Published in final edited form as:

Can J Stat. 2010 June ; 38(2): 256–270. doi:10.1002/cjs.10062.

Longitudinal functional principal component modeling via Stochastic Approximation Monte Carlo

Josue G. Martinez, Faming Liang, Lan Zhou, and Raymond J. Carroll

Department of Statistics Texas A&M University College Station, Texas USA, 77843-3143

Abstract

The authors consider the analysis of hierarchical longitudinal functional data based upon a functional principal components approach. In contrast to standard frequentist approaches to selecting the number of principal components, the authors do model averaging using a Bayesian formulation. A relatively straightforward reversible jump Markov Chain Monte Carlo formulation has poor mixing properties and in simulated data often becomes trapped at the wrong number of principal components. In order to overcome this, the authors show how to apply Stochastic Approximation Monte Carlo (SAMC) to this problem, a method that has the potential to explore the entire space and does not become trapped in local extrema. The combination of reversible jump methods and SAMC in hierarchical longitudinal functional data is simplified by a polar coordinate representation of the principal components. The approach is easy to implement and does well in simulated data in determining the distribution of the number of principal components, and in terms of its frequentist estimation properties. Empirical applications are also presented.

Keywords

Functional data analysis; Hierarchical models; Longitudinal data; Markov chain monte carlo; Principal Components; Stochastic approximation

1. INTRODUCTION

Because longitudinal data can be viewed as sparsely observed functional data (Rice, 2004), functional principal components analysis becomes an important tool for analyzing longitudinal data. Since the early contributions of Besse & Ramsay (1986), Ramsay & Dalzell (1991), and Rice & Silverman (1991), a substantial literature has developed for functional principal components analysis, including but not limited to the work of Silverman (1996), Cardot (2000), James, Hastie & Sugar (2000), Rice & Wu (2001), Boente & Fraiman (2002), Ramsay & Silverman (2005, Chapters 8–10), Yao, Müller & Wang (2005ab), Hall & Hosseini-Nasab (2006), Hall, Müller & Wang (2006), Jank & Shmueli (2006), Ocaña, Aguilera & Escabias (2007), Reiss & Ogden (2007), and Huang, Shen & Buja (2008).

Let $Y_{i\ell}$ be the ℓ^{th} response for the i^{th} subject at time $t_{i\ell}$. Then the hierarchical model we are considering takes the general form

Josue G. Martinez: jgmartinez@stat.tamu.edu
 Faming Liang: fliang@stat.tamu.edu
 Lan Zhou: lzhou@stat.tamu.edu
 Raymond J. Carroll: carroll@stat.tamu.edu

$$Y_{i\ell} = \mu(t_{i\ell}) + g_i(t_{i\ell}) + \epsilon_{i\ell}, \quad (1)$$

where $\mu(\cdot)$ is the fixed effects population-level function, $g_i(\cdot)$ represents random functions associated with each individual, and $\epsilon_{i\ell}$ is noise. The goal is to estimate the fixed-effects function $\mu(\cdot)$ and to understand the variability of the random effects functions $g_i(\cdot)$.

It is standard in the functional data analysis literature to model the correlations among responses at different times through a mixed effects model, e.g., via the Karhunen-Lve expansion used in a number of the papers referenced above. As a modification of work by James et al. (2000) and Rice & Wu (2001), to fit models such as (1), Zhou, Huang & Carroll (2008) introduced penalized spline estimation of functional principal components for longitudinal data using a mixed effects model framework. Suppose that there are $i = 1, \dots, n$ subjects, each of which contribute $\ell = 1, \dots, m_i$ observations at times or locations $t_{i\ell}$. Let $b(t)$ be a $p \times 1$ orthogonal basis with the property that $\int b(t)b^T(t)dt = I$, and let B_i be a $m_i \times p$ matrix with ℓ^{th} row $b^T(t_{i\ell})$. Let Y_i be the $m_i \times 1$ vector of responses for the i^{th} observation. Then the model becomes

$$\begin{aligned} Y_i &= B_i \theta_\mu + B_i \Theta_f \alpha_i + \epsilon_i; \\ E(\epsilon_i | B_i) &= 0; \\ \text{cov}(\epsilon_i | B_i) &= \sigma_\epsilon^2 I_{m_i}; \\ \alpha_i &= \text{Normal}\{0, D_\alpha = \text{diag}(d_1, \dots, d_M)\}, \end{aligned} \quad (2)$$

where the principal components matrix Θ_f is $p \times M$ with $p > M$, the main effects parameter θ_μ is $p \times 1$, and the vector of principal components scores α_i is $M \times 1$. In addition to the orthogonality constraint on the basis functions, for identifiability we need Θ_f to be orthogonal, $\Theta_f^T \Theta_f = I_M$. Thus, comparing (1) and (2), we are making the approximations that $\mu(t) \approx b^T(t)\theta_\mu$ and $g_i(t) \approx b^T(t)\Theta_f \alpha_i$.

It is interesting to note that in common with the Karhunen-Lve approaches, