Journal of the
Royal Statistical Society

J. R. Statist. Soc. B (2008)
70, *Part* 4, *pp.* 703–723

Modelling sparse generalized longitudinal observations with latent Gaussian processes

Peter Hall,

University of California, Davis, USA, and University of Melbourne, Australia

Hans-Georg Müller

University of California, Davis, USA

and Fang Yao

University of Toronto, Canada

[Received April 2006. Final revision December 2007]

Summary. In longitudinal data analysis one frequently encounters non-Gaussian data that are repeatedly collected for a sample of individuals over time. The repeated observations could be binomial, Poisson or of another discrete type or could be continuous. The timings of the repeated measurements are often sparse and irregular. We introduce a latent Gaussian process model for such data, establishing a connection to functional data analysis. The functional methods proposed are non-parametric and computationally straightforward as they do not involve a likelihood. We develop functional principal components analysis for this situation and demonstrate the prediction of individual trajectories from sparse observations. This method can handle missing data and leads to predictions of the functional principal component scores which serve as random effects in this model. These scores can then be used for further statistical analysis, such as inference, regression, discriminant analysis or clustering. We illustrate these non-parametric methods with longitudinal data on primary biliary cirrhosis and show in simulations that they are competitive in comparisons with generalized estimating equations and generalized linear mixed models.

Keywords: Binomial data; Eigenfunction; Functional data analysis; Functional principal component; Prediction; Random effect; Repeated measurements; Smoothing; Stochastic process

1. Introduction

1.1. Preliminaries

When undertaking predictive inference in longitudinal data analysis involving irregularly spaced and infrequent measurements, relatively little information is often available about each subject, owing to sparse and irregular measurements. Irregularity of measurements for individual subjects is an inherent difficulty of such studies. Therefore it is especially important to use all the information that can be accessed. This requires us to model the relationships between measurements that are made at widely separated time points. We aim at a flexible non-parametric functional data analysis approach, which is in contrast with commonly used parametric models such as generalized linear mixed models (GLMMs) or generalized estimation equations

---

Address for correspondence: Hans-Georg Müller, Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA.
E-mail: mueller@wald.ucdavis.edu

© 2008 Royal Statistical Society 1369–7412/08/70703

(GEEs) (see, for example, Heagerty (1999) for recent discussions on applying such models to repeated binary measurements, Pourahmadi (2000) for related aspects of covariance modelling and Heagerty and Zeger (2000), Heagerty and Kurland (2001) and Chiou and Müller (2005) for discussions on limitations, modifications and feasibility of the underlying parametric assumptions).

A non-parametric functional approach for the analysis of longitudinal data, with its philosophy to let the data speak for themselves and its inherent flexibility, is expected to perform better than the parametric GEE or GLMM approaches in many situations. However, it faces difficulties due to the potentially large gaps between repeated measurements in typically sparse longitudinal data. The parametric methods overcome this easily by postulating a parametric form of the underlying functions. In contrast, in the presence of such gaps, the classical non-parametric approach to smooth individual trajectories in a first step is not feasible (Yao *et al.*, 2005). The problems that are caused by gaps are exacerbated in the commonly encountered case of non-Gaussian longitudinal responses such as binomial or Poisson responses (see Section 5).

We demonstrate how one can overcome the difficulties that are posed by such data for non-parametric approaches, by applying suitably modified methods of functional data analysis. Functional data analysis methods have been primarily developed for smooth and densely sampled data (Ramsay and Silverman, 2002, 2005). The basic idea to connect the data that we wish to analyse to functional data analysis methodology is to postulate an underlying latent Gaussian process (LGP) (for other examples of latent processes modelling for longitudinal studies compare, for example, Diggle *et al.* (1998), Jowaheer and Sutradhar (2002), Hashemi *et al.* (2003) and Proust *et al.* (2006)). Specifically, the Gaussian property makes it possible to overcome sparseness by a conditioning argument. Relevant features of the stochastic relationships of the observed data are reflected by the mean and covariance properties of this LGP. Simulations indicate that the method is in practice quite insensitive to the Gaussian assumption for the latent process.

Since sufficiently flexible parameterizations of the underlying Gaussian process would suffer from a large number of parameters, making corresponding maximum likelihood approaches computationally demanding and unstable, we propose instead to connect the LGP to random trajectories for individual observations directly by means of a link function. These subject-specific trajectories correspond to the probabilities of a response in the binary response case. Whereas the link function is assumed known, the mean and covariance of the Gaussian process are assumed to be unknown but smooth. This proposition is attractive on grounds of flexibility, but it raises the challenging problem of constructing appropriate estimators.

The methodology proposed is a first attempt to extend functional data analysis technology to the case of non-Gaussian repeated measurements. Prominent examples for such data are repeated binary measurements or repeated counts. The methods proposed are motivated by several considerations: the variation of random coefficients may be relatively low, and in this case a simple Taylor approximation motivates simple, explicit and non-parametric mean and covariance function estimators; and these estimators are elementary to compute, irrespective of whether the low variation assumption is satisfied or not. The simple, low variation estimators that we propose are attractive owing to their flexibility and numerical simplicity.

The analysis of continuous Gaussian sparse longitudinal data by functional methods has been considered previously (e.g. Shi *et al.* (1996), Rice and Wu (2000), James *et al.* (2001) and James and Sugar (2003)). Our main tool from functional data analysis is functional principal component (FPC) analysis, where observed trajectories are decomposed into a mean function and eigenfunctions (e.g. Rice and Silverman (1991) and Boente and Fraiman (2000)). Various aspects of the relationship between functional and longitudinal data are discussed in Staniswalis and Lee (1998), Rice (2004) and Zhao *et al.* (2004); an early study of modelling longitudinal trajectories in biological applications with FPCs is Kirkpatrick and Heckman (1989). FPC analysis allows us to achieve three major goals:

(a) dimension reduction of functional data by summarizing the data in a few FPCs;
(b) the prediction of individual trajectories from sparse data, by estimating the FPC scores of the trajectories;
(c) further statistical analysis of longitudinal data based on the FPC scores.

In the next subsection, we introduce the LGP model; then in Section 2 the proposed estimators, followed by applications to prediction (Section 3). The results from a simulation study, including a comparison of the method proposed with GLMMs and GEEs, are reported in Section 4. The analysis of non-Gaussian sparse longitudinal data is illustrated in Section 5, with the longitudinal analysis of the occurrence of hepatomegaly in primary biliary cirrhosis. This is followed by a brief discussion (Section 6) and an appendix, which contains derivations and some theoretical results about estimation.

1.2. *Latent Gaussian process model*

Generally, denoting the generalized responses by Y_{ij}, we observe independent copies of Y, but, in each case, only for a few sparse time points. In particular, the data are pairs (T_{ij}, Y_{ij}), for $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant m_i$, where $Y_{ij} = Y_i(T_{ij})$ for an underlying random trajectory Y_i, and each $T_{ij} \in \mathcal{I} = [0, 1]$. The sparse and scattered nature of the observation times T_{ij} may be expressed theoretically by noting that the m_is are uniformly bounded, if these quantities have a degenerate distribution, or that they represent the values of independent and identically distributed random variables with sufficiently light tails, if the m_is originate stochastically. We are aiming at the seemingly difficult task of making such sparse designs amenable to functional methods, which have been primarily aimed at densely collected smooth data.

