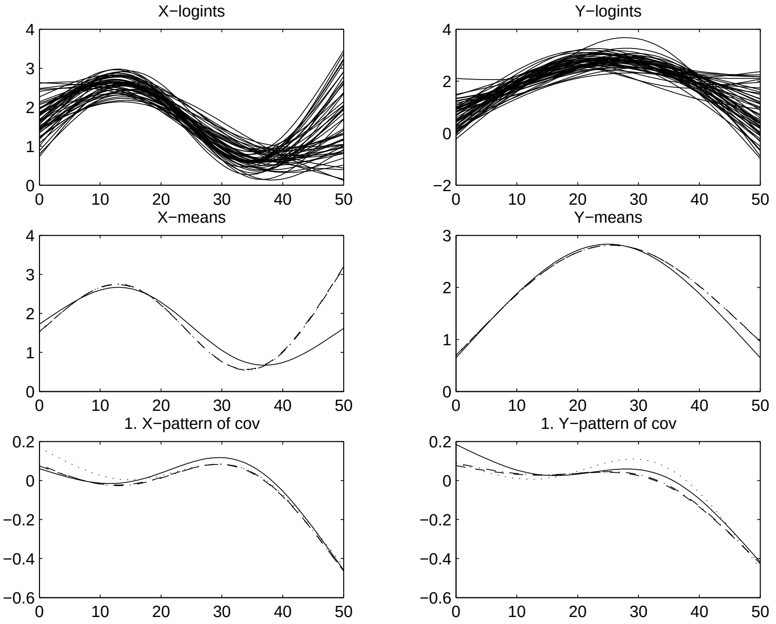

Figure 7: Ex.2. Log-intensity functions (1.row), mean functions (2.row) and first pattern of covariation (3.row), obtained under different sampling schemes: true (solid line), poi (dashed line), poipar1 (dashed-dotted line), poipar2 (dotted line).

Additionally, posterior sampling in WinBUGS again for $K = 1$, $r_Q^X = 4$, $r_Q^Y = 3$ and $r_R^X = r_R^Y = 4$ with initial values from variational inference was carried out. For this example the difference can be expected to be more pronounced, because in order to facilitate variational inference the Poisson-likelihood is approximated by second order Taylor expansions with working observations. Indeed, figure 8 shows that in FCCA more accurate estimates of log-intensity functions are obtained with MCMC (based on 20000 updates after a burn-in of 40000, taking about 11 hours). In contrast, for the multivariate inner model with simulated coefficients θ_m^Z as data hardly any improvement in the extraction of patterns of covariation could be achieved (with 50000 samples after a burn-in of 50000 taking about 15 minutes).

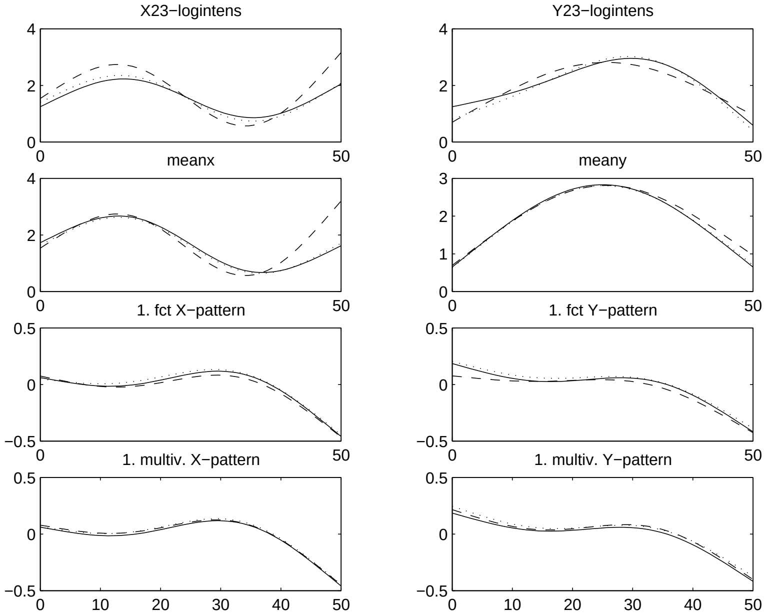

Figure 8: Ex.2. Data set ‘poi’. Log-intensity functions f_{23}^Z (1.row), mean functions (2.row) and first pattern of covariation in FCCA (3.row) and CCA (4.row). True functions are represented by solid lines, estimates by variational inference by dashed lines, MCMC estimates by dotted lines.

4.2 Functional prediction

Next consider the problem of predicting $Y-$ curves from $X-$ curves, that is of predicting f_m^Y from X_m based on FCCA or an inner regression model with $\Sigma_{\theta^Z} = \sigma_{\theta^Z}^2 I_{r_R^Z}$ (INDXY).

The two examples previously used to illustrate FCCA are continued in this section.

Example 1 (continued)

For illustration consider again example 1 with three different data sets. There are four $Y-$ curves common to all data sets to be predicted from observed corresponding $X-$ curves. Thus the data sets 1-1/96 and 1-2/96 are augmented to comprise 100 $X-$ curves and 96 $Y-$ curves each. The comparison of these two data sets is to show that the higher precisions in the second one are indeed reflected in higher accuracy of predictions. The comparison of the augmented data sets 1-1/96 and 1-1/16 (with 20 $X-$ curves and 16 $Y-$ curves) is to illustrate how loss of information due to a smaller number of curves results in increased uncertainty about the predicted curves.

In order to corroborate the choice of K based on the lower bound in INDXY in table 3 precisions of the columns of G^Z are listed (with high precisions pointing to unnecessary latent variables according to ARD). Furthermore, the given values of average absolute errors in reconstructing the true function values $f^Z_{md^Z}$ by the posterior mean $Q_{d^Z}\mu^*_{\delta^Z} + R_{d^Z}E(G^Z|\widetilde{x}, D')E(s_m|\widetilde{x}, D')$ allow for an assessment of how well the lower bound reflects the fit with latent variables. All given values result from variational inference after 10 iterations for the augmented data sets with true smoothing parameters.

For dataset 1-1/96 ARD suggests at least 3 latent variables, maybe 4. The lower bound also points to $K = 4$, and the fit of functions is best for $K = 5$. The prediction error is minimized for $K = 5$ (0.174), but only slightly smaller than for $K = 4$ (0.176). Hence for this data set $K = 4$ is a well justified choice.

For data set 1-2/96 at least $K = 3$ and at most $K = 4$ are chosen according to ARD. Maximization of the lower bound yields $K = 3$. The fit of $X-$curves is best for $K = 4$, that of $Y-$curves is best for $K = 5$. The prediction error points to $K = 5$. Thus the lower bound results in the smallest K but the average absolute prediction error differs by only 0.0043 (with function values ranging about -2.5 and 5).

For data set 1-1/16 a cautious interpretation of ARD suggests $K = 3$, the lower bound indicates the smaller $K = 2$. $X-$ functions are best fitted using $K = 2$, $Y-$ functions using $K = 3$. The prediction error is minimized for $K = 5$, but again only slightly smaller than for $K \in \{2, 3\}$. Here $K = 3$ maybe a compromising choice.

In all three data sets the lower bound suggests fewer latent variables than ARD but the resulting differences in the prediction errors are not pronounced.

Next, predictions based on FCCA and INDXY are compared. The subsequent analyses of the augmented data sets result from the following findings: model search based on the lower bound yielded the true smoothing parameters $r^X_Q = 3, r^X_R = 5, r^Y_Q = 5, r^Y_R = 7$ for all data sets. For data set 1-1/96 $K = 3$ was suggested by canonical correlation coefficients for FCCA, $K = 4$ by the lower bound for INDXY. For data set 1-2/96 $K = 3$ was indicated in both FCCA and (by the lower bound) INDXY. With fewer curves (1-1/16) only two latent variables were identified by FCCA as well as INDXY, again along with the correct smoothing parameters except that $r^Y_Q = 3$ (rather than $r^Y_Q = 5$) was

Table 3: Ex.1. Functional prediction with INDXY for data sets 1-1/96, 1-2/96 and 1-1/16. Posterior means of $\lambda_{\gamma_k^Z}$, values lb(K) of the lower bound (in units of 10^3), average absolute errors (fit $f^Z_{md^Z}$) and average absolute prediction errors (prederr) in reconstructing $f^Z_{md^Z}$ with latent variables.

