Modeling Sparse Generalized Longitudinal Observations With Latent Gaussian Processes

Peter Hall[1,2], Hans-Georg Müller[1,3] and Fang Yao[4]

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 proposed functional methods are nonparametric 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 nonparametric methods with longitudinal data on primary biliary cirrhosis and show in simulations that they are competitive in comparisons with Generalized Estimating Equations (GEE) and Generalized Linear Mixed Models (GLMM).

KEY WORDS: Binomial Data, Eigenfunction, Functional Data Analysis, Functional Principal Component, Prediction, Random Effect, Repeated Measurements, Smoothing, Stochastic Process.

---

[1] Department of Statistics, University of California, Davis, USA

[2] Department of Mathematics, University of Melbourne, Australia

[3] Corresponding author, e-mail: mueller@wald.ucdavis.edu; address: Department of Statistics, UC Davis, One Shields Avenue, Davis, CA 95616, USA.

[4] Department of Statistics, University of Toronto, Canada

1. Introduction

1.1. Preliminaries

When undertaking prediction in longitudinal data analysis involving irregularly-spaced and infrequent measurements, there is often relatively little information available about each subject, due 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 among measurements made at widely separated time points. We aim at a flexible nonparametric functional data analysis approach, which is in contrast to commonly used parametric models such as the generalized linear mixed models (GLMM) or generalized estimation equations (GEE) – see, e.g., Heagerty (1999) for recent discussions on applying such models to repeated binary measurements, Pourahmadi (2000) for related aspects of covariance modeling, 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 nonparametric 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/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 nonparametric approach to smooth individual trajectories in a first step is not feasible (Yao *et al.*, 2005). The problems 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 posed by such data for nonparametric approaches, by applying suitably modified methods of functional data analysis (FDA). FDA methods have been primarily developed for smooth and densely sampled data (Ramsay and Silverman, 2002, 2005). The basic idea to connect the data we wish to analyze to FDA methodology is to postulate an underlying latent Gaussian process (for other examples of latent process modeling for longitudinal studies compare, e.g., Diggle *et al.*, 1998; Jowaheer and Sutradhar, 2002; Hashemi, Jacqmin-Gadda and Commenges, 2003; 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 latent Gaussian process. Simulations indicate that the method is in practice quite insensitive to the Gaussian assumption for the latent process.

Since sufficiently flexible parametrizations 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 directly connect the latent Gaussian process to random trajectories for individual observations by means of a link function. These subject-specific trajectories correspond to the probabilities of a response in the binary response case. While 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 proposed methodology 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 proposed methods 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 nonparametric 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 due 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; James and Sugar, 2003). Our main tool from functional data analysis is functional principal component analysis, where observed trajectories are decomposed into a mean function and eigenfunctions (e.g., Rice and Silverman, 1991; 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 and Marron (2004); an early study of modeling longitudinal trajectories in biological applications with functional principal components is Kirkpatrick and Heckman (1989). Functional principal component analysis allows us to achieve three major goals: Dimension reduction of functional data by summarizing the data in a few functional principal components; the prediction of individual trajectories from sparse data, by estimating the functional principal component scores of the trajectories; and further statistical analysis of longitudinal data based on the functional principal component scores.

In the next subsection, we introduce the latent Gaussian process model, then in section 2 the proposed estimates, followed by applications to prediction (section 3). The results from a

simulation study, including a comparison of the proposed method with GLMM and GEE, 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 $1 \leq i \leq n$ and $1 \leq j \leq m_i$, 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_i's are uniformly bounded, if these quantities have a deterministic origin, or that they represent the values of independent and identically distributed random variables with sufficiently light tails, if the m_i's 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 dependency 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

$$E\big\{Y(t_1) \dots Y(t_m) \,\big|\, X\big\} = \prod_{j=1}^{m} g\{X(t_j)\}\,, \quad E\big\{Y(t)^2 \,\big|\, X\big\} \leq g_1\{X(t)\} \tag{1}$$

for $0 \leq t_1 < \ldots < t_m \leq 1$ 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. While 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 $1 \leq i \leq n$ and $1 \leq j \leq m_i$, 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_i's are supposed to be identically distributed as X. When interpreted for the data (T_{ij}, Y_{ij}), the model (1) implies that

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

The assumption that X at (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 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 one might select the logit link in the binary-data case, $g(x) = e^x/(1+e^x)$, and the log link for count data; under some circumstances, the link can also be estimated nonparametrically. An important special case of the model at (1) is that of binary responses, i.e., zero-one data, where the first identity in (1) simplifies to:

$$P\big\{Y(t_1) = \ell_1, \ldots, Y(t_m) = \ell_m \,\big|\, X\big\} = \prod_{j=1}^{m} g\{X(t_j)\}^{\ell_j} \, [1 - g\{X(t_j)\}]^{1-\ell_j} \,, \tag{3}$$

for all sequences $\ell_1, \ldots, \ell_m$ of zeros and ones. In this case, the link function g would be chosen as a distribution function and the proposed methodology corresponds to an extension of functional data analysis to longitudinal binary data.