A central assumption for our approach is that the dependence between the observations Y_{ij} is inherited from an underlying unobserved Gaussian process X: let $Y(t)$, for $t \in \mathcal{T}$, where $\mathcal{T}$ is a compact interval, denote a stochastic process satisfying

$$
\begin{aligned}
E\{Y(t_1)\ldots Y(t_m)|X\} &= \prod_{j=1}^{m} g\{X(t_j)\}, \\
E\{Y(t)^2|X\} &\leqslant g_1\{X(t)\}
\end{aligned}
\tag{1}
$$

for $0 \leqslant t_1 < \ldots < t_m \leqslant 1$ and $0 < t < 1$. Here, X denotes a Gaussian process on $\mathcal{I}$, g is a smooth, monotone increasing link function, from the real line to the range of the distribution of the Y_{ij}, and g_1 is a bounded function. Although we observe independent copies of Y, these are accessible only for a few sparse time points for each subject. The Gaussian processes X_i and measurement times T_{ij}, for $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant m_i$, are assumed to be totally independent, the T_{ij}s are taken to be identically distributed as T, say, with support $\mathcal{I}$ and the X_is are supposed to be identically distributed as X. When interpreted for the data (T_{ij}, Y_{ij}), model (1) implies that

$$
E\{Y_i(T_{i1})\ldots Y_i(T_{im_i})|X_i(T_{i1}), \ldots, X_i(T_{im_i})\} = \prod_{j=1}^{m_i} g\{X_i(T_{ij})\}. \tag{2}
$$

The assumption that X at model (1) is Gaussian provides a plausible way of linking stochastic properties of $Y(t)$ for values t in different parts of $\mathcal{I}$, so that data that are observed at each time point can be used for inference about future values of $Y(t)$ for any specific value of t. The idea of pooling data across subjects to overcome the sparseness problem is motivated as in Yao *et al.* (2005). The link function g is assumed known; for example we might select the logit link in the binary data case, $g(x)=\exp(x)/\{1+\exp(x)\}$, and the log-link for count data; under some circumstances, the link can also be estimated non-parametrically. An important special case of model (1) is that of binary responses, i.e. 0–1 data, where the first identity in model (1) simplifies to

$$P\{Y(t_1)=l_1,\ldots,Y(t_m)=l_m\,|\,X\}=\prod_{j=1}^{m} g\{X(t_j)\}^{l_j}[1-g\{X(t_j)\}]^{1-l_j}, \tag{3}$$

for all sequences $l_1,\ldots,l_m$ of 0s and 1s. In this case, the link function g would be chosen as a distribution function and the methodology proposed corresponds to an extension of functional data analysis to longitudinal binary data.

2. Estimating mean and covariance of latent Gaussian processes

To use model (1) to make predictive inference about future values of $Y(t)$, we need to estimate the defining characteristics of the process X, i.e. its mean and covariance structure. In a setting where the distribution of Y can be completely specified, e.g. in the binary data model (3), one possible approach would be maximum likelihood. This is, however, a difficult proposition in the irregular case, where it would necessitate the specification of a large number of parameters for the various means and covariances that are involved, a difficulty which can only be overcome by invoking restrictive assumptions, limiting the flexibility of the approach. Moreover, we are considering a non-stationary case, and the number of parameters would need to increase with n, the sample size. Finally, another major motivation is to extend the functional approach to non-Gaussian longitudinal data. To sustain the non-parametric flavour, we prefer not to make stronger assumptions than model (1), and in particular we do not wish to make the restrictive assumptions that would be necessary to employ maximum likelihood methods.

Our approach is based on the supposition that the variation of X_i about its mean is relatively small. In particular, we assume that

$$X_i(t)=\mu(t)+\delta\,Z_i(t), \qquad \mu=E(X_i), \tag{4}$$

Z_i is a Gaussian process with zero mean and bounded covariance and $\delta>0$ is an unknown small constant. In this case, assuming that g has four bounded derivatives, and writing (X,Z) for a generic pair (X_i,Z_i), we have

$$g(X)=g(\mu)+\delta Z\,g^{(1)}(\mu)+\tfrac{1}{2}\delta^2Z^2\,g^{(2)}(\mu)+\tfrac{1}{6}\delta^3Z^3\,g^{(3)}(\mu)+O_p(\delta^4), \tag{5}$$

$$E[g\{X(t)\}]=g(\mu)+\tfrac{1}{2}\delta^2\,E\{Z^2(t)\}\,g^{(2)}\{\mu(t)\}+O(\delta^4) \tag{6}$$

and

$$\operatorname{cov}[g\{X(s)\},g\{X(t)\}]=\delta^2\,g^{(1)}\{\mu(s)\}\,g^{(1)}\{\mu(t)\}\operatorname{cov}\{Z(s),Z(t)\}+O(\delta^4). \tag{7}$$

Here and throughout we make the assumption that $g^{(1)}$ does not vanish, and that $\inf_{s\in D}\{g^{(1)}(s)\}>0$, where D is the (compact) range of the mean function μ. Setting

$$\left.\begin{aligned}\alpha(t)&=E[g\{X(t)\}],\\ \nu(t)&=g^{-1}\{\alpha(t)\},\\ \tau(s,t)&=\operatorname{cov}[g\{X(s)\},g\{X(t)\}]/g^{(1)}\{\mu(s)\}\,g^{(1)}\{\mu(t)\},\end{aligned}\right\} \tag{8}$$

we obtain

$$\mu(t) = E\{X(t)\} = g^{-1}(E[g\{X(t)\}]) + O(\delta^2) = \nu(t) + O(\delta^2), \tag{9}$$

$$\sigma(s,t) = \operatorname{cov}\{X(s), X(t)\} = \frac{\operatorname{cov}[g\{X(s)\}, g\{X(t)\}]}{g^{(1)}\{\mu(s)\}\, g^{(1)}\{\mu(t)\}} + O(\delta^4) = \tau(s,t) + O(\delta^4). \tag{10}$$

These formulae immediately suggest estimators of μ and σ, if we are willing to neglect the effect of order $O(\delta^2)$. Indeed, we may estimate

$$\alpha(t) = E\{Y(t)\} = E[E\{Y(t)|X(t)\}] = E[g\{X(t)\}], \tag{11}$$

by passing a smoother through the data (T_{ij}, Y_{ij}), and estimate

$$\beta(s,t) = E\{Y(s)Y(t)\} = E[g\{X(s)\}g\{X(t)\}] \tag{12}$$

(by using model (1)) by passing a bivariate smoother through the data $((T_{ij}, T_{ik}), Y_{ij}Y_{ik})$ for $1 \leqslant i \leqslant n$ such that $m_i \geqslant 2$, and $1 \leqslant j, k \leqslant m_i$ with $j \neq k$. It is necessary to omit the diagonal terms in this smoothing step, since according to model (1) we have

$$E\{Y^2(t)\} = E[E\{Y^2(t)|X(t)\}] > E[E\{Y(t)|X(t)\}]^2 = E[g\{X(t)\}]^2,$$

whenever $\operatorname{var}\{Y(t)|X(t)\} > 0$, so the variance along the diagonal in general will have an extra component, leading to a covariance surface that has a discontinuity along the diagonal. More details about this phenomenon can be found in Yao *et al.* (2005). Implementation of these smoothing steps, by using local least squares estimators, is discussed in Appendix A.