1-1/96

	$\lambda_{\gamma_1^Z}$	$\lambda_{\gamma_2^Z}$	$\lambda_{\gamma_3^Z}$	$\lambda_{\gamma_4^Z}$	$\lambda_{\gamma_5^Z}$
X	1.33	1.21	2.85	36.80	53.28
Y	0.31	1.57	0.82	2.55	8.50
	K=1	K=2	K=3	K=4	K=5
lb(K)	6.26	6.66	6.75	**6.80**	6.76
fit $f^X_{md^X}$	0.27	0.09	0.07	0.0478	**0.0475**
fit $f^Y_{md^Y}$	0.30	0.24	0.18	0.1268	**0.1259**
prederr	0.38	0.26	0.20	0.176	**0.174**

1-2/96

	$\lambda_{\gamma_1^Z}$	$\lambda_{\gamma_2^Z}$	$\lambda_{\gamma_3^Z}$	$\lambda_{\gamma_4^Z}$	$\lambda_{\gamma_5^Z}$
X	1.52	1.12	2.55	12.81	17.68
Y	0.34	1.83	0.91	3.41	20.44
	K=1	K=2	K=3	K=4	K=5
lb(K)	6.48	7.68	**7.72**	7.67	7.64
fit $f^X_{md^X}$	0.25	0.029	0.023	0.021	**0.020**
fit $f^Y_{md^Y}$	0.21	0.074	0.058	**0.044**	0.045
prederr	0.45	0.084	0.074	0.072	**0.067**

1-1/16

	$\lambda_{\gamma_1^Z}$	$\lambda_{\gamma_2^Z}$	$\lambda_{\gamma_3^Z}$	$\lambda_{\gamma_4^Z}$	$\lambda_{\gamma_5^Z}$
X	1.45	21.08	3.25	135.06	135.17
Y	0.36	3.72	69.83	144.94	122.49
	K=1	K=2	K=3	K=4	K=5
lb(K)	1,062	**1,074**	1,049	1,030	1,016
fit $f^X_{md^X}$	0.19	0.087	**0.081**	0.086	0.088
fit$f^Y_{md^Y}$	0.22	0.21	**0.156**	0.18	0.162
prederr	0.37	0.339	0.340	0.41	**0.32**

suggested in INDXY. The corresponding predictions of one curve are shown in figure 9.

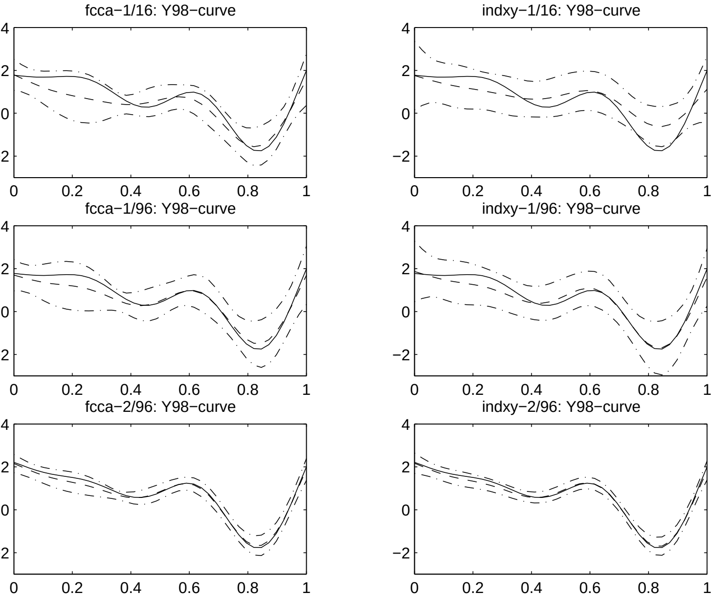

Figure 9: Ex.1. Curve 98 predicted based on different modelling assumptions. FCCA: 1.column, INDXY: 2.column. Three data sets: 1/16: 1.row, 1/96: 2.row, 2/96: 3.row. Each panel displays the true curve (solid line), the posterior predictive mean function (dashed line) and a pointwise prediction error of +/- 3 times the posterior standard deviation (dashed-dotted lines).

Note that with only 16 complete pairs of curves the true curve 98 is not covered by the prediction interval obtained with model INDXY. In order to further compare the performance of the two approaches relative errors $(||\widetilde{f}^Z_{md^Z} - \widehat{\widetilde{f}^Z_{md^Z}}||^2_{\Re^{N^Z}}/||\widetilde{f}^Z_{md^Z}||^2_{\Re^{N^Z}})$ for the vectors of de-noised function values are given for all four pairs of curves ($m \in \{1, 2, 3, 4\}$) in table 4.

Table 3 shows that $\widetilde{f}^Y_1$ is particularly hard to predict and only caught with the augmented data set 1-2/96. Applied to this highly informative data set the two approaches hardly differ. A general conclusion which model should be preferred does not seem to be justified. FCCA is a less restrictive model but the choice of K hinges on the estimation

Table 4: Ex.1. Relative errors in prediction

	1-	1/96	1-	2/96	1-	1/16
	FCCA	INDXY	FCCA	INDXY	FCCA	INDXY
$\widetilde{f}_1^X$	0.10	0.009	0.003	0.001	0.81	0.02
$\widetilde{f}_2^X$	0.05	0.02	0.004	0.005	0.28	0.12
$\widetilde{f}_3^X$	0.07	0.007	0.002	0.002	0.25	0.08
$\widetilde{f}_4^X$	0.18	0.04	0.004	0.005	0.49	0.03
$\widetilde{f}_1^Y$	0.42	0.44	0.027	0.039	1.06	2.26
$\widetilde{f}_2^Y$	0.05	0.02	0.008	0.005	0.18	0.18
$\widetilde{f}_3^Y$	0.04	0.03	0.009	0.007	0.08	0.25
$\widetilde{f}_4^Y$	0.05	0.03	0.008	0.005	0.09	0.19

of canonical correlation coefficients which is not very robust. INDXY allows for the choice of K based on the lower bound and performs as well as FCCA if the data set is sufficiently informative.

Example 2 (continued)

In example 2 Poisson data generated from given log-intensity functions were considered. To illustrate prediction three out of the last four pairs of curves are chosen (curves 97, 98, 99 from in total 100 curves, some displayed in figure 7), such that $M^X = 100$ $X-$ curves and $M^Y = 96$ $Y-$ curves are used to estimate the model parameters. Predictions are obtained according to INDXY and based on partial designs (poipar1) with $N_m^X = 40$ and $N_m^Y = 35$. The model parameters obtained in FCCA with 100 pairs of curves after 3 runs, $K = 1$, $r_Q^X = 4$, $r_R^X = 4$, $r_Q^Y = 3$, $r_R^Y = 4$ were used, and again three runs applied. Resulting curves are displayed in figure 10.

The predicting $X-$ log-intensity curves were reasonably well identified with individual estimates differing from the mean function (figure 10, first row). The errors that did occur are however considerably enlarged by the exponential transformation (figure 10, second row) yielding the intensity-functions. Pointwise the errors in estimating the mean $Q_{d^Y}\delta^Y$ are very small such that the prediction error is determined by the uncertainty about the residual log-intensity functions. The inner regression model did not extract a sufficiently strong relation between θ_m^X and θ_m^Y to discriminate the $Y-$ log-intensities to be predicted (figure 10, 3.row), and the $Y-$ log-intensity functions are essentially estimated by the mean function (figure 10, 4.row). This of course also holds on a larger scale for the predicted $Y-$intensity functions (figure 10, 5.row). The prediction intervals do inform about the remaining uncertainty. In summary, the data set with $N_m^X = 40$ and $N_m^Y = 35$ is not informative enough to allow for precise prediction of individual curves. The chosen sample sizes are not uncommon in biometrical applications and thus the example cautions against too much optimism with respect to prediction even if a major part of the covariance structure is discovered.