2. Estimating mean and covariance of latent Gaussian processes

In order to use (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, for example 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 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 nonparametric flavor, we prefer not to make stronger assumptions than (1), and in particular 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

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

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

generic pair (X_i, Z_i), one has:

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

whence it may be deduced that

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

and

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

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

$$\alpha(t) = E\{g(X(t))\}, \quad \nu(t) = g^{-1}(\alpha(t)), \quad \tau(s,t) = \frac{\operatorname{cov}[g\{X(s)\}, g\{X(t)\}]}{g^{(1)}\{\mu(s)\}\, g^{(1)}\{\mu(t)\}}, \tag{8}$$

we obtain

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

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

These formulas immediately suggest estimators of μ and σ, if we are willing to neglect the effect of orders $O\big(\delta^2\big)$. 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

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

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

$$\widehat{\nu}(t) = g^{-1}(\hat{\alpha}(t)), \quad \hat{\tau}(s,t) = \frac{\hat{\beta}(s,t) - \hat{\alpha}(s)\,\hat{\alpha}(t)}{g^{(1)}\{\widehat{\nu}(s)\}\,g^{(1)}\{\widehat{\nu}(t)\}} \tag{13}$$

for

$$\nu(t) = g^{-1}(\alpha(t)) \;\text{ and }\; \tau(s,t) = \frac{\beta(s,t) - \alpha(s)\,\alpha(t)}{g^{(1)}\{\nu(s)\}\,g^{(1)}\{\nu(t)\}}, \tag{14}$$

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

$$\hat{\mu}(t) = \hat{\nu}(t), \quad \hat{\sigma}(s,t) = \hat{\tau}(s,t). \tag{15}$$

Note that these estimators do not depend on the constant δ, which therefore does not need to be known or estimated. While 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 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 the following, we work with the resulting estimators $\tilde{\tau}$ as defined in Appendix B. Properties of the estimators $\hat{\alpha}$, $\hat{\beta}$, $\widehat{\nu}$ and $\hat{\tau}$, defined at (32), (33) and (13), respectively, and of estimators $\widehat{\mu}$ and $\hat{\sigma}$ at (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 proposed functional data analysis model is dimension reduction through predicted functional principal component scores. These lead to predicted trajectories of the underlying hidden Gaussian process for the subjects in a study. Specifically, the predicted functional principal component 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 latent Gaussian process,

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

where ψ_j are the orthonormal eigenfunctions of the linear integral operator B with kernel $\sigma(s,t)$, that maps a L^2 function f to $Bf(s) = \int \sigma(s,t)f(t)\,dt$, i.e., the solutions of

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

where θ_j is the eigenvalue associated with eigenfunction ψ_j. The $\xi_{ij} = \int X_i(t)\psi_j(t)\,dt$ are the functional principal component (FPC) scores that play the role of random effects, with $E(\xi_{ij}) = 0$, $\mathrm{var}(\xi_{ij}) = \theta_j$, where θ_j is the eigenvalue corresponding to eigenfunction ψ_j. Once the estimator $\hat{\sigma}(s,t)$ (15) is determined, the corresponding estimates $\hat{\theta}_j$, $\hat{\psi}_j$ 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, focussing 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 (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 the following 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

$$Ee_{ik} = 0, \quad \mathrm{var}(e_{ik}) = \gamma^2 v[g\{X_i(T_{ik})\}]. \tag{19}$$

Here, γ^2 is an unknown variance (overdispersion) 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 the following, we implicitly condition on the measurement times T_{ij}.

With a Taylor expansion of g, using (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

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

(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

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

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

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

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,

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

is a reasonable predictor for the random effect ξ_{ij}, where $\tilde{X}_i = (X_i(T_{i1}) + \tilde{e}_{i1}, \ldots, X_i(T_{im_i}) + \tilde{e}_{im_i})^T$ 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 (21) is given in Appendix D.

3.2 Predicting trajectories

Motivated by (16) and (21), predicted trajectories for the latent Gaussian processes are obtained as

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

and predicted trajectories for the observed process Y as

$$\widehat{Y}_i(t) = \widehat{E}\{Y_i(t)|Y_{i1}, \ldots, Y_{im_i}\} = g\{\widehat{X}_i(t)\}, \tag{23}$$

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}, obtained by leaving that observation out, with Y_{ik} itself. Computing

$$\widehat{Y}_{ik}^{(-ik)} = \widehat{E}(Y_{ik}|Y_{i1},\ldots,Y_{i,k-1},Y_{i,k+1},\ldots,Y_{im_i}) = g\{\widehat{X}_i^{(-ik)}(T_{ik})\},\ 1 \le i \le n,\ 1 \le k \le m_i, \quad (24)$$

where

$$\widehat{X}_i^{(-ik)}(T_{ik}) = \widehat{\mu}(t) + \sum_{j=1}^{M} \widehat{E}(\xi_{ij}|Y_{i1},\ldots,Y_{i,k-1},Y_{i,k+1},\ldots,Y_{im_i})\widehat{\psi}_j(t), \quad (25)$$

we define the Pearson-type weighted prediction error

$$\mathrm{PE}(\gamma^2) = \sum_{i,k} \frac{(\widehat{Y}_{ik}^{(-ik)} - Y_{ik})^2}{v[g\{\widehat{X}_i^{(-ik)}(T_{ik})\}]}, \quad (26)$$

which will depend on the variance parameter γ^2 and implicitly also on the number of eigenfunctions M that are included in the model.

We found the following iterative selection procedure, for choosing the number of eigenfunctions M and the overdispersion parameter γ^2 simultaneously, to lead 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,

$$\hat{\gamma} = \mathrm{argmin}_\gamma\, \mathrm{PE}(\gamma^2). \quad (27)$$

Then in a subsequent step, update M by the criterion 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 $M = 2, 3$.

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 for functional data (compare 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

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

The penalty $2M$ corresponds to that used in AIC; other penalties such as those corresponding to 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 (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 preferrable to estimate it nonparametrically. This can be done via semiparametric quasi-likelihood regression (Chiou and Müller, 2005).