From the resulting estimators $\hat{\alpha}$ and $\hat{\beta}$ of α and β respectively, we obtain estimators

$$\begin{aligned} \hat{\nu}(t) &= g^{-1}\{\hat{\alpha}(t)\}, \\ \hat{\tau}(s,t) &= \{\hat{\beta}(s,t) - \hat{\alpha}(s)\,\hat{\alpha}(t)\}/g^{(1)}\{\hat{\nu}(s)\}\, g^{(1)}\{\hat{\nu}(t)\} \end{aligned} \tag{13}$$

for

$$\begin{aligned} \nu(t) &= g^{-1}\{\alpha(t)\}, \\ \tau(s,t) &= \{\beta(s,t) - \alpha(s)\,\alpha(t)\}/g^{(1)}\{\nu(s)\}\, g^{(1)}\{\nu(t)\} \end{aligned} \tag{14}$$

respectively. By virtue of approximations (9) and (10) we may interpret $\hat{\nu}$ and $\hat{\tau}$ as estimators of μ and σ respectively, i.e. we set

$$\begin{aligned} \hat{\mu}(t) &= \hat{\nu}(t), \\ \hat{\sigma}(s,t) &= \hat{\tau}(s,t). \end{aligned} \tag{15}$$

These estimators do not depend on the constant δ, which therefore does not need to be known or estimated. Although the estimator $\hat{\tau}(s,t)$ is symmetric, it will generally not enjoy the positive semidefiniteness property that is required of a covariance function. This deficiency can be overcome by implementing a method that was described in Yao *et al.* (2003), which is to drop from the spectral decomposition of $\hat{\tau}$ those terms that correspond to negative eigenvalues. It is easy to show that, in doing so, the mean-squared error of $\hat{\tau}$ is strictly improved by omitting a term that corresponds to a negative eigenvalue; details can be found in Appendix B. In what follows, we work with the resulting estimators $\tilde{\tau}$ as defined in Appendix B. Properties of the estimators $\hat{\alpha}$ and $\hat{\beta}$, and $\hat{\nu}$ and $\hat{\tau}$, which are defined at expressions (32), (33) and (13) respectively, and of estimators $\hat{\mu}$ and $\hat{\sigma}$ at expression (15) are discussed in Appendix C.

3. Predicting individual trajectories and random effects

3.1. Predicting functional principal component scores

One of the main purposes of the functional data analysis model proposed is dimension reduction through predicted FPC scores. These lead to predicted trajectories of the underlying hidden Gaussian processes for the subjects in a study. Specifically, the predicted FPC scores provide a means for regularizing the irregular data, and also for dimension reduction, and can be used for inference, discriminant analysis or regression.

The starting point is the Karhunen–Loève expansion of random trajectories X_i of the LGP,

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_{ij}\, \psi_j(t), \tag{16}$$

where ψ_j are the orthonormal eigenfunctions of the linear integral operator B with kernel $\sigma(s,t)$, that maps an L^2-function f to $B\,f(s) = \int \sigma(s,t)\, f(t)\, \mathrm{d}t$, i.e. the solutions of

$$\int \operatorname{cov}\{X(s), X(t)\}\, \psi_j(t)\, \mathrm{d}s = \theta_j\, \psi_j(t),$$

where θ_j is the eigenvalue that is associated with eigenfunction ψ_j. The $\xi_{ij} = \int \{X_i(t) - \mu(t)\}\, \psi_j(t)\, \mathrm{d}t$ are the FPC scores that play the role of random effects, with $E(\xi_{ij}) = 0$ and $\operatorname{var}(\xi_{ij}) = \theta_j$, where θ_j is the eigenvalue corresponding to eigenfunction ψ_j. Once the estimator $\sigma(s,t)$ (15) has been determined, the corresponding estimates θ_j and ψ_j of eigenvalues and eigenfunctions of latent processes X are obtained by a standard discretization procedure, whereby these estimates are derived from a discrete principal component analysis step.

We aim to estimate the best linear predictor

$$E\{X_i(t)|Y_{i1}, \ldots, Y_{im}\} = \sum_{j=1}^{\infty} E(\xi_{ij}|Y_{i1}, \ldots, Y_{im})\, \psi_j(t) \tag{17}$$

of the trajectory X_i, given the data $Y_{i1}, \ldots, Y_{im_i}$. Here a truncation of the expansion to include only the first M components is needed. Then, focusing on the first M conditional FPC scores will allow us to reduce the dimension of the problem and also to regularize the highly irregular data. According to equation (17), the task of representing and predicting individual trajectories can be reduced to that of estimating $E(\xi_{ij}|Y_{i1}, \ldots, Y_{im})$. In what follows we develop a suitable approximation in the non-Gaussian case by means of a moment-based approach, as follows. The repeated measurements per subject are assumed to be generated by

$$Y_{ik} = Y_i(T_{ik}) = g\{X_i(T_{ik})\} + e_{ik}, \tag{18}$$

with independent errors e_{ik}, satisfying

$$\begin{aligned} E(e_{ik}) &= 0, \\ \operatorname{var}(e_{ik}) &= \gamma^2\, v[g\{X_i(T_{ik})\}]. \end{aligned} \tag{19}$$

Here, γ^2 is an unknown variance (or dispersion) parameter and $v(\cdot)$ is a known smooth variance function, which is determined by the characteristics of the data. For example, in the case of repeated binary observations, one would choose $v(u) = u(1-u)$. In what follows, we implicitly condition on the measurement times T_{ij}.

With a Taylor series expansion of g, using expression (4) and assuming as before that $\inf\{g^{(1)}(\cdot)\} > 0$, we obtain

$$g\{X(t)\} = g\{\mu(t)\} + g^{(1)}\{\mu(t)\}\{X(t) - \mu(t)\} + O(\delta^2). \tag{20}$$

Defining

$$\varepsilon_{ik}=\frac{e_{ik}}{g^{(1)}\{\mu(T_{ik})\}},$$

$$U_{ik}=\mu(T_{ik})+\frac{Y_{ik}-g\{\mu(T_{ik})\}}{g^{(1)}\{\mu(T_{ik})\}},$$

expressions (19) and (20) lead to $U_{ik}=X_i(T_{ik})+\varepsilon_{ik}+O(\delta^2)$. We next substitute estimates (15) and errors ε_{ik} by

$$\tilde{e}_{ik}=Z_{ik}\gamma\frac{v[g\{\mu(T_{ik})\}]^{1/2}}{g^{(1)}\{\mu(T_{ik})\}},$$

where the Z_{ik} are independent copies of a standard Gaussian $N(0,1)$ random variable, so that the first two moments of $\tilde{e}_{ik}$ are approximating those of ε_{ik}. Then, for small δ, $U_{ik}\approx X_i(T_{ik})+\tilde{e}_{ik}$, implying that

$$E(\xi_{ij}|Y_{i1},\ldots,Y_{im_i})=E(\xi_{ij}|U_{i1},\ldots,U_{im_i})\approx E\{\xi_{ij}|X_i(T_{i1})+\tilde{e}_{i1},\ldots,X_i(T_{im_i})+\tilde{e}_{im_i}\}.$$

Owing to the Gaussian assumption for latent processes X_i, the last conditional expectation is seen to be a linear function of the terms on the right-hand side, and therefore

$$E(\xi_{ij}|Y_{i1},\ldots,Y_{im_i})=A_{ij}\tilde{X}_i \tag{21}$$

is a reasonable predictor for the random effect ξ_{ij}, where $\tilde{X}_i=(X_i(T_{i1})+\tilde{e}_{i1},\ldots,X_i(T_{im_i})+\tilde{e}_{im_i})^{\mathrm{T}}$ and the A_{ij} are matrices depending only on γ, μ, v, g and $g^{(1)}$. These quantities are either known or estimates are available, with the sole exception of γ, the estimation of which is discussed below. The explicit form of equation (21) is given in Appendix D.

3.2. Predicting trajectories

Motivated by equations (16) and (21), predicted trajectories for the LGPs are obtained as

$$X_i(t)=E\{X_i(t)|Y_{i1},\ldots,Y_{im_i}\}=\mu(t)+\sum_{j=1}^{M}A_{ij}\tilde{X}_i\,\psi_j(t), \tag{22}$$

and predicted trajectories for the observed process Y as

$$Y_i(t)=E\{Y_i(t)|Y_{i1},\ldots,Y_{im_i}\}=g\{X_i(t)\}, \tag{23}$$

where t may be any time point within the range of processes Y, including times for which no response was observed. Predicted values for $Y(t)$ can sometimes be used to predict the entire response distribution when the mean determines the entire distribution, such as in binomial and Poisson cases. This method could also be employed for the prediction of missing values in a situation where missing data occur totally at random.