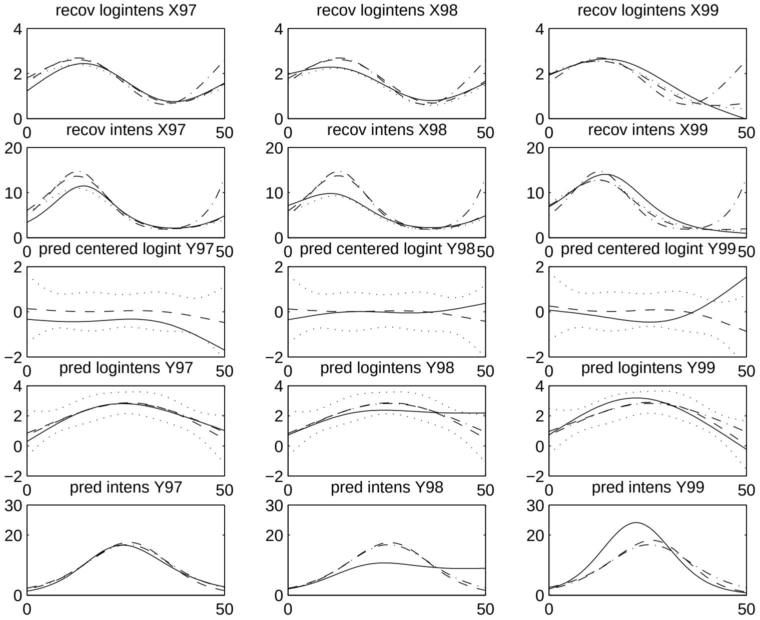

Figure 10: Ex.2: Columns illustrate fits of $f^X_{md^X}$ and predictions of $f^Y_{md^Y}$ from X_m for curve 97 (1.column), curve 98 (2.column) and curve 99 (3.column). 1.row: true log-intensity (solid line) $f^X_{md^X}$, mean X - log-intensity $Q_{d^X}\mu^*_{\delta^X}$ (dashed-dotted line), estimate $\widehat{\widehat{f}}^X_{md^X} = Q_{d^X}\mu^*_{\delta^X} + R_{d^X}\mu^*_{\theta^X_m}$ (dotted line), estimate $\widehat{f}^X_{md^X} = Q_{d^X}\mu^*_{\delta^X} + R_{d^X}E(G^X|\widetilde{x}, D')E(\widetilde{s}_m|\widetilde{x}, D')$ (dashed line). 2.row: The same functions as in the first row, but exponentially transformed. 3.row: True residual curve $R_{d^Y}\theta^Y_m$ (solid line), posterior predictive mean $R_{d^Y}E(G^Y|\widetilde{x}, D')E(\widetilde{s}_m|\widetilde{x}, D')$ (dashed line) and a pointwise prediction error of +/- 3 times the posterior standard deviation (dotted lines). 4.row: true log-intensity $f^Y_{md^Y}$ (solid line), mean Y-log-intensity $Q_{d^Y}\mu^*_{\delta^Y}$ (dashed-dotted line), posterior predictive mean function $\widehat{f}^Y_{md^Y} = Q_{d^Y}\mu^*_{\delta^Y} + R_{d^Y}E(G^Y|\widetilde{x}, D')E(\widetilde{s}_m|\widetilde{x}, D')$ (dashed line) and pointwise prediction error of +/- 3 times the posterior standard deviation (dotted line). 5.row: The first three functions as in the fourth row exponentially transformed.

4.3 Scalar prediction

The special case of a univariate continuous response Y and a functional predictor X is considered. Example 1 is continued applying the models INDY and INDXY.

Example 1 (continued).

Again the augmented data set 1-2 is considered with 100 $X-$ and 96 $Y-$ curves which are strongly related. Two scalar targets are defined as

$$T_{1m} = f_m^Y(t_0), \quad t_0 = 0.677$$

and

$$\begin{aligned} T_{2m} &= 2 + \alpha^T R_{d^X} \theta_m^X, \\ \alpha^T &= (\alpha(t_1), ..., \alpha(t_{N^X})), \quad \alpha(t_n) = 5\cos(2\pi t^2). \end{aligned}$$

T_1 is a function value by which the $Y-$ curves can be distinguished although it is not the value with maximum spread (cp. figure 4). T_2 mimics a functional regression with a functional regression coefficient α. Corresponding Y_1- values are given by $Y_m(t_0)$ (with $\sigma_Y = 0.06$). Y_2- values are generated adding Gaussian noise with $\sigma_Y = 0.01$. Predictions of the last four $T-$ values are obtained using the (true) smoothing parameters for $X-$ curves ($r_Q^X = 3$, $r_R^X = 5$) detected in the previous analyses with both INDXY and FCCA. As in FCCA the lower bound in INDY turned out to be decreasing in the number K of latent variables, and hence cannot be used to determine K. INDY differs from FCCA only by the hyperparameters of the Gamma prior on λ_{θ^Y}, and in scalar prediction with FCCA $K = 1$ is the only choice. Hence it may be tried in INDY as well. In fact, with $K > 2$ the prediction error varies only slightly. In INDXY $K = 2$ was chosen based on the lower bound for both targets. As shown in figure 11 both targets are predicted well by INDY and INDXY.

In table 5 the choice of K in INDXY by maximization of the lower bound and alternatively ARD are compared, and additionally measures of fit for the underlying functions and prediction errors are given. The values were obtained with the true smoothing parameters $r_Q^X = 3$, $r_R^X = 5$ after 10 iterations of the variational algorithm.

For T_1 ARD is not conclusive: using a cutpoint of 3 two latent variables would suffice to represent the target, but $K > 2$ variables are needed to represent the $X-$curves. The choice of $K = 2$ by the lower bound goes along with the best reconstruction of the $X-$curves by the latent variables, but the fit of $Y-$values as well as the prediction can be improved with $K > 2$.

For T_2 it is again not clear which cutpoint to use in ARD. A cautious decision would be $K = 4$. With the lower bound $K = 2$ is (slightly) preferred to higher values, and again corresponds to the best reconstruction of $X-$curves. The reconstruction of $Y-$values can be (slightly) improved with $K > 2$, and also the prediction error is smaller for $K > 2$.

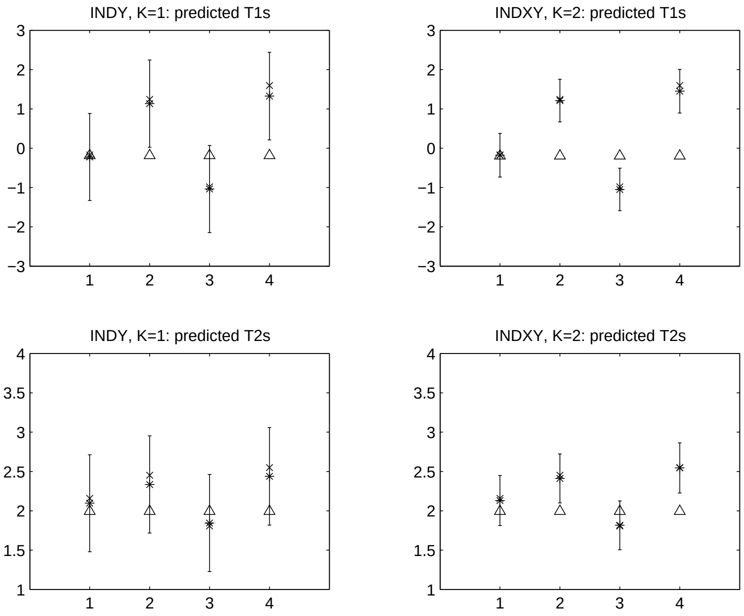

Figure 11: Ex.1: Prediction of two scalar targets. 1.row: function values T_{1m}, 2.row: sums T_{2m}. Crosses mark the values to be predicted, stars the estimates. The vertical solid lines describe the prediction intervals based on 3 standard deviations. The triangles represent the estimated mean value.