4. Simulation Results

4.1 Comparisons with GEE and GLLM

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 $\phi_1(t) = -\cos(\pi t/10)/\sqrt{5}$, $0 \le t \le 10$ with eigenvalues $\lambda_1 = 2$ ($\lambda_k = 0$, $k \ge 2$). 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 $\mathcal{N}(0, 2)$, while the ξ_{i1} for the non-Gaussian samples were simulated from a mixture of two normals, $\mathcal{N}(\sqrt{2}, 2)$ with probability 1/2 and $\mathcal{N}(-\sqrt{2}, 2)$ with probability 1/2. Binary outcomes Y_{ij} were generated as Bernoulli variables with the probability $E\{Y_{ij}|X_i(t_{ij}\} = g\{X_i(t_{ij})\}$, using the canonical 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 $\mathbf{1}_A(x) = 1$ if $x \in A$ and 0 otherwise for any set A. The number of eigenfunctions K and the overdispersion parameter γ^2 were separately selected for each run by the iteration (27), (28). These iterations converged fast, requiring only 2-4 iteration steps in most cases.

We compare the proposed nonparametric latent Gaussian process (LGP) method with the

popular parametric approaches provided by GLMM (Generalized Linear Mixed Models) and GEE (Generalized Estimating Equations). For GEE, we used the unstructured correlation option and both GEE and GLMM were run with linear (GEE-L and GLMM-L) and in addition with quadratic (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 $\mu = EX$ resp. $g(\mu)$ and predictions of subject-specific trajectories X_i resp. $g(X_i)$. The latter are available for LGP and GLMM, but not for GEE, which aims at marginal modeling. The specific criteria for the comparisons are as follows:

$$\text{XMSE} = \frac{\int_{\mathcal{I}} \{\hat{\mu}(t) - \mu(t)\}^2 dt}{\int_{\mathcal{I}} \mu^2(t) dt}, \quad \text{YMSE} = \frac{\int_{\mathcal{I}} [g\{\hat{\mu}(t)\} - g^{-1}\{\mu(t)\}]^2 dt}{\int_{\mathcal{I}} g^2\{\mu(t)\} dt}, \tag{29}$$

$$\text{XPE}_i = \frac{\int_{\mathcal{I}} \{\widehat{X}_i(t) - X_i(t)\}^2 dt}{\int_{\mathcal{I}} X_i^2(t) dt}, \quad \text{YPE}_i = \frac{\int_{\mathcal{I}} [g\{\widehat{X}_i(t)\} - g\{X_i(t)\}]^2 dt}{\int_{\mathcal{I}} g^2\{X_i(t)\} dt}, \tag{30}$$

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 proposed LGP method is not sensitive to the Gaussian assumption for latent processes. While 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 (PACE; see Yao et al., 2005). Secondly, the nonlinearity in the target functions throws the parametric methods off track, even when the more flexible quadratic fixed effects versions are used. We find that LGP conveys clear advantages in estimation and especially in predicting individual trajectories in such situations. While the parametric methods are sensitive to violations of assumptions, LGP 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) and (30), we also evaluated the estimation error for the single eigenfunction in the model (noting $\int_{\mathcal{I}} \phi_1^2(t) dt = 1$),

$$\text{EMSE} = \int_{\mathcal{I}} \{\hat{\phi}_1(t) - \phi_1(t)\}^2 dt. \tag{31}$$

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

		XMSE	XPE_i			YMSE	YPE_i		
			25th	50th	75th		25th	50th	75th
Gaussian	LGP	.1242	.1529	.2847	.7636	.0076	.0101	.0205	.0433
	GLMM-L	.4182	.3405	.5843	1.283	.0265	.0278	.0369	.0577
	GLMM-Q	.4323	.3479	.5990	1.319	.0271	.0285	.0377	.0584
	GEE-L	.4168	—	—	—	.0264	—	—	—
	GEE-Q	.4308	—	—	—	.0272	—	—	—
Non-Gaussian (Mixture)	LGP	.1272	.1664	.3166	.9556	.0078	.0109	.0228	.0459
	GLMM-L	.4209	.3309	.5943	1.364	.0266	.0280	.0372	.0589
	GLMM-Q	.4373	.3385	.6118	1.404	.0274	.0287	.0380	.0597
	GEE-L	.4227	—	—	—	.0268	—	—	—
	GEE-Q	.4396	—	—	—	.0277	—	—	—

Using the same simulation design as in subsection 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. 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.

	δ	XMSE	EMSE	XPE_i			YMSE	YPE_i		
				25th	50th	75th		25th	50th	75th.
Normal	.5	.1106	.7662	.1188	.1815	.3366	.0068	.0077	.0119	.0205
	.8	.1205	.3801	.1430	.2437	.5710	.0076	.0094	.0171	.0338
	1	.1280	.2434	.1513	.2809	.7857	.0077	.0101	.0203	.0431
	2	.1616	.0429	.2025	.3851	.8137	.0102	.0144	.0362	.0752
Mixture	.5	.1134	.7198	.1243	.1913	.3651	.0071	.0081	.0126	.0217
	.8	.1258	.3910	.1498	.2563	.6691	.0078	.0100	.0188	.0366
	1	.1323	.2256	.1624	.2986	.7944	.0081	.0113	.0227	.0450
	2	.1633	.0397	.2041	.3840	.8140	.0103	.0158	.0387	.0768