To evaluate the effect of auxiliary quantities on the prediction, we use a cross-validation criterion where we compare predictions of Y_{ik}, which are obtained by leaving that observation out, with Y_{ik} itself. Computing

$$Y_{ik}^{(-ik)}=E(Y_{ik}|Y_{i1},\ldots,Y_{i,k-1},Y_{i,k+1},\ldots,Y_{im_i})=g\{X_i^{(-ik)}(T_{ik})\},\qquad 1\leqslant i\leqslant n,\quad 1\leqslant k\leqslant m_i, \tag{24}$$

where

$$X_i^{(-ik)}(T_{ik}) = \mu(t) + \sum_{j=1}^{M} E(\xi_{ij}|Y_{i1},\dots,Y_{i,k-1},Y_{i,k+1},\dots,Y_{im_i})\,\psi_j(t), \tag{25}$$

we define the Pearson-type weighted prediction error

$$\mathrm{PE}(\gamma^2) = \sum_{i,k} \frac{(Y_{ik}^{(-ik)} - Y_{ik})^2}{v[g\{X_i^{(-ik)}(T_{ik})\}]}, \tag{26}$$

which will depend on the variance parameter γ^2 and implicitly also on the number of eigenfunctions M that are included in the model; see equation (19).

We found that the following iterative selection procedure, for choosing the number of eigenfunctions M and the overdispersion parameter γ^2 simultaneously, led to good practical results: choose a starting value for M; then obtain γ^2 by minimizing the cross-validated prediction error PE with respect to γ^2,

$$\gamma = \arg\min_{\gamma}\{\mathrm{PE}(\gamma^2)\}. \tag{27}$$

Then, in a subsequent step, update M by the criterion that is described below, and repeat these two steps until the values of M and γ^2 stabilize. This iterative algorithm worked very well in practice; typical starting values for M would be 2 or 3.

Specifically, for the choice of M, we adopt a quasi-likelihood-based functional information criterion FIC that is an extension of the Akaike information criterion AIC for functional data (see Yao *et al.* (2005) for a related pseudo-Gaussian likelihood-based criterion). The number of eigenfunctions M, to be included in the model, is chosen in such a way as to minimize

$$\mathrm{FIC}(M) = -2\sum_{i,k} \int_{Y_{ik}}^{Y_{ik}} \frac{Y_{ij} - t}{\gamma^2\, v(t)}\,\mathrm{d}t + 2M. \tag{28}$$

The penalty $2M$ corresponds to that used in AIC; other penalties such as those corresponding to the Bayes information criterion BIC could be used as well.

Some simple algorithmic restrictions can be imposed in this iteration for the choice of M and γ, so that loops cannot happen, although we never observed this to occur. We also investigated direct minimization of equation (26), simultaneously for both γ and M. Besides being considerably more computing intensive, this alternative minimization scheme tended to choose more components and resulted in less parsimonious fits without obtaining better predictions. Instead of making a parametric assumption about the variance function v, in some cases it may be preferable to estimate it nonparametrically. This can be done via semiparametric quasi-likelihood regression (Chiou and Müller, 2005).

4. Simulation results

4.1. Comparisons with generalized estimating equations and generalized linear mixed models

The simulations were based on latent processes $X(t)$ with mean function $E\{X(t)\} = \mu(t) = 2\sin(\pi t/5)/\sqrt{5}$, and $\mathrm{cov}\{X(s), X(t)\} = \lambda_1\,\phi_1(s)\,\phi_1(t)$ derived from a single eigenfunction $\phi_1(t) = -\cos(\pi t/10)/\sqrt{5}$, $0 \leqslant t \leqslant 10$, with eigenvalues $\lambda_1 = 2$ $(\lambda_k = 0,\ k \geqslant 2)$. Then 200 Gaussian and 200 non-Gaussian samples of latent processes consisting of $n = 100$ random trajectories each were generated by $X_i(t) = \mu(t) + \xi_{i1}\,\phi_1(t)$, where for the 200 Gaussian samples the FPC scores ξ_{i1} were simulated from $\mathcal{N}(0, 2)$, whereas the ξ_{i1} for the non-Gaussian samples were simulated from a mixture of two normal distributions: $\mathcal{N}(\sqrt{2}, 2)$ with probability $\frac{1}{2}$ and $\mathcal{N}(-\sqrt{2}, 2)$

with probability $\frac{1}{2}$. Binary outcomes Y_{ij} were generated as Bernoulli variables with probability $E\{Y_{ij}|X_i(t_{ij})\} = g\{X_i(t_{ij})\}$, using the canonical logit link function $g^{-1}(p) = \log\{p/(1-p)\}$ for $0 < p < 1$.

To generate the sparse observations, each trajectory was sampled at a random number of points, chosen uniformly from $\{8,\ldots,12\}$, and the locations of the measurements were uniformly distributed over the domain $[0,10]$. For the smoothing steps, univariate and bivariate product Epanechnikov weight functions were used, i.e. $K_1(x) = (3/4)(1-x^2)\,\mathbf{1}_{[-1,1]}(x)$ and $K_2(x,y) = (9/16)(1-x^2)(1-y^2)\,\mathbf{1}_{[-1,1]}(x)\,\mathbf{1}_{[-1,1]}(y)$, where $\mathbf{1}_A(x)$ equals 1 if $x \in A$ and 0 otherwise for any set A. The number of eigenfunctions M and the overdispersion parameter γ^2 were separately selected for each run by the iteration (27) and equation (28). These iterations converged fast, requiring only 2–4 iteration steps in most cases.

We compare the non-parametric LGP method proposed with the popular parametric approaches provided by GLMMs and GEEs. For the GEE method, we used the unstructured correlation option and both GEEs and GLMMs were run with linear (methods GEE-L and GLMM-L) and in addition with quadratic (methods GEE-Q and GLMM-Q) fixed effects. We use four criteria for the comparisons, measuring discrepancies between estimates and targets both in terms of latent processes X and response processes $Y = g(X)$, and comparing both estimates for mean functions $\mu = E(X)$ and $g(\mu)$ respectively and predictions of subject-specific trajectories X_i and $g(X_i)$ respectively. The latter are available for the LGP and GLMM methods, but not for GEEs, which aim at marginal modelling. The specific criteria for the comparisons are as follows:

$$\mathrm{XMSE} = \int_{\mathcal{I}} \{\mu(t) - \mu(t)\}^2 \,\mathrm{d}t \Big/ \int_{\mathcal{I}} \mu^2(t)\,\mathrm{d}t,$$

$$\mathrm{YMSE} = \int_{\mathcal{I}} [g\{\mu(t)\} - g\{\mu(t)\}]^2 \,\mathrm{d}t \Big/ \int_{\mathcal{I}} g^2\{\mu(t)\}\,\mathrm{d}t, \tag{29}$$

$$\mathrm{XPE}_i = \int_{\mathcal{I}} \{X_i(t) - X_i(t)\}^2 \,\mathrm{d}t \Big/ \int_{\mathcal{I}} X_i^2(t)\,\mathrm{d}t,$$

$$\mathrm{YPE}_i = \int_{\mathcal{I}} [g\{X_i(t)\} - g\{X_i(t)\}]^2 \,\mathrm{d}t \Big/ \int_{\mathcal{I}} g^2\{X_i(t)\}\,\mathrm{d}t, \tag{30}$$

for $i = 1,\ldots,n$. Summary statistics for the values of these criteria from 200 Monte Carlo runs are shown in Table 1.