Table 5: Ex.1. Scalar prediction with INDXY for data sets 1-1/96, 1-2/96 and 1-1/16. Posterior means of $\lambda_{\gamma_k^Z}$, values lb(K) of the lower bound (in units of 10^3), average absolute errors (fit $f^Z_{md^Z}$) and average absolute prediction errors (prederr) in reconstructing $f^Z_{md^Z}$ with latent variables.

T$_1$

	λ_{γ_1}	λ_{γ_2}	λ_{γ_3}	λ_{γ_4}	λ_{γ_5}
X	3.03	2.52	2.93	1.56	4.31
Y	2.60	9.16	3.73	2.68	3.71
	K=1	K=2	K=3	K=4	K=5
lb(K)	3.23	**3.81**	3.78	3.75	3.71
fit $f^X_{md^X}$	0.25	**0.0262**	0.0266	0.265	0.0274
fit $f^Y_{md^Y}$	0.18	0.0621	0.0611	0.0607	**0.0605**
prederr	0.54	0.062	0.059	**0.055**	0.056

T$_2$

	λ_{γ_1}	λ_{γ_2}	λ_{γ_3}	λ_{γ_4}	λ_{γ_5}
X	10.58	1.94	2.03	2.13	3.30
Y	89.18	155.27	19.18	67.56	34.99
	K=1	K=2	K=3	K=4	K=5
lb(K)	3.32	**3.828**	3.827	3.79	3.75
fit $f^X_{md^X}$	0.25	0.0264	0.0265	**0.0262**	0.0279
fit $f^Y_{md^Y}$	0.11	0.019	0.0178	**0.0175**	0.0177
prederr	0.24	0.020	**0.0170**	0.0171	0.0170

Thus in INDXY (modelling FPCA in regression) the choice of K resulting from the maximization of the lower bound is dominated by the reconstruction of the $X-$ curves. Larger values of K may contribute to a more adequate regression and hence to better predictions. Here the effect is weak though: for $K = 2$ an average absolute deviance of the estimated $\widehat{T}_{1m}$ from the true T_{1m} of 0.062 is obtained in the validation set while for $K = 4$ an error of 0.055 is achieved. Similarly, for $K = 2$ the error in predicting T_{2m} is 0.020 and drops to 0.017 for $K > 2$.

4.4 Functional discriminant analysis

In this section classification of observational units in J groups based on functional covariates is considered. Example 1 is continued posing a problem of binary classification. The solutions resulting from the approximation of the lower bound based on working observations and from the softmax parametrization are compared. In the second real data example from the field of speech recognition classification into more than two groups is required and only the softmax parametrization is used.

Example 1 (continued).

A binary variable Y' was derived, indicating whether the value taken by T_1 was positive or not, that is $Y'_m = 1 \Leftrightarrow T_{1m} > 0$. First, using working observations based on an expansion in random perturbations of the logit of the relative frequencies of ones in the $Y-$test data set, logits $\widehat{\xi}^{Y'}_m$ were predicted, transformed to probabilities $\widehat{\pi}^{Y'}_m$ and predicted values of Y'_m obtained setting $\widehat{Y}'_m = 1$ whenever $\widehat{\pi}^{Y'}_m > 0.5$. In table 6 misclassification rates "*testmc*(wo)", "*valmc*(wo)" are given for INDXY with several splits of the M^X =100 pairs (x_m, y'_m) into a test set of size M and a validation set of size m^X. Smoothing parameters found in previous analyses of the $X-$ curves ($r^X_Q = 3$, $r^X_R = 5$) were used. The model parameter K was searched in a range of 1 to 10 for $m^X \in \{10, 20, 30\}$ and in a range of 1 to 20 for $m^X \in \{40, 50\}$ and chosen corresponding to the minimum misclassification rate in the test set. Based on ARD $K = 7$ should have been chosen for all m^X.

Table 6: Ex.1: Misclassification rates in fitting/predicting $Y'_1 > 0$

m^X	$K(wo)$	$testmc(wo)$	$valmc(wo)$	$testmc(lb)$	$valmc(lb)$	$K(lb)$
10	3	0.00	0.00	0.01	0.00	2
20	4	0.00	0.00	0.04	0.00	2
30	4	0.00	0.00	0.06	0.00	5
40	18	0.05	0.05	0.05	0.00	8
50	11	0.06	0.00	0.04	0.18	8

For all test sets iterating over the points of expansion that generated working observations did not improve the misclassification rates. Comparison to classification based on the softmax parametrization (that is, based on the maximum predicted log probability $\widehat{\eta}^{Y'}_{jm}, j = 1, 2$) yields that the two approaches perform similarly well. The corresponding misclassification rates "*testmc*(*lb*)", "*valmc*(*lb*)" are listed in table 6. The value $K(lb) = 8$ yielding a relatively high misclassification rate of 0.18 with $m^X = 50$ in the validation set is too low: with $K(lb) = 13$ the rates $testmc(lb) = 0.04$ and $valmc(lb) = 0.08$ and with $K(lb) = 17$ the rates $testmc(lb) = 0.06$ and $valmc(lb) = 0.00$ could be achieved. Analyses with unrestricted Σ_{θ^X} (INDY) yielded similar results in terms of misclassification rates but required fewer latent variables in Bouchard's approach ("lb"). For example, for $m^X = 40$, $K(lb) = 4$ was sufficient to obtain $testmc(lb) = 0.03, valmc(lb) = 0.03$, and for $m^X = 50$, $K(lb) = 3$ resulted in $testmc(lb) = 0.04$ and $valmc(lb) = 0.02$.

Example 3 (speech recognition)

As an example of functional classification with more than two groups log-periodograms corresponding to the phonemes "sh", "iy", "dcl", "aa" and "ao" were analyzed. The comprehensive data set is available at
http://www.stat.stanford.edu/ElemStatLearn. It was introduced by Hastie et al. (1995) and also discussed by Ferraty and Vieu (2006, ch.8). For each of the five phonemes 50 log-periodograms were randomly selected for the test set and another 20 for the validation set. The log-periodograms of original length 256 were cut to a length of 150. Thus the log-periodograms used here look very much the same as those displayed in figure 2.4 in the book by Ferraty and Vieu (2006, p.16) with the classes "aa" and "ao" being most similar. As the log-periodograms are rather rough the interpolation design d^X was chosen to be equal to the individual designs d^X_m with $N^X_m = N^X = 150$ points (frequencies). A preliminary fit of the mean $X-$ function within the spline basis yielded the smoothing parameter $r^X_Q(0) = 13$, and similar fits for the mean residual curve within each group suggested $r^X_R(0) = 18$. To reduce the computational burden the search for the smoothing parameters was not further refined and the given values kept in all analyses. In table 7 misclassification rates for a range of K both for FCCA and INDXY are given, separately for the test set ("*testmc*") and the validation set ("*valmc*").

The misclassification rates of INDX are intermediate: best $testmc = 0.20$ with $valmc = 0.42$ for K=7. In FCCA the correlation coefficients point to $K = 2$ canonical variates only. INDXY performs strikingly better than FCCA. $K = 11$ would have been chosen corresponding to the smallest *testmc*. A stabilization of rates can be observed, however, for $K \geq 8$, and the smaller $K = 8$ even yields a slightly better *valmc*. The values of *valmc* for $K \geq 8$ are well in the range of good misclassification rates reported by Ferraty and Vieu (2006, fig.8.2, p.121). ARD for INDXY points to $K = 7$.

5 Discussion

Finally, some issues are to be discussed systematically which have been addressed only occasionally in the previous sections.