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 gets smaller, more and more of the variation in the observed data is due to error rather than the patterns of the underlying latent Gaussian process, and therefore it gets harder and harder to estimate the eigenfunction. This phenomenon is also observed in ordinary FPCA, 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. While large values of δ increase the errors in predicting individual trajectories, this is within expectations: For the predictor processes X, this is due to the fact that the variation of individual trajectories increases, while 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 due to the fact that 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 $\delta \leq 1$. 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 exact value has a small impact on the errors in mean function estimates and a modest impact on the errors in individual predictions, and 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 six months, one 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 proposed methods, we restrict the analysis to the participants who survived at least 10 years (3650 days) since they entered the study and were alive and not transplanted at the end of the tenth year. We carry out our analysis on the domain from 0 to 10 years, exploring the dynamic behavior 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 where 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 Figure 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 data range. We also provide pointwise bootstrap confidence intervals which broaden (not unexpectedly) near the endpoints 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

described in section 3.2 yielded the choices $M = 3$ for the number of included components and $\hat{\gamma}^2 = 1.91$ for the overdispersion parameter. The leave-one-point-out cross-validated prediction error $\mathrm{PE}(\gamma^2)$, as in (26), obtained for the final iteration (3rd iteration), is shown in the left panel of Figure 2 in dependence on γ^2, and the dependence of the FIC scores (28) on the number of included components M is shown in the right panel.

Smooth estimates of the first three eigenfunctions of the underlying Gaussian process X, resulting from the choices made in the iterative selection procedure, are shown in the left panels of Figure 3. The variation is mainly captured by the first two leading eigenfunctions. The first eigenfunction is roughly similar to the mean function, accounting for 74.2% of total variation, and the second eigenfunction essentially is a contrast between early and late times, explaining 23.2% of total variation.

The predicted trajectories $X_i(t)$, defined by (22), for the three patients with the largest projections in the directions of the respective eigenfunctions are shown in the top right panels of Figure 3. The original data and the predicted trajectories (23) are illustrated in the bottom right panels of that figure. Note that the sign of the eigenfunctions is arbitrary. These extreme cases clearly reveal how the individual trajectories X_i and Y_i are influenced by the dominant modes of variation. The predicted trajectories of $Y_i(t)$, obtained by (23) for nine randomly selected subjects, are shown in Figure 4. The predicted trajectories $\widehat{Y}_i(t)$ 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.

We find that the overall trend of the predicted trajectories $Y_i(t)$ agrees well with the observed longitudinal binary outcomes, and one-leave-out analysis using (24) confirmed this. In making the comparison between observed data and fitted probabilities, one needs to keep in mind that the Bernoulli observations consist of zeros or ones, while the fitted probabilities and response processes are constrained to be strictly between 0 and 1. Therefore, long "runs" are expected for extreme cases such as the one shown in the middle panel of the first row of Figure 4, where the fitted function is bound to be always larger than the data. Generally, in generalized response models, the variation in the data that corresponds to the conditional variance of the observations, given their Bernoulli probability, is in principle unexplained by any model, and only the probabilities themselves and their variation can be modeled, which may explain only a relatively small portion of the overall observed variation seen in the data.

To illustrate further statistical analysis after estimates for the FPC scores have been obtained, we regress the first two FPC scores of the underlying Gaussian process on the variable age at

entry into the study S. For this regression of response curves on a scalar predictor we use the model $E(X(t)|S) = \mu(t) + \sum_{j=1}^{M} E(\xi_j|S)\psi_j(t)$ (Chiou, Müller and Wang, 2004). We demonstrate the estimated regression functions $E(\hat{\xi}_j|S)$ for two components $j = 1, 2$ in Figure 5. The fits are obtained by local linear smoothing of the scatterplots $\hat{\xi}_j$ versus S by local linear smoothing. The regression fits indicate that the second FPC of the latent process is not much influenced by age at entry, while the first FPC remains flat for lower ages but then increases nonlinearly for ages after 45. For age at entry above 45, the conditional response curves therefore move increasingly upwards as age at entry increases, where the shape of the average increase corresponds to the first eigenfunction in Figure 3. This means older age at entry is associated with increasing probability of hepatomegaly.

6. Discussion

The assumption of small δ implies that the variation of 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 proposed methodology, for which the value of δ is not needed and plays no role. The proposed estimators always target and are consistent for the unique latent Gaussian process $\tilde{X}$, characterized by mean function $\nu(t)$ and covariance function $\tau(s,t)$, as defined in (8). On the other hand, biases may be accrued for response process estimates and especially predicting individual response trajectories for the case of large δ.

Processes $\tilde{X}$ characterize the data, and their functional principal component scores can be used for further statistical analysis. When δ is small, then $X \sim \tilde{X}$ so that (1)-(3) are satisfied (approximately) for $\tilde{X}$ as well. While the proposed approach is always useful to represent the data, even in the case where δ is not small, the small δ assumption is needed to obtain reasonably accurate estimates of probability trajectories $Y(t)$.

Simulation results demonstrate that the proposed methodology outperforms classical parametric models such as GEE and GLMM in situations where their parametric assumptions do not apply. The proposed nonparametric method relies on far fewer assumptions which makes it more universally applicable. Further statistical analysis such as exploring the effect of subject-specific covariates can be based on the estimated functional principal component scores. We note that in the data example, mean function and subject-specific trajectories are highly nonlinear, emphasizing the need for nonparametric methodology to analyze such data.

Acknowledgements

We wish to thank two reviewers for insightful comments which led to many improvements. This research was supported in part by NSF grants DMS03-54448 and DMS05-05537.