These results indicate that, first of all, the LGP method proposed is not sensitive to the Gaussian assumption for latent processes. Although there is some deterioration in the non-Gaussian case, it is minimal. This non-sensitivity to the Gaussian assumption has been described before in functional data analysis in the context of principal analysis by conditional expectation (see Yao *et al.* (2005)). Secondly, the non-linearity in the target functions throws the parametric methods off track, even when the more flexible quadratic fixed effects versions are used. We find that the LGP method conveys clear advantages in estimation and especially in predicting individual trajectories in such situations. Whereas the parametric methods are sensitive to violations of assumptions, the LGP method is designed to work under minimal assumptions and therefore provides a useful alternative approach.

4.2. Effect of the size of variation

Here we examine the influence of the size of the variation constant δ on model estimation, including mean function, eigenfunctions and individual trajectories. In addition to criteria (29)

Table 1. Simulation results for the comparisons of mean estimates and individual trajectory predictions obtained by the proposed non-parametric LGP method with those obtained for the established parametric methods GLMM-L, GLMM-Q, GEE-L and GEE-Q, with linear and quadratic fixed effects (see Section 4.1)

Distribution	*Method*	*XMSE*	XPE_i			*YMSE*	YPE_i		
			25th	*50th*	*75th*		*25th*	*50th*	*75th*
Gaussian	LGP	0.1242	0.1529	0.2847	0.7636	0.0076	0.0101	0.0205	0.0433
	GLMM-L	0.4182	0.3405	0.5843	1.283	0.0265	0.0278	0.0369	0.0577
	GLMM-Q	0.4323	0.3479	0.5990	1.319	0.0271	0.0285	0.0377	0.0584
	GEE-L	0.4168				0.0264			
	GEE-Q	0.4308				0.0272			
Non-Gaussian (mixture)	LGP	0.1272	0.1664	0.3166	0.9556	0.0078	0.0109	0.0228	0.0459
	GLMM-L	0.4209	0.3309	0.5943	1.364	0.0266	0.0280	0.0372	0.0589
	GLMM-Q	0.4373	0.3385	0.6118	1.404	0.0274	0.0287	0.0380	0.0597
	GEE-L	0.4227				0.0268			
	GEE-Q	0.4396				0.0277			

Simulations were based on 200 Monte Carlo runs with $n=100$ trajectories per sample, generated for both Gaussian and non-Gaussian latent processes. Simulation results are reported through summary statistics for error criteria XMSE and YMSE (29) for relative squared errors of the mean function estimates of latent processes X and of response processes Y, and the 25th, 50th and 75th percentiles of relative prediction errors XPE$_i$ and YPE$_i$ (30) for individual trajectories of latent and response processes.

and (30), we also evaluated the estimation error for the single eigenfunction in the model (noting that $\int_{\mathcal{I}} \phi_1^2(t)\, \mathrm{d}t = 1$),

$$\mathrm{EMSE} = \int_{\mathcal{I}} \{\phi_1(t) - \phi_1(t)\}^2\, \mathrm{d}t. \tag{31}$$

Using the same simulation design as in Section 4.1 and generating latent processes $X(t;\delta) = \mu(t) + \delta\xi_1\phi_1(t)$ for varying δ, we simulated 200 Gaussian and 200 non-Gaussian samples (as described before) for each of $\delta = 0.5, 0.8, 1, 2$. The Monte Carlo results over 200 runs for the various values of δ are presented in Table 2.

Table 2. Simulation results for the effect of the variation parameter δ

Distribution	δ	*XMSE*	*EMSE*	XPE_i			*YMSE*	YPE_i		
				25th	*50th*	*75th*		*25th*	*50th*	*75th*
Normal	0.5	0.1106	0.7662	0.1188	0.1815	0.3366	0.0068	0.0077	0.0119	0.0205
	0.8	0.1205	0.3801	0.1430	0.2437	0.5710	0.0076	0.0094	0.0171	0.0338
	1	0.1280	0.2434	0.1513	0.2809	0.7857	0.0077	0.0101	0.0203	0.0431
	2	0.1616	0.0429	0.2025	0.3851	0.8137	0.0102	0.0144	0.0362	0.0752
Mixture	0.5	0.1134	0.7198	0.1243	0.1913	0.3651	0.0071	0.0081	0.0126	0.0217
	0.8	0.1258	0.3910	0.1498	0.2563	0.6691	0.0078	0.0100	0.0188	0.0366
	1	0.1323	0.2256	0.1624	0.2986	0.7944	0.0081	0.0113	0.0227	0.0450
	2	0.1633	0.0397	0.2041	0.3840	0.8140	0.0103	0.0158	0.0387	0.0768

Design and outputs of the simulation are the same as in Table 1. EMSE denotes the average integrated mean-squared error for estimating the first eigenfunction.

We find substantial sensitivity of the error EMSE in estimating the eigenfunction on the value of δ. This is caused by the fact that, as δ grows smaller, increasingly more of the variation in the observed data is due to error, rather than to the patterns of the underlying LGP, and therefore it becomes increasingly difficult to estimate the eigenfunction. This is also observed in ordinary FPC analysis, where the error in estimating an eigenfunction is tied to the size of its associated eigenvalue: the larger, the better the eigenfunction can be estimated. Although larger values of δ increase the errors in predicting individual trajectories, this is within expectations: for the predictor processes X, this is because the variation of individual trajectories increases, whereas the binary nature of the responses imposes constraints on how much of this variation is reflected in the sparse observations; for the response processes, the error increases much more, which is because the biases in the approximations that are used for these predictions are increasing with δ.

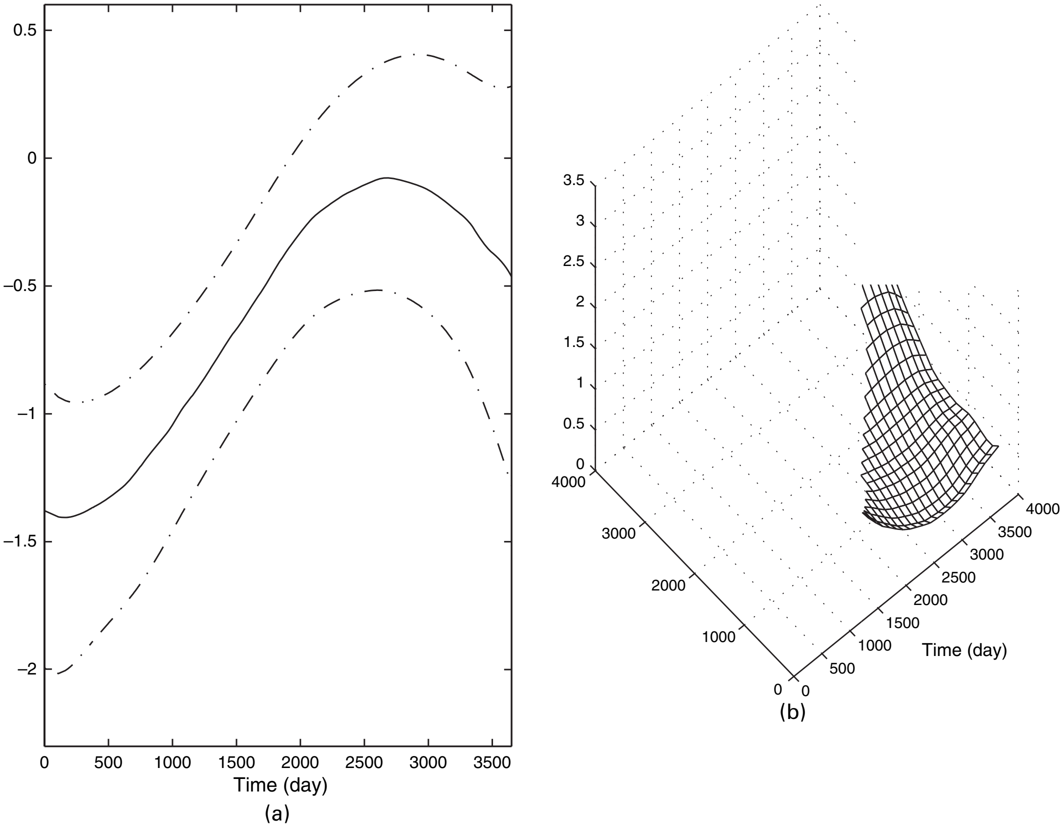
The errors in estimating the mean functions remain fairly stable as long as $\delta \leqslant 1$. This is especially and not unexpectedly observed for the mean of predictor processes X, since this mean estimate is not affected by any approximation error. We conclude that, unless δ is large, its actual value has a small effect on the errors in mean function estimates and a modest effect on the errors in individual predictions, and we note that the strong effect on the error in eigenfunction estimation does not spill over into the predictions for individual trajectories or the mean function estimates, as the effect is mitigated by the multiplication with δ.