Table 7: Ex 3: Misclassification rates for phonemes

K	FCCA		INDXY	
	testmc	valmc	testmc	valmc
1	0.77	0.68	0.68	0.68
2	0.67	0.62	0.38	0.38
3	0.65	0.70	0.21	0.29
4	0.70	0.62	0.16	0.15
5	0.55	0.57	0.13	0.11
6			0.13	0.11
7			0.13	0.09
8			0.11	0.07
9			0.11	0.09
10			0.11	0.08
11			0.10	0.09
12			0.13	0.07
13			0.11	0.07

5.1 Model and inference

The functional RRR model and the variational algorithm are two different, in principle independent aspects of the given analyses. Although just the Demmler-Reinsch like basis of functions and the variational algorithm combine well to yield a convenient tool for approximate Bayesian inference covering many standard applications, a variational algorithm may become inapplicable if the model is to be extended or refined, and sampling methods may have to be used instead. The referees expressed the concern that using mainly variational inference to analyze RRR models in this paper the model and the method of inference may be confounded. However, at a basic level of explorative analysis for the examples with simulated data point estimates, credible intervals or predictions resulting from variational inference could be validated and hence justified. At a more subtle level of posterior dependencies, however, doubts remain. The crucial convergence of the parameters of the inner (CCA) regression model in WinBUGS even with initial values obtained as posterior means in variational inference indicates that more sophisticated and targeted MCMC sampling strategies are needed to tell apart peculiarities of the computational approach to inference and features of the RRR model.

5.2 Limitations and extensions of the model

The model studied in this paper is rather flexible with respect to both the error distributions and the designs under which the functions may be observed. It is limited in as much the inner regression model is a parametric, even linear model. Several model extensions are of interest which have already been tried occasionally but have not been incorporated within the proposed class of models.

Basic functions

Instead of interpolation splines with domains in $\Re$ spatial or higher dimensional interpolation splines could be used in the same set-up. However, smoothing parameters may then be dimension specific, and the number of basis functions as smoothing parameter can be too restrictive. Some of the examples with highly oscillating trigonometric functions considered in the literature (e.g. by Cardot et al., 2007) cannot be handled with the basis of splines used here. Other basis functions like the Fourier basis can be used instead, but these are driven by different smoothing parameters and hence induce different problems of model choice.

Error schemes

In order to robustify inference, t-distributions instead of Gaussian error distributions are popular. They do not belong to one-parameter exponential families but can be incorporated as mixtures of Gaussian distributions with another level of hierarchy (Archambeau et al., 2006). De la Cruz (2008) suggests a Bayesian analysis for classifying longitudinal data under a skew elliptical error scheme.

Allowing for unknown link functions rather than the canonical link functions would extend the flexibility of the model (Amato et al., 2006). Note however, that the choice of the link function and the distributional assumption for the latent variables are closely related and flexibility in one model component can compensate simplicity of the other one.

Distribution of latent variables

The specification of the distribution of s_m as a mixture of Gaussians may allow to extend canonical correlation analysis aiming at uncorrelated pairs of canonical variates to an analysis aiming at independent pairs of variates, in analogy to the transition from principal component analysis to independent component analysis (Choudrey and Roberts, 2001). Indeed, this problem has been addressed by Karhunen (2007), though not based on a probabilistic model. However, as shown in section 3.1, the identification of patterns of covariation ex post hinges on the Gaussian concept of covariance, and it is not clear how to derive interpretable patterns of covariation reflecting general dependence with non-Gaussian latent variables.

In the prediction models considered in this paper the latent vectors s_m were assumed to be multivariate Gaussian. For the different purpose of joint clustering the latent variable may in contrast be assumed to be multinomial. This case was studied by Klami and Kaski (2008) and by Bigelow and Dunson (2006).

Non-functional covariates

Although formally the block bivariate model appears to be symmetric in X and Y only sub-models with functional X were considered. In particular the case of a functional response Y and non-functional covariates X (comprising ANOVA as treated by Reinsel and Velu (2003) or Kaufman and Sain (2010)) and functional mixed models (dealt with by Chiou, Müller and Wang (2004) or Morris et al. (2006)) were omitted. Bayesian methodology for regression models with non-functional manifest regressors is advanced and the class of models discussed in this paper may not be the most appropriate class in that case.

5.3 Model identification

In non-functional PCA and factor analysis as well as RRR models with manifest regressors (orthogonality) constraints that ensure identifiability of the model have been proposed and might be incorporated in a prior distribution (Minka, 2001; Šmídl and Quinn, 2007; Lopes and West, 2004; Geweke, 1996). In functional approaches the orthogonality constraints are not immediately applicable conflicting with the representation in a basis of functions. Moreover in CCA the derivation of patterns of covariation corresponds to a particular identification of the RRR. This is involved because of the pre-standardization of the random vectors as the formula for L^Z (31) demonstrates. Canonical variates, (linear) discriminant functions and predictor effects are not incorporated in the parameterization and the prior, and the resulting necessity to derive patterns of covariation in a post processing step is a principled weakness of the RRR model. Further work is needed to resolve these problems.

5.4 Inference

The relative simplicity of the RRR model, compared to more ambitious non-parametric approaches, allows for the approximate closed form posterior inference and thus for a pragmatic Bayesian approach to FDA which is not computationally intensive and may be a good starting point also for the development of more complex models and more elaborate Bayesian inference. Efficient MCMC algorithms or hybrid methods still have to be developed, and model choice based on posterior sampling needs further investigation. Furthermore, the accuracy of the approximate Bayesian inference is to be evaluated in comparison to MCMC inference.

Performance of variational inference

The performance of the proposed approach cannot be evaluated analyzing only three examples. Therefore, many more examples particularly with real data sets and different sampling schemes were analyzed which are described in the technical report (van der Linde, 2010). The following comments are based on this more comprehensive experience. Also, as to its performance the cautionary remarks by van der Linde (2009) concerning the interactions of required sample sizes and degrees of smoothness of the underlying curves and about the appropriateness of the chosen basis functions in functional principal component analysis apply here as well, in particular to functional canonical correlation analysis.

With the examples for FCCA it was demonstrated that the proposed computational approach works well: interpretable functional patterns of covariation can be extracted and inference is sensitive to loss of information in the data. The approach is coherent with multivariate CCA for discretized curves (if applicable), but various error schemes and partial designs can also be handled.

Prediction in general requires that not only common features of curves like the principal modes of (co-)variation can be detected, but that individual curves can be discriminated by the latent variables, and - moreover - that the dependence structure modelled in regression is strong enough to transfer identification of of $X-$ curves by latent variables to identification of $Y-$ values by latent variables. With the first (running) example a synthetic data set with strong dependence in regression was devised, and prediction was satisfactory for all $Y-$ scales (functional, scalar, multinomial). This example thus demonstrates that the approach does work if there is a pronounced structure of dependence. In other examples (example 2 and gait data (not shown)), however, (poor) prediction by mean values was observed. The introduction of the error Σ_{θ^Y} not only in FCCA but also in predictive models like INDXY provides a diagnostic tool to assess the strength of a tendency towards the mean: a comparison of $E(\theta_m^Z|\widetilde{x}, D')$ to $E(G^Z|\widetilde{x}, D')E(s_m|\widetilde{x}, D')$ for observed Z_m often reveals a good fit of $E(\theta_m^Z|\widetilde{x}, D')$, that is, good filtering or de-noising, but a poor explanation of $E(\theta_m^Z|\widetilde{x}, D')$ by $E(G^Z|\widetilde{x}, D')E(s_m|\widetilde{x}, D')$.