Appendix A: Local linear smoothers

Local-linear versions of the estimators $\hat{\alpha}$ and $\hat{\beta}$, introduced in Section 2.1, are given explicitly by

$$\hat{\alpha}(t) \;=\; \frac{P_2(t)\,Q_0(t) - P_1(t)\,Q_1(t)}{P_0(t)\,P_2(t) - P_1(t)^2}, \tag{32}$$

$$\hat{\beta}(s,t) \;=\; \bar{Z} + \frac{1}{R}\left(\frac{s-\bar{T}_{10}}{h}, \frac{t-\bar{T}_{01}}{h}\right) \begin{pmatrix} R_{02} & - & R_{11} \\ -R_{11} & & R_{20} \end{pmatrix} \begin{pmatrix} S_{10} \\ S_{01} \end{pmatrix}, \tag{33}$$

where

$$\begin{aligned}
P_r(t) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} (t-T_{ij})^r\,K_{ij}(t)\,, \quad Q_r(t) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} (t-T_{ij})^r\,Y_{ij}\,K_{ij}(t)\,, \\
R_{qr}(s,t) &= \sum_{i\,:\,m_i\geq 2}\;\sum_{j,k\,:\,j\neq k} \left\{\frac{T_{ij}-\bar{T}_{10}(s,t)}{h}\right\}^q \left\{\frac{T_{ik}-\bar{T}_{01}(s,t)}{h}\right\}^r K_{ij}(s)\,K_{ik}(t)\,, \\
S_r(s,t) &= \sum_{i\,:\,m_i\geq 2}\;\sum_{j,k\,:\,j\neq k} \{Z_{ijk}-\bar{Z}(s,t)\} \left\{\frac{T_{ij}-\bar{T}_{10}(s,t)}{h}\right\}^q \left\{\frac{T_{ik}-\bar{T}_{01}(s,t)}{h}\right\}^r \\
&\qquad\qquad \times K_{ij}(s)\,K_{ik}(t)\,, \\
U_{qr}(s,t) &= \sum_{i\,:\,m_i\geq 2}\;\sum_{j,k\,:\,j\neq k} T_{ij}^q\,T_{ik}^r\,K_{ij}(s)\,K_{ik}(t)\,, \qquad \bar{T}_{qr} = U_{qr}/U_{00}\,, \\
\bar{Z} &= U_{00}^{-1} \sum_{i\,:\,m_i\geq 2}\;\sum_{j,k\,:\,j\neq k} Z_{ijk}\,K_{ij}(s)\,K_{ik}(t)\,, \quad R = R_{20}\,R_{02} - R_{11}^2\,,
\end{aligned}$$

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

Both $\hat{\alpha}$ and $\hat{\beta}$ are conventional, except that diagonal terms are omitted when constructing the latter. The data within the ith block, i.e. $\mathcal{B}_i = \{Y_{ij} \text{ for } 1 \leq i \leq m_i\}$, 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 $\hat{\tau}(s,t)$ is symmetric, we may write

$$\hat{\tau}(s,t) = \sum_{j=1}^{\infty} \hat{\theta}_j \, \widehat{\psi}_j(s) \, \widehat{\psi}_j(t) \,, \tag{34}$$

where $(\hat{\theta}_j, \widehat{\psi}_j)$ are (eigenvalue, eigenfunction) pairs of a linear operator A in L^2 which maps a function f to the function $A(f)$, defined by $A(f)(s) = \int_{\mathcal{I}} \hat{\tau}(s,t) \, f(t) \, dt$. It is explained after equation (16) how these estimates are obtained. Assuming that only a finite number of the $\hat{\theta}_j$'s are nonzero, the operator A will be positive semi-definite, or equivalently, $\hat{\tau}$ will be a proper covariance function, if and only if each $\hat{\theta}_j \geq 0$. To ensure this property we compute (34) numerically, and drop those terms that correspond to negative $\hat{\theta}_j$'s, giving the estimator

$$\tilde{\tau}(s,t) = \sum_{j \geq 1 \,:\, \hat{\theta}_j > 0} \hat{\theta}_j \, \widehat{\psi}_j(s) \, \widehat{\psi}_j(t) \,. \tag{35}$$

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

Theorem 1. *Under regularity conditions, it holds that*

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

In order to prove this result, we show that (36) holds with strict inequality whenever $\tilde{\tau}$ is a nontrivial modification of $\hat{\tau}$, i.e. when $\tilde{\tau} \neq \hat{\tau}$. In the series on the right-hand side of (34) we may, without loss of generality, order the terms so that those corresponding to nonzero $\hat{\theta}_j$'s are listed first, for $1 \leq j \leq J$ say, and $\hat{\theta}_j = 0$ only for $j \geq J+1$. The sequence $\widehat{\psi}_1, \ldots, \widehat{\psi}_J$ is necessarily orthonormal, and we may choose $\widehat{\psi}_{J+1}, \widehat{\psi}_{J+2}, \ldots$ so that the full sequence $\widehat{\psi}_1, \widehat{\psi}_2, \ldots$ 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:

$$\tau(s,t) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} \, \widehat{\psi}_j(s) \, \widehat{\psi}_k(t) \,, \tag{37}$$

where $a_{jk} = \int_{\mathcal{I}^2} \tau(s,t) \, \widehat{\psi}_j(s) \, \widehat{\psi}_k(t) \, ds \, dt$. Expansions (34), (35) and (37) imply that

$$\int_{\mathcal{I}^2} (\tilde{\tau} - \tau)^2 = \sum_{j,k \,:\, j \neq k} \sum a_{jk}^2 + \sum_{j=1}^{\infty} (a_{jj} - \tilde{\theta}_j)^2, \quad \int_{\mathcal{I}^2} (\hat{\tau} - \tau)^2 = \sum_{j,k \,:\, j \neq k} \sum a_{jk}^2 + \sum_{j=1}^{\infty} (a_{jj} - \hat{\theta}_j)^2 \,, \tag{38}$$

where $\tilde{\theta}_j = \hat{\theta}_j$ if $\hat{\theta}_j \geq 0$ and $\tilde{\theta}_j = 0$ otherwise. The fact that τ is a proper covariance function, and so enjoys the positive-semidefiniteness property, implies that $a_{jj} \geq 0$ for each j. Result (36) follows from this property and (38).