5. Application

Primary biliary cirrhosis (Murtaugh *et al.*, 1994) is a rare but fatal chronic liver disease of unknown cause, with a prevalence of about 50 cases per million population. The data were collected between January 1974 and May 1984 by the Mayo Clinic (see also Appendix D of Fleming and Harrington (1991)). The patients were scheduled to have measurements of blood characteristics at 6 months, 1 year and annually thereafter post diagnosis. However, since many individuals missed some of their scheduled visits, the data are sparse and irregular with unequal numbers of repeated measurements per subject and also varying measurement times T_{ij} across individuals.

To demonstrate the usefulness of the methods proposed, we restrict the analysis to the participants who survived at least 10 years (3650 days) since they entered the study and were alive and had not had a transplant at the end of the 10th year. We carry out our analysis on the domain from 0 to 10 years, exploring the dynamic behaviour of the presence of hepatomegaly (0, no; 1, yes), which is a longitudinally measured Bernoulli variable with sparse and irregular measurements. Presence or absence of hepatomegaly is recorded on the days when the patients are seen. We include 42 patients for whom a total of 429 binary responses were observed, where the number of recorded observations ranged from 3 to 12, with a median of 11 measurements per subject.

We employ a logistic link function, and the smooth estimates of the mean and covariance functions for the underlying process $X(t)$ are displayed in Fig. 1. The mean function of the underlying process shows an increasing trend until about 3000 days, except for a short delay at the beginning, and a subsequent decrease towards the end of the range of the data. We also provide pointwise bootstrap confidence intervals which broaden (not unexpectedly) near the end points of the domain. The estimated covariance surface of $X(t)$ displays rapidly decreasing correlation as the difference between measurement times increases. With variance function $v(\mu)=\mu(1-\mu)$, the iterative procedure for selecting the number of eigenfunctions and the variance parameter γ that is described in Section 3.2 yielded the choices $M=3$ for the number of components included and $\gamma^2=1.91$ for the overdispersion parameter. The leave one point out



[illegible]

which were obtained by equation (23) for nine randomly selected subjects, are shown in Fig. 4. The predicted trajectories $\hat{Y}_i(t)$ describe the time evolution of the probability of the presence of hepatomegaly for each individual; it is often increasing, but there are also subjects with mild or strong declines.

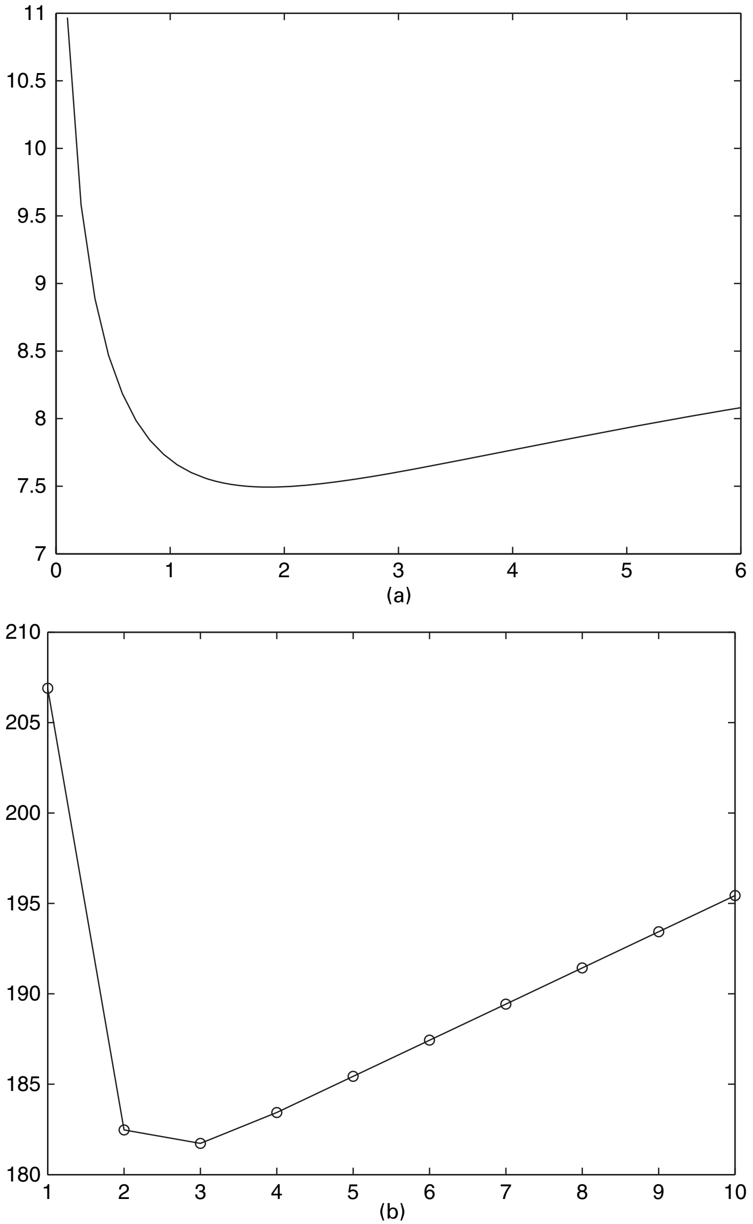
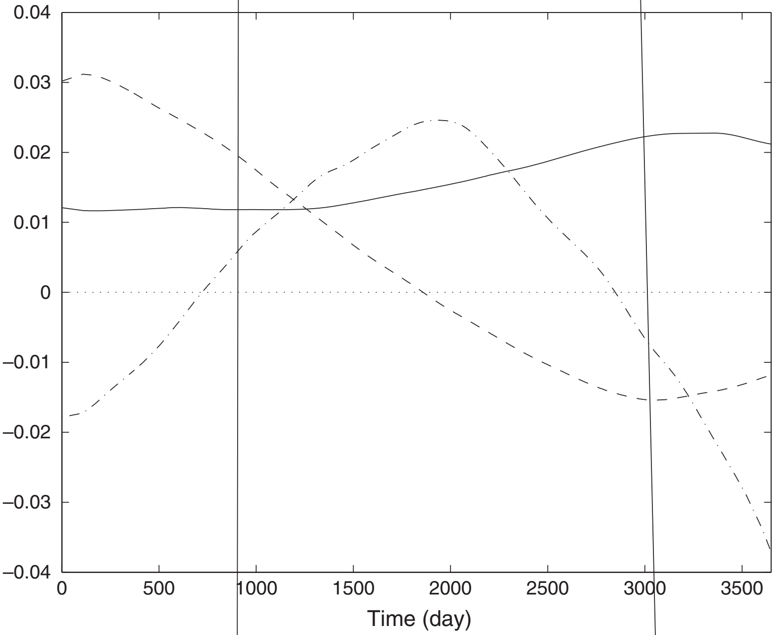
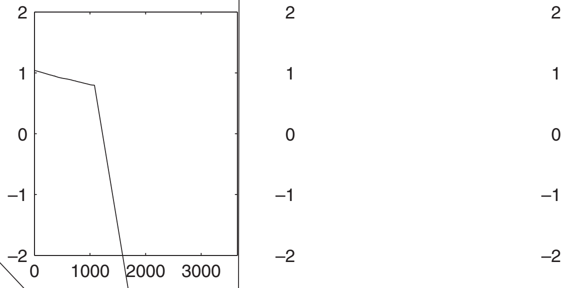
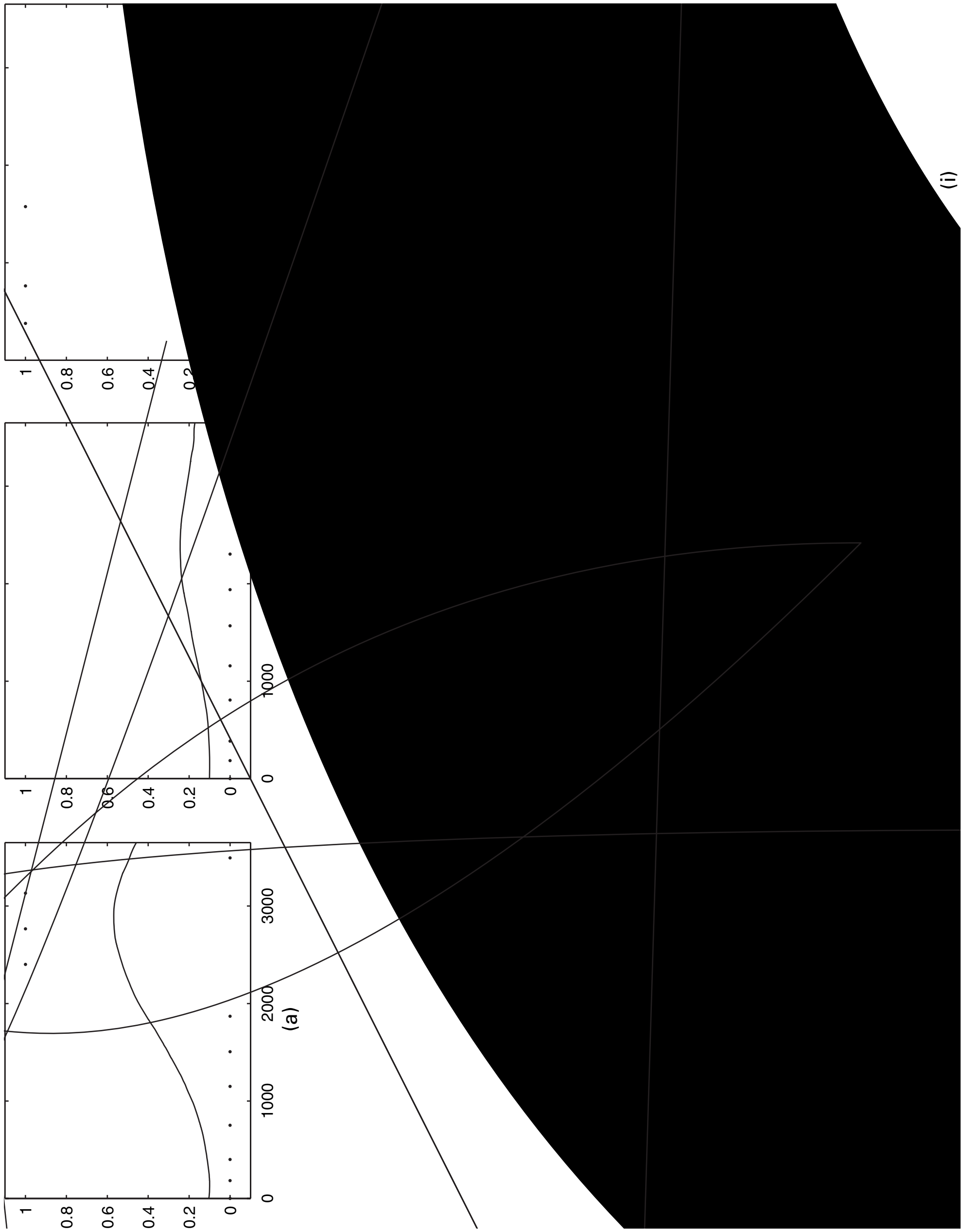