In functional prediction poor predictive performance was attributed to a weak relation between $X-$ and $Y-$ curves in the data set. In scalar prediction in contrast not a single weak but too many strong relations between X and Y may pose a problem. With spectrometric data (not shown)) the inherent prediction by principal components was not competitive in comparison to more sophisticated methods. Although the analyses always did yield reasonable indications of structure there is some inertia in the reproduction of individual values (or curves) by latent variables. This was already observed for FPCA by van der Linde (2009) and showed up again in regression where prediction based on multivariate manifest principal components outperformed prediction based on "latent principal components". There was no evidence that this performance was related to unlucky or misleading initial values resulting possibly in inappropriate or insufficient model identification. However, the curves over-determined the scalar responses. The analyses of spectrometric data thus point to the limits of the class of models discussed here (that is, to the limits of PCA). In functional and scalar prediction no marked difference between FCCA or INDY and INDXY could be observed.

In functional classification the variational algorithms based on an approximate lower bound induced by working observations and based on a valid lower bound induced by pseudo observations using the softmax parametrization were compared only for example 1 where both approaches performed similarly well. The results using working observations, also those obtained for Poisson data, are encouraging and working observations should be tried particularly if there is no alternative. Focusing on Bouchard's approach INDXY and INDY outperformed FCCA. This may be explained by the fact that in the softmax parametrization there is a dependence between the parameters which is caught in s_m by the decomposition $cov(\theta_m^Y|G^Y, s_m) = \sigma^2_{\theta^Y} I_{r_R^Y} + G^Y s_m s_m^T (G^Y)^T$ in IND(X)Y.

In contrast, in FCCA Σ_{θ^Y} instead of $\sigma^2_{\theta^Y} I_{r^Y_R}$ occurs in the decomposition, and with s_m only the dependency within θ^Y_m may not be fully recovered. (In functional prediction dependence between function values is caught in the basis functions, not in the coefficients θ^Z_m.) The misclassification rates obtained with the variational algorithm applied to the RRR model were excellent for simulated data and comparable to those reported for similar methods in speech recognition. Hence prediction on the rougher nominal scale worked better than on the continuous scale where higher accuracy is required.

Model choice based on variational inference

Model choice remains a crucial issue. A major advantage of the lower bound as a criterion for model choice is that is computationally cheap being calculated as a byproduct for monitoring programming and convergence of variational inference anyway. Together with the simplicity of the smoothing parameters related to the Demmler-Reinsch like basis of functions it is a main ingredient of a fast and pragmatic approach. However, it turned out to be of no use for choosing the number of latent variables K in FCCA and INDY (seemingly due to the Wishart distribution). In scalar prediction the asymmetry between X and Y affects the lower bound, and in classification the inclusion of variational parameters does not allow for comparisons. Only in functional prediction with INDXY it turned out to work. Hence, in most cases the lower bound does not point to an adequate value of K. The alternative of ARD is computationally not more expensive, but can be ambiguous. Canonical correlation coefficients again can easily be computed, but turned out to be instable with an increasing number of iterations. More experiences need to be reported in this respect.

Implementation in WinBUGS

The RRR model can easily be implemented in WinBUGS and posterior sampling initialized with estimates obtained by variational inference. In the multivariate CCA model with simulated coefficients θ^Z_m the extraction of patterns of covariation could not be essentially improved in this way, but the error assessment might be more accurate. Convergence of the values of the latent variables s_m remained doubtful even after a long period (350000 updates) of burn-in. In FCCA the situation is worse because the inner regression model is specified at the third level of hierarchy (and coefficients θ^Z_m are estimated). Convergence proceeds from the outer model (function values close to the data) to the inner model. Computing times in WinBUGS for the full functional model were in the range of 5.5 to 7.5 hours for 20000 updates used for inference after 10000 samples of burn-in. This is prohibitive for model choice based on a criterion like DIC. Posterior sampling is computer intensive, but in FCCA MCMC did improve variational inference. Sampling should be initialized by preliminary estimates (which can be obtained using variational inference with two minutes of computing time), in order to avoid even longer chains to achieve convergence.

5.5 Dependence between multivariate non-Gaussian variables

In the hierarchical model investigated in this paper the dependence between observed vectors is specified by a latent regression model for conditionally Gaussian parameter vectors, that is, by a covariance structure. No attempt has been made to extend CCA directly to non-Gaussian multivariate distributions like the multivariate Bernoulli or multivariate Poisson distribution. A particularly interesting extension with applications in ecology would be the analysis of dependence between sets of (a priori) Dirichlet distributed profiles observed under "multinomial noise". "Principal profiles" fitting into the concepts discussed here were already studied by Hastie and Little (1987).

Appendices

Appendix 1: Update rules for functional CCA

with $M^X = M + m^X \geq M^Y = M$.

- $$s_m \sim N(\mu^*_{s_m}, \Sigma^*_{s_m}).$$

  For $m = 1, ..., M$

  $$\Sigma^*_{s_m} = \left(\left\langle (G^X)^T \Lambda_{\theta^X} G^X + (G^Y)^T \Lambda_{\theta^Y} G^Y \right\rangle + I_K\right)^{-1},$$

  $$\mu^*_{s_m} = \Sigma_{s_m} \left(\left\langle G^X \right\rangle^T \langle \Lambda_{\theta^X} \rangle \langle \theta^X_m \rangle + \left\langle G^Y \right\rangle^T \langle \Lambda_{\theta^Y} \rangle \langle \theta^Y_m \rangle \right).$$

  For $m = M + 1, ..., M^X$

  $$\Sigma^*_{s_m} = \left(\left\langle (G^X)^T \Lambda_{\theta^X} G^X + \right\rangle + I_K\right)^{-1},$$

  $$\mu^*_{s_m} = \Sigma_{s_m} \left(\left\langle G^X \right\rangle^T \langle \Lambda_{\theta^X} \rangle \langle \theta^X_m \rangle \right).$$

- $$\delta^Z \sim N(\mu^*_{\delta^Z}, \Sigma^*_{\delta^Z}).$$

  $$\Sigma^*_{\delta^Z} = \left(\sum_{m=1}^{M^Z} (Q^Z_m)^T \langle \Lambda_{Z_m} \rangle Q^Z_m + \lambda^0_{\delta^Z} I_{r^Z_Q} \right)^{-1},$$

  $$\mu^*_{\delta^Z} = \Sigma^*_{\delta^Z} \left(\sum_{m=1}^{M^Z} (Q^Z_m)^T \langle \Lambda_{Z_m} \rangle (Z_m - R^Z_m \langle \theta^Z_m \rangle) \right)$$

  where $\Lambda_{Z_m} = \lambda_Z I_{N^Z_m}$.

- For $m = 1, ..., M^Z$

$$\theta_m^Z \sim N(\mu^*_{\theta_m^Z}, \Sigma^*_{\theta_m^Z}),$$

$$\Sigma^*_{\theta_m^Z} = \left((R_m^Z)^T \left\langle \Lambda_{Z_m} \right\rangle R_m^Z + \left\langle \Lambda_{\theta^Z} \right\rangle\right)^{-1},$$

$$\mu^*_{\theta_m^Z} = \Sigma^*_{\theta_m^Z} \left((R_m^Z)^T \left\langle \Lambda_{Z_m} \right\rangle (Z_m - Q_m^Z \left\langle \delta^Z \right\rangle) + \left\langle \Lambda_{\theta^Z} \right\rangle \left\langle G^Z \right\rangle \left\langle s_m \right\rangle\right).$$

-

$$\lambda_Z \sim \Gamma(\alpha^*_{\lambda_Z}, \beta^*_{\lambda_Z}).$$

$$\alpha^*_{\lambda_Z} = \alpha^0_{\lambda_Z} + \frac{1}{2}\sum_{m=1}^{M^Z} N_m^Z,$$

$$\beta^*_{\lambda_Z} = \beta^0_{\lambda_Z} + \frac{1}{2}\sum_{m=1}^{M^Z} \left\langle ||\widetilde{z}_m||^2 \right\rangle$$

where $\widetilde{z}_m = z_m - Q_m^Z \delta^Z - R_m^Z \theta_m^Z$.