Appendix C: Some theoretical properties of estimators (32), (33), (13) and (15)

Standard arguments show that local-linear forms of the estimators $\hat{\alpha}$ and $\hat{\beta}$, given in Appendix A, converge to α and β at mean-square rates $\rho_\alpha(h) = (nh)^{-1} + h^4$ and $\rho_\beta(h) = (nh^2)^{-1} + h^4$, respectively, where h denotes the bandwidth used to construct either estimator. Therefore, the optimal bandwidths are of sizes $n^{-1/5}$ and $n^{-1/6}$, respectively, and the optimal mean-square convergence rates are $n^{-4/5}$ and $n^{-2/3}$, for $\hat{\alpha}, \widehat{\nu}$ and $\hat{\beta}$, respectively. Hence, in view of the manner of construction (13) of $\hat{\tau}$ in terms of $\hat{\alpha}$ and $\hat{\beta}$, the optimal mean-square convergence rate of $\hat{\tau}$ to τ is also $n^{-2/3}$. To obtain these results it is necessary to incorporate a small ridge parameter into the denominators of estimators, to guard against difficulties with data sparsity among the observation times T_{ij}. The ridge may be taken as small as n^{-c}, for sufficiently large $c > 0$. Adjustments of this type are common for local-linear estimators (Fan, 1993; Seifert and Gasser, 1996; Cheng, Hall and Titterington, 1997).

The above results are exact, for example in the sense that upper and lower bounds to mean squared errors of $\hat{\alpha}, \widehat{\nu}$ and for $\hat{\beta}, \hat{\tau}$ are of sizes $\rho_\alpha(h)$ and $\rho_\beta(h)$, respectively, provided the m_i's are uniformly bounded and the number of m_i's that strictly exceed 1 is bounded above a constant multiple of n. However, the mean squared errors will not admit standard asymptotic formulae, for example $\rho_\beta(h) \sim C_1\,(nh^2)^{-1} + C_2\,h^4$ for positive constants C_1 and C_2, unless additional conditions are imposed to ensure, for instance, that the m_i's that strictly exceed 1, and the proportion of times that they exceed 1, have well-defined long-run "average" values in an appropriate sense. It is sufficient, but not necessary, that the m_i's represent conditioned-upon values of independent and identically distributed random variables distributed as the integer-valued variable M, where $P(M \geq 2) > 0$ and, for some integer $k \geq 2$, $P(M \leq k) = 1$. Additionally, more conventional regularity conditions should be assumed. In particular, both α and β should have two continuous derivatives, and the moment conditions (1) should hold. Standard methods may also be used to show that leave-one-block-out cross-validation achieves asymptotic optimality, in the estimation of α and β, to first order and in an L_2 sense.

We remark that if we leave the longitudinal situation, and contrary to what we assumed before and the conditions we discussed earlier, assume for a moment the m_i's can take very large values,

with high frequency as n increases, then convergence rates can be faster than those discussed above. In particular, if the number of values of $m_1, \ldots, m_n$ that exceed a divergent quantity is bounded below by a fixed constant multiple of n; that is, if $\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I\{m_i > p(n)\} > 0$, where $p(n) \to \infty$ and $I(\cdot)$ denotes the indicator function of the indicated property; then the mean squared errors of $\hat{\alpha}, \widehat{\nu}$ and of $\hat{\beta}, \hat{\tau}$ equal $o\{(nh)^{-1}\} + O(h^4)$ and $o\{(nh^2)^{-1}\} + O(h^4)$, respectively, rather than simply the values $O\{(nh)^{-1} + h^4\}$ and $O\{(nh^2)^{-1} + h^4\}$ discussed in the second paragraph of this section. In these formulas the terms in $(nh)^{-1}$ and $(nh^2)^{-1}$ represent variance contributions to mean squared error. The fact that variance contributions are of relatively small order if the proportion of large m_i's is sufficiently high reflects the additional information that is available about the process X_i in such cases.

Appendix D: Details for (21)

Let $\tilde{X}_i = (\tilde{X}_{i1}, \ldots, \tilde{X}_{im_i})^T$ and $\psi_{i,j} = (\psi_j(T_{i1}), \ldots, \psi_j(T_{im_i}))^T$, referring to expansion (16). One has $\mathrm{cov}(\xi_{ij}, \tilde{X}_i) = \theta_j \psi_{i,j}^T$,

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

where $\delta_{kl} = 1$ if $k = l$ and 0 otherwise, and

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

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

$$\widehat{E}(\xi_{ij} | Y_{i1}, \ldots, Y_{im_i}) = \hat{\theta}_j \hat{\psi}_{i,j} \widehat{\Sigma}_i^{-1} \hat{d}_i, \tag{39}$$

where we substitute μ by $\widehat{\mu}$ at (15), γ by $\hat{\gamma}$ at (27), and θ_j, ψ_j by the corresponding estimates for eigenvalues and eigenfunctions, derived from $\hat{\sigma}(s,t)$ to obtain the estimated version.