Fig. 2. (a) Plot of PE(γ^2) values (26) of the final iteration *versus* corresponding candidate values of γ^2, where $\hat{\gamma}^2$ minimizes PE(γ^2) and (b) FIC scores (28) for final iteration based on quasi-likelihood by using the binomial variance function for 10 possible leading eigenfunctions, where $M = 3$ is the minimizing value (for the primary biliary cirrhosis data)

We find that the overall trend of the predicted trajectories $Y_i(t)$ agrees well with the observed longitudinal binary outcomes, and leave-one-out analysis using equation (24) confirmed this. In making the comparison between observed data and fitted probabilities, we need to keep in mind that the Bernoulli observations consist of 0s or 1s, whereas the fitted probabilities and response processes are constrained to be strictly between 0 and 1. Therefore, long runs are expected for


0.04
0.03
0.02
0.01
0
–0.01
–0.02
–0.03
–0.04
0
500
1000
1500
2000
2500
3000
3500
Time (day)


2
1
0
–1
–2
0
1000
2000
3000
2
1
0
–1
–2
2
1
0
–1
–2


(a)
(i)
0
0.2
0.4
0.6
0.8
1
0
1000
2000
3000

6. Discussion

The assumption of small δ implies that the variation in the latent process X is assumed to be limited, according to the assumption $X(t) = \mu(t) + \delta Z(t)$. We note that the small δ assumption does not affect the methodology proposed, for which the value of δ is not needed and plays no role. The estimators proposed always target and are consistent for the unique LGP $\tilde{X}$, which is characterized by mean function $\nu(t)$ and covariance function $\tau(s,t)$, as defined in expression (8). However, biases may be accrued for response process estimates and especially predicting individual response trajectories for the case of large δ.

$$U_{qr}(s,t)=\sum_{i:m_i\geqslant 2}\sum_{j,k:j\neq k}\sum T_{ij}^q T_{ik}^r\,K_{ij}(s)\,K_{ik}(t),$$

$$\bar{T}_{qr}=U_{qr}/U_{00},$$

$$\bar{Z}=U_{00}^{-1}\sum_{i:m_i\geqslant 2}\sum_{j,k:j\neq k}\sum Z_{ijk}K_{ij}(s)K_{ik}(t),$$

$$R=R_{20}R_{02}-R_{11}^2,$$

$Z_{ijk}=Y_{ij}Y_{ik}$, $K_{ij}(t)=K\{(t-T_{ij})/h\}$, K is a kernel function and h a bandwidth. Of course, we would not use the same bandwidth to construct α and β; we expect the appropriate bandwidth for β to be larger than that for α.

Both α and β are conventional, except that diagonal terms are omitted when constructing the latter. The data within the ith block, i.e. $\mathcal{B}_i=\{Y_{ij}$ for $1\leqslant i\leqslant m_i\}$, are not independent of one another, but the n blocks, or trajectories $\mathcal{B}_1,\ldots,\mathcal{B}_n$ are independent. Therefore, a leave-one-trajectory-out version of cross-validation (Rice and Silverman, 1991) can be used to select the bandwidths for either estimator.