-

$$\Lambda_{\theta^Z} \sim W(\nu_Z^*, (C_Z^*)^{-1})$$

$$\nu_Z^* = \nu_Z^0 + M^Z$$

$$C_Z^* = C_Z^0 + \left\langle \sum_{m=1}^{M^Z} (\theta_m^Z - G^Z s_m)(\theta_m^Z - G^Z s_m)^T \right\rangle$$

-

$$\gamma_k^Z \sim N(\mu^*_{\gamma_k^Z}, \Sigma^*_{\gamma_k^Z})$$

$$\Sigma^*_{\gamma_k^Z} = \left(\left\langle ||\widetilde{s}_k||^2 \Lambda_{\theta^Z} + \lambda_{\gamma_k^Z} I_{r_R^Z} \right\rangle\right)^{-1}$$

where $\widetilde{s}_k$ denotes the k-th row of S,

$$\mu^*_{\gamma_k^Z} = \Sigma^*_{\gamma_k^Z} \left\langle \Lambda_{\theta^Z} \right\rangle \left\langle \sum_{m=1}^{M^Z} \theta_m^Z s_{km} - G_{-k}^Z S_{-k} \widetilde{s}_k^T \right\rangle$$

where G_{-k}^Z equals G^Z without the k-th column and S_{-k} equals S without the k-th row. See also (van der Linde, 2008, section 3.3).

- $$\lambda_{\gamma_k^Z} \sim \Gamma(\alpha^*_{\lambda_{\gamma_k^Z}}, \beta^*_{\lambda_{\gamma_k^Z}})$$
$$\alpha^*_{\lambda_{\gamma_k^Z}} = \alpha^0_{\lambda_{\gamma_k^Z}} + \frac{1}{2} r_R^Z$$
$$\beta^*_{\lambda_{\gamma_k^Z}} = \beta^0_{\lambda_{\gamma_k^Z}} + \frac{1}{2} \left\langle ||\gamma_k^Z||^2 \right\rangle .$$

The number of iterations needed varied between 3 and 20.

Appendix 2: Modification of update of Λ_{θ^Z} under independence in functional prediction

If $\Lambda_{\theta^Z} = \lambda_{\theta^Z} I_{r_R^Z}$ and apriori (17) holds, then aposteriori independently

$$\lambda_{\theta^Z} \sim \Gamma(\alpha^*_{\lambda_{\theta^Z}}, \beta^*_{\lambda_{\theta^Z}}),$$

$$\alpha^*_{\lambda_{\theta^Z}} = \alpha^0_{\lambda_{\theta^Z}} + \frac{1}{2} M^Z r_R^Z,$$

$$\beta^*_{\lambda_{\theta^Z}} = \beta^0_{\lambda_{\theta^Z}} + \frac{1}{2} tr \left\langle \sum_{m=1}^{M^Z} (\theta_m^Z - G^Z s_m)(\theta_m^Z - G^Z s_m)^T \right\rangle .$$

Appendix 3: Default values

Denote again by I_n the identity matrix of dimension n and by $J_{n\times m}$ a matrix of ones with n rows and m columns. Let further $randU(n,m)$ be a $n \times m-$ matrix with entries which are simulated from a uniform distribution on (0,1) and similarly $randN(n,m)$ a matrix the entries of which are simulated from a standard Normal distribution.

A3.1 Hyperparameters

In (8): $\lambda^0_{\delta^Z} = 10^{-3}$.

In (9): $\alpha^0_{\lambda_Z} = \beta^0_{\lambda_Z} = 10^{-3}$.

In (15): $\alpha^0_{\lambda_{\gamma_k^Z}} = \beta^0_{\lambda_{\gamma_k^Z}} = 10^{-3}$, for $k = 1, ..., K$.

In (16): $\nu_Z^0 = r_R^Z + 2$, $C_Z^0 = 10^{-3} I_{r_R^Z}$,

alternatively in (17): $\alpha^0_{\theta^Z} = \beta^0_{\theta^Z} = 10^{-3}$.

A3.2 Initial values

$\mu^{+}_{s_m} = randN(K,1),\ \Sigma^{+}_{s_m} = I_K$ for $m = 1, ..., M$.

$\mu^{+}_{\delta^Z} = J_{r^Z_Q \times 1},\ \Sigma^{+}_{\delta^Z} = I_{r^Z_Q}$.

$\mu^{+}_{\theta^Z_m} = randN(r^Z_R, 1),\ \Sigma^{+}_{\theta^Z_m} = I_{r^Z_R}$ for $m = 1, ..., M$.

$\alpha^{+}_{\lambda_Z} = \beta^{+}_{\lambda_Z} = 1$.

$\nu^{+}_{Z} = r^Z_R + 2,\ C^{+}_{Z} = I_{r^Z_R}$, alternatively $\alpha^{+}_{\theta^Z} = \beta^{+}_{\theta^Z} = 1$.

$\mu^{+}_{\gamma^Z_k} = randU(r^Z_R, 1),\ \Sigma^{+}_{\gamma^Z_k} = I_{r^Z_R}$, for $k = 1, ..., K$.

$\alpha^{+}_{\lambda_{\gamma^Z_k}} = \beta^{+}_{\lambda_{\gamma^Z_k}} = 1$, for $k = 1, ..., K$.

Appendix 4: Update of variational parameters in Bouchard's algorithm

Denote the initial (approximate factorized) posterior distribution by $q^{(0)}$ and an initial variational parameter a by $a^{(0,1)} = (a^{(0,1)}_1, ..., a^{(0,1)}_{M^Y})$.

For $l = 1, ...L$:

From $q^{(h)}, a^{(h,l)}$ obtain the variational parameters $\zeta^{(h,l)}_m = (\zeta^{(h,l)}_{1m}, ..., \zeta^{(h,l)}_{Jm})$, $m = 1, ..., M^Y$ by

$$\zeta^{(h,l)}_{jm} = [E_{q^{(h)}}(\eta^2_{jm}) + (a^{(h,l)}_m)^2 - 2a^{(h,l)}_m E_{q^{(h)}}(\eta_{jm})]^{1/2}.$$

For $l = 2, ...L$:

From $q^{(h)}, \zeta^{(h,l-1)}$ obtain the variational parameters $a^{(h,l)}_m$, $m = 1, ..., M^Y$ by

$$a^{(h,l)}_m = \frac{\frac{1}{2}(\frac{J}{2} - 1) + \sum_{j=1}^{J} g(\zeta^{(h,l-1)}_{jm}) E_{q^{(h)}}(\eta_{jm})}{\sum_{j=1}^{J} g(\zeta^{(h,l-1)}_{jm})}$$

using g from (27).

Given $a^{(h,L)}_m$ and $\zeta^{(h,L)}_m$ calculate pseudo observations $v^{(h)}_m$ according to equations (24) to (28) for $m = 1, ..., M^Y$. $q^{(h)}$ and $v^{(h)} = (v^{(h)}_1, ..., v^{(h)}_{M^Y})$ then yield a new posterior distribution $q^{(h+1)}$, and setting $a^{(h+1,1)} := a^{(h,L)}$ the process can be iterated.

References

Amato, U., Antoniadis, A., and Feis, I. D. (2006). "Dimension Reduction in Functional Regression with Applications." *Computational Statistics and Data Analalysis*, 50: 2422–2446.

Archambeau, C., Delannay, N., and Verlysen, M. (2006). "Robust Probabilistic Projections." In *Proceedings of the 23rd international conference on Machine Learning, ACM International Conference Proceeding Series*, volume 148, 33 – 40.

Attias, H. (1999). "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes." In Laskey, K. and Prade, H. (eds.), *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 21–30.

Bach, F. and Jordan, M. (2005). "A probabilistic interpretation of canonical correlation analysis." *Technical Report 688*. Dept. of Statistics, University of California, Berkeley.

Baladandayuthapani, V., Mallick, B., Hong, M., Lupton, J., Turner, N., and Carroll, R. (2007). "Bayesian Hierarchical Spatially Correlated Functional Data Analysis with Application to Colon Carcinogesis." *Biometrics*, 64: 64–73.