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/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. Sinica*, **14**, 675-693.

De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I. and De Keyser, P. (1998) Twelve weeks of continuous onychomycosis caused by dermatophytes: a double blind comparative trial of terbafine 250mg/day versus itraconazole 200mg/day. *J. Am. Acad. Derm.*, **38**, S57-S63.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299350.

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.* Wiley, New York.

Hashemi, R., Jacqmin-Gadda H. and Commenges D. (2003) A latent process model for joint modeling of events and marker. *Lifetime 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. Assoc.*, **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.

Lesaffre, E. and Spiessens, B. (2001) On the effect of the number of quadrature points in a logistic random-effects model: an example. *Appl. Statist.*, **50**, 325-335.

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*, Second Edition. New York: Springer.

Rice, J. (2004) Functional and longitudinal data analysis: Perspectives on smoothing. *Statist. Sinica*, **14**, 631-647.

Rice, J. and Silverman, B. (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. Assoc.*, **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. Assoc.*, **93**, 1403-1418.

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

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

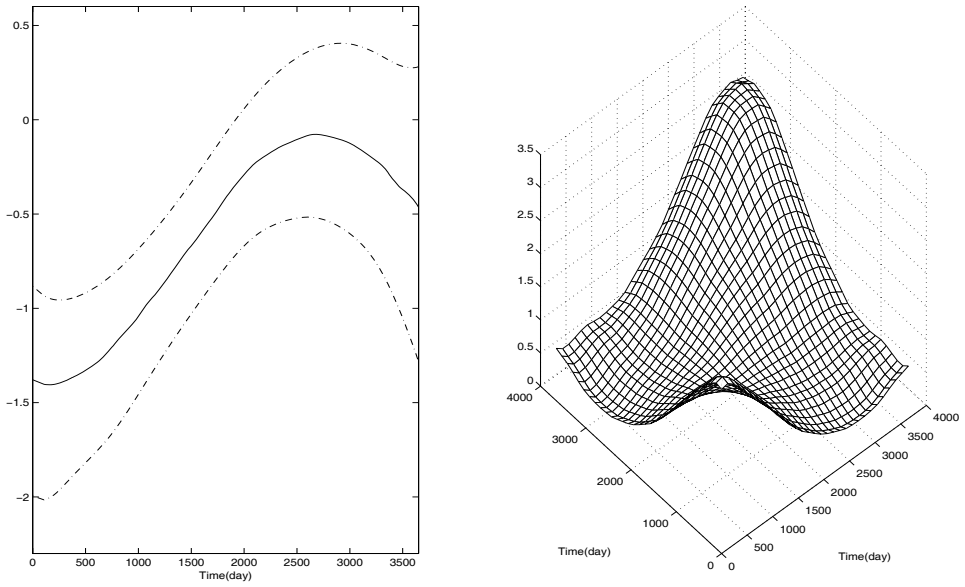

Figure 1: Left panel: Smooth estimate $\widehat{\mu}(t)$ (15) of the mean function of the latent process $X(t)$ with pointwise 95% bootstrap confidence intervals. Right panel: Smooth estimate of the covariance function $\hat{\sigma}(s,t)$ of $X(t)$ (for PBC data).

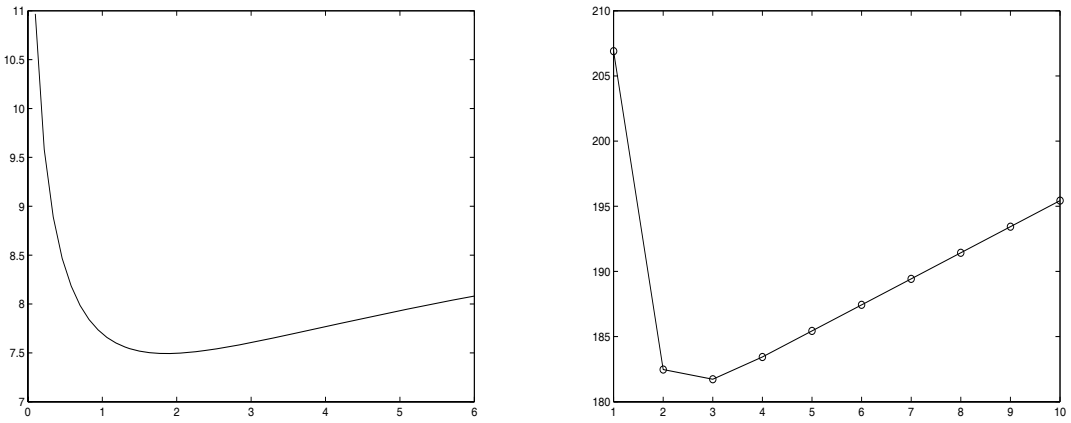

Figure 2: Left panel: Plot of PE(γ^2) values (26) of the final iteration versus corresponding candidate values of γ^2, where $\hat{\gamma}^2 = 1.91$ minimizes PE(γ^2). Right panel: FIC scores (28) for final iteration based on quasi-likelihood using the binomial variance function for 10 possible leading eigenfunctions, where $M = 3$ is the minimizing value (for PBC data).