Appendix B: Positive definiteness of covariance estimation

Since the estimator $\tau(s,t)$ is symmetric, we may write

$$\tau(s,t)=\sum_{j=1}^{\infty}\theta_j\,\psi_j(s)\,\psi_j(t), \tag{34}$$

where (θ_j,ψ_j) are (eigenvalue, eigenfunction) pairs of a linear operator A in L^2 which maps a function f to the function $A(f)$, which is defined by $A(f)(s)=\int_{\mathcal{I}}\tau(s,t)f(t)\,\mathrm{d}t$. It is explained after equation (16) how these estimates are obtained. Assuming that only a finite number of the θ_js are non-zero, the operator A will be positive semidefinite or, equivalently, τ will be a proper covariance function, if and only if each $\theta_j\geqslant 0$. To ensure this property we compute equation (34) numerically and drop those terms that correspond to negative θ_js, giving the estimator

$$\tilde{\tau}(s,t)=\sum_{j\geqslant 1:\theta_j>0}\theta_j\,\psi_j(s)\,\psi_j(t). \tag{35}$$

The modified estimator $\tilde{\tau}$ is not identical to τ if one or more of the eigenvalues θ_j are strictly negative. In such cases, the estimator $\tilde{\tau}$ has strictly greater L_2-accuracy than τ, when viewed as an estimator of τ.

Theorem 1. Under regularity conditions, it holds that

$$\int_{\mathcal{I}^2}(\tilde{\tau}-\tau)^2\leqslant\int_{\mathcal{I}^2}(\tau-\tau)^2. \tag{36}$$

To prove this result, we show that condition (36) holds with strict inequality whenever $\tilde{\tau}$ is a non-trivial modification of τ, i.e. when $\tilde{\tau}\neq\tau$. In the series on the right-hand side of equation (34) we may, without loss of generality, order the terms so that those corresponding to non-zero θ_js are listed first, for $1\leqslant j\leqslant J$ say, and $\theta_j=0$ only for $j\geqslant J+1$. The sequence $\psi_1,\ldots,\psi_J$ is necessarily orthonormal, and we may choose $\psi_{J+1},\psi_{J+2},\ldots$ so that the full sequence $\psi_1,\psi_2,\ldots$ is orthonormal and also complete in the class of square integrable functions on $\mathcal{I}$.

We may therefore express the true covariance τ in terms of this sequence, as a conventional expansion in a generalized Fourier series:

$$\tau(s,t)=\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}a_{jk}\psi_j(s)\,\psi_k(t), \tag{37}$$

where $a_{jk}=\int_{\mathcal{I}^2}\tau(s,t)\,\psi_j(s)\,\psi_k(t)\,\mathrm{d}s\,\mathrm{d}t$. Expansions (34), (35) and (37) imply that

$$\int_{\mathcal{I}^2} (\tilde{\tau} - \tau)^2 = \sum_{j,\,k:\,j\neq k}\sum a_{jk}^2 + \sum_{j=1}^{\infty} (a_{jj} - \tilde{\theta}_j)^2,$$

$$\int_{\mathcal{I}^2} (\tau - \tau)^2 = \sum_{j,k:\,j\neq k}\sum a_{jk}^2 + \sum_{j=1}^{\infty} (a$$

$$\sigma_{ikl} \equiv \mathrm{cov}(\tilde{X}_{ik}, \tilde{X}_{il}) = \sum_j \theta_j \, \psi_j(T_{ik}) \, \psi_j(T_{il}) + \delta_{kl} \frac{\gamma^2 \, v[g\{\mu(T_{ik})\}]}{g^{(1)}\{\mu(T_{ik})\}^2},$$

where δ_{kl} equals 1 if $k=l$ and 0 otherwise, and

$$d_i \equiv \tilde{X}_i - E(\tilde{X}_i) = \left(\frac{Y_{i1} - g\{\mu(T_{i1})\}}{g^{(1)}\{\mu(T_{i1})\}}, \dots, \frac{Y_{i1} - g\{\mu(T_{i1})\}}{g^{(1)}\{\mu(T_{i1})\}} \right)^{\mathrm{T}}.$$

Denote $\mathrm{cov}(\tilde{X}_i, \tilde{X}_i)$ by $\Sigma_i = (\sigma_{ikl})_{1 \leqslant j, l \leqslant m_i}$. Then the explicit form of the matrices A_{ij} in equation (21) is given by

$$E(\xi_{ij}|Y_{i1}, \dots, Y_{im_i}) = \theta_j \psi_{i,j} \Sigma_i^{-1} d_i, \tag{39}$$

where estimates are substituted for μ by μ as in expression (15), γ by γ as in expression (27), and θ_j and ψ_j by the corresponding estimates for eigenvalues and eigenfunctions, derived from $\sigma(s,t)$ to obtain the estimated expression.

References

Boente, G. and Fraiman, R. (2000) Kernel-based functional principal components. *Statist. Probab. Lett.*, **48**, 335–345.

Cheng, M. Y., Hall, P. and Titterington, D. M. (1997) On the shrinkage of local linear curve estimators. *Statist. Comput.*, **7**, 11–17.

Chiou, J.-M. and Müller, H.-G. (2005) Estimated estimating equations: semiparametric inference for clustered and longitudinal data. *J. R. Statist. Soc.* B, **67**, 531–553.

Chiou, J. M., Müller, H. G. and Wang, J. L. (2004) Functional response models. *Statist. Sin.*, **14**, 675–693.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299–350.

Fan, J. (1993) Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, **21**, 196–216.

Fleming, T. R. and Harrington, D. P. (1991) *Counting Processes and Survival Analysis*. New York: Wiley.

Hashemi, R., Jacqmin-Gadda, H. and Commenges, D. (2003) A latent process model for joint modeling of events and marker. *Liftime Data Anal.*, **9**, 331–343.

Heagerty, P. J. (1999) Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.

Heagerty, P. J. and Kurland, B. F. (2001) Misspecified maximum likelihood estimation and generalized linear mixed models. *Biometrika*, **88**, 973–985.

Heagerty, P. J. and Zeger, S. L. (2000) Marginalized multilevel models and likelihood inference. *Statist. Sci.*, **15**, 1–26.

James, G., Hastie, T. G. and Sugar, C. A. (2001) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

James, G. and Sugar, C. A. (2003) Clustering for sparsely sampled functional data. *J. Am. Statist. Ass.*, **98**, 397–408.

Jowaheer, V. and Sutradhar, B. (2002) Analysing longitudinal count data with overdispersion. *Biometrika*, **89**, 389–399.

Kirkpatrick, M. and Heckman, N. (1989) A quantitative genetic model for growth, shape, reaction norms and other infinite-dimensional characters. *J. Math. Biol.*, **27**, 429–450.

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L. and Gips, C. H. (1994) Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, **20**, 126–134.

Pourahmadi, M. (2000) Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425–435.

Proust, C., Jacqmin-Gadda, H., Taylor, J. M. G., Ganiayre, J. and Commenges, D. (2006) A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*, **62**, 1014–1024.

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

Ramsay, J. and Silverman, B. (2005) *Functional Data Analysis*, 2nd edn. New York: Springer.

Rice, J. (2004) Functional and longitudinal data analysis: perspectives on smoothing. *Statist. Sin.*, **14**, 631–647.

Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B, **53**, 233–243.

Rice, J. and Wu, C. (2000) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.

Seifert, B. and Gasser, T. (1996) Finite-sample variance of local polynomials: analysis and solutions. *J. Am. Statist. Ass.*, **91**, 267–275.

Shi, M., Weiss, R. E. and Taylor, J. M. G. (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Appl. Statist.*, **45**, 151–163.

Staniswalis, J. G. and Lee, J. J. (1998) Nonparametric regression analysis of longitudinal data. *J. Am. Statist. Ass.*, **93**, 1403–1418.

Yao, F., Mnller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A. and Vogel, J. S. (2003) Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, **59**, 676–685.

Yao, F., Mnller, H. G. and Wang, J. L. (2005) Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.*, **100**, 577–590.

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