Bigelow, J. and Dunson, D. (2006). "Bayesian semiparametric clustering of functional predictors." URL: http://ftp.isds.duke.edu/WorkingPapers/06-13.pdf.

Bouchard, G. (2007). "Efficient bounds for the Softmax Function and Applications to Approximate Inference in Hybrid models." In *Neural Information Processing Systems Conference, Whistler, Canada, December 7-8*.

Bougeard, S., Hanafi, M., and Qannari, E. (2008). "Continuum redundancy-PLS regression: A simple continuum approach." *Computational Statistics and Data Analysis*, 52: 3686–3696.

Brooks, R. and Stone, M. (1994). "Joint continuum regression with multiple predictands." *Journal of the American Statistical Association*, 89: 1374–1377.

Cardot, H., Crambes, C., Kneip, A., and Sarda, P. (2007). "Smoothing splines estimators in functional linear regression with errors-in-variables." *Computational Statistics and Data Analysis*, 51: 4832–4848.

Chiou, J.-M., Mueller, H.-G., and Wang, J.-L. (2004). "Functional response models." *Statistica Sinica*, 14: 659–677.

Choudrey, W., Penny, W., and Roberts, S. (2000). "An ensemble learning approach to independent component analysis." In *Proceedings IEEE Workshop on Neural Networks for Signal Processing, Sydney, Australia, December 2000*, 435–444. IEEE Press.

Choudrey, W. and Roberts, S. (2001). "Flexible Bayesian Independent Component Analysis for Blind Source Separation." In *Proceedings of ICA-2001, San Diego, December 2001*.

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

De la Cruz, R. (2008). "Bayesian non-linear regression models with skew-elliptical errors: Applications to the classification of longitudinal profiles." *Computational Statistics and Data Analysis*, 53: 436–449.

Dunson, D., Herring, A., and Siega-Riz, A. (2008). "Bayesian inference on changes in response densities over predictor clusters." *Journal of the American Statistical Association*, 103: 1508–1517.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. New York: Springer.

Frank, I. and Friedman, H. (1993). "A Statistical View of Some Chemometrics Regression Tools." *Technometrics*, 35: 109–148. (with discussion).

Geweke, J. (1996). "Bayesian reduced rank regression in econometrics." *Journal of Econometrics*, 75: 121–146.

Hastie, T., Buja, A., and Tibshirani, R. (1995). "Penalized discriminant analysis." *Annals of Statistics*, 13: 435–475.

Hastie, T. and Little, F. (1987). "Principal Profiles." URL: http://www-stat.stanford.edu/˜hastie/Papers/PrincipalProfiles.pdf.

James (2002). "Generalized linear models with functional predictors." *Journal of the Royal Statistal Society B*, 64: 411–432.

James, G. and Hastie, T. (2001). "Functional linear discriminant analysis for irregularly sampled curves." *Journal of the Royal Statistal Society B*, 63: 533–550.

Karhunen, J. (2007). "Extending ICA for finding jointly dependent components for two related data sets." *Neurocomputing*, 70: 2969–2979.

Kaufman, C. and Sain, S. (2010). "Bayesian Functional ANOVA Modeling Using Gaussian Process Prior Distributions." *Bayesian Analysis*, 5: 123–150.

Klami, A. and Kaski, S. (2008). "Probabilistic approach to detecting dependencies between data sets." *Neurocomputing*, 72: 39–46.

Lopes, H. and West, M. (2004). "Bayesian Model Assessment in Factor Analysis." *Statistica Sinica*, 14: 41–67.

MacLehose, R. and Dunson, D. (2009). "Nonparametric Bayes kernel-based priors for functional data analysis." *Statistica Sinica*, 19: 611–629.

Manteiga, W. and Vieu, P. (2007). "Statistics for Functional Data." *Computational Statistics and Data Analysis*, 51: 4788–4792.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. London: Academic Press.

Merola, M. and Abraham, B. (2001). "Dimensionality reduction approach to multivariate prediction." *Canadian Journal of Statistics*, 29: 191–200.

Minka, T. (2001). "Automatic choice of dimensionality for PCA." In Leen, T., Dieterich, T., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems*, volume 13, 598–604. MIT Press.

Morris, J., Brown, P., Baggerly, K., and Coombes, K. (2006). "Analysis of Mass Spectrometry Data Using Bayesian Wavelet-Based Functional Mixed Models." In Do, K., Mueller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 269–288. New York: Cambridge University Press.

Petrone, S., Guindani, M., and Gelfand, A. (2009). "Hybrid Dirichlet mixture models for functional data." *Journal of the Royal Statistal Society B*, 71: 755–782.

Preda, C. (2007). "Regression models for functional data by reproducing kernel Hilbert space methods." *Journal of Statistical Planning and Inference*, 137: 829–840.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer.

— (2005). *Functional Data Analysis*. New York: Springer, 2nd edition.

Ray, S. and Mallick, B. (2006). "Functional clustering by Bayesian wavelet methods." *Journal of the Royal Statistal Society B*, 68: 305–332.

Reinsel, G. and Velu, R. (1998). *Multivariate reduced-rank regression*. New York: Springer.

— (2003). "Reduced-rank growth curve models." *Journal of Statistical Planning and Inference*, 114: 107–129.

Rish, I., Grabarnik, G., Cecchi, G., Pereira, F., and Gordon, G. J. (2008). "Closed-form supervised dimensionality reduction with generalized linear models." In *ACM Int. Conf. Proc. Series*, volume 307, 832–839. New York: ACM.

Rodriguez, A., Dunson, D., and Gelfand, A. (2009). "Bayesian non-parametric functional data analysis through density estimation." *Biometrika*, 96: 149–162.

Schmidli, H. (1995). *Reduced Rank Regression*. Heidelberg: Physica-Verlag.

Smidl, V. and Quinn, A. (2007). "On Bayesian principal component analysis." *Computational Statistics and Data Analysis*, 51: 4101–4123.

Srivastava, M. (2007). "Reduced Rank Discrimination." *Scandinavian Journal of Statistics*, 24: 115–124.

Sundberg, R. (2002). "Continuum Regression." In *Encyclopedia of Statistical Sciences*. New York: Wiley.

Thompson, W. and Rosen, O. (2008). "A Bayesian model for sparse functional data." *Biometrics*, 64: 54–63.

Tipping, M. and Bishop, C. (1999). "Probabilistic principal component analysis." *Journal of the Royal Statistal Society B*, 21: 611–622.

Valderrama, M. (2007). "An overview to modelling functional data." *Computational Statistics*, 22: 331–334.

van der Linde, A. (2008). "Variational Bayesian functional PCA." *Computational Statistics and Data Analysis*, 53: 517–533.

— (2009). "Bayesian functional principal components analysis for binary and count data." *Advances in Statistical Analysis*, 93: 307–333.

— (2010). "Reduced rank regression in Bayesian FDA." *Technical report. Department of Mathematics, Institute of Statistics, University of Bremen, Germany*.

von Storch, H. and Zwiers, F. (1999). *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press.

Wang, C. (2007). "Variational Bayesian approach to canonical correlation analysis." *IEEE Transactions on Neural Networks*, 18: 905–910.

West, M. (2003). "Bayesian Factor Regression Models in the Large p, Small n Paradigm." In et al., J. B. (ed.), *Bayesian Statistics 7*, 733–742. Oxford University Press.

Yee, T. and Hastie, T. (2003). "Reduced-rank Vector Generalized Linear Models." *Statistical Modelling*, 3: 15–41.

Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., and Wu, M. (2006). "Supervised probabilistic principal component analysis." In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA*, 464–473. New York: ACM.

Zhang, D., Lin, X., and Sowers, M. (2007). "Two-Stage Functional Mixed Models for Evaluating the Effect of Longitudinal Covariate Profiles on Scalar Outcome." *Biometrics*, 63: 351–362.

Acknowledgments

The comments and suggestions of referees and editors helped to improve earlier drafts of this paper and are gratefully acknowledged.