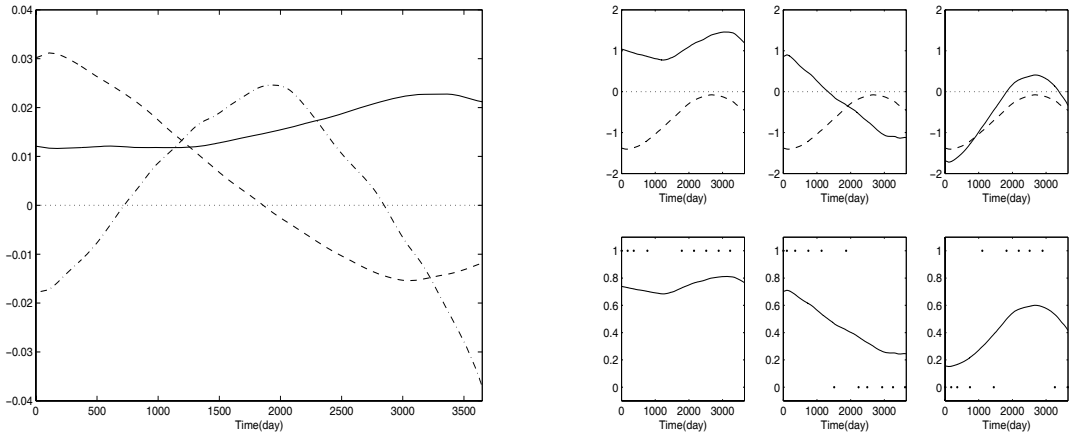

Figure 3: Left panels: Smooth estimates of the first (solid), second (dashed) and third (dash-dotted) eigenfunctions which explain 69.6%, 26.0% and 3.9% of total variation. Top right panels: Predicted trajectories of $X_i(t)$ (solid) as in (22) for the three individuals with the largest projections on the respective eigenfunctions in the left panel, overlaid with the overall estimated mean function (dashed). Bottom right panels: Observations (dots) and predicted trajectories of $Y_i(t)$ as given in (23), corresponding to the above three subjects (for PBC data).

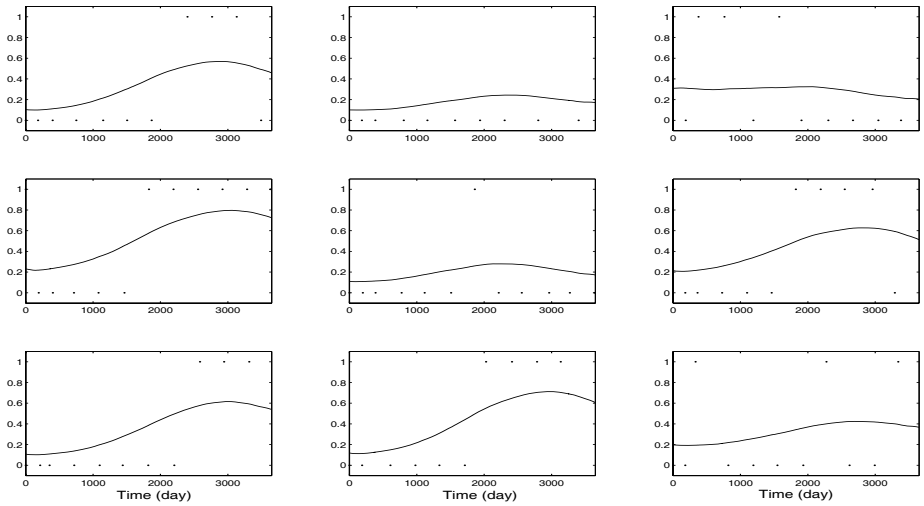

Figure 4: Observed responses (dots) and predicted subject-specific trajectories obtained as in (23) for nine randomly selected subjects (PBC data).

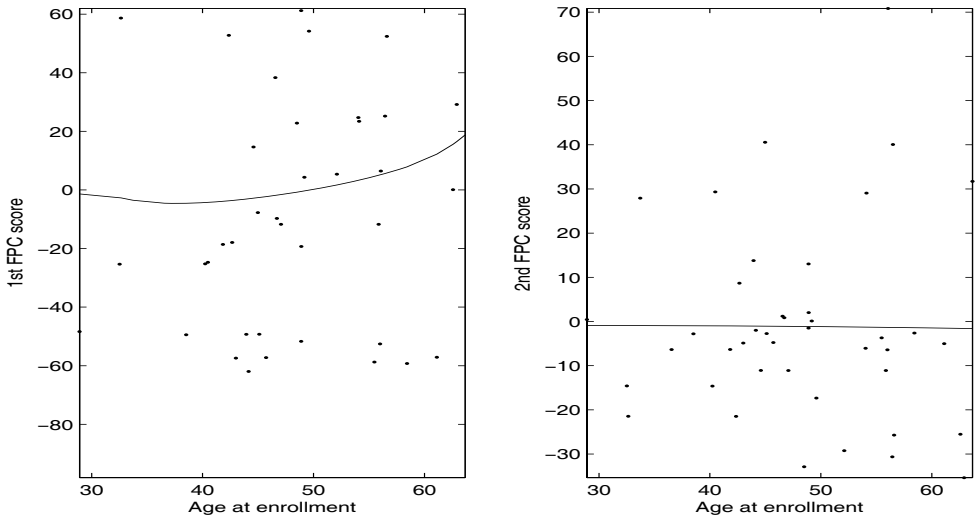

Figure 5: Scatterplot (dots) and fitted nonparametric regression of the first (left) and second (right) FPC scores on age at enrollment into the PBC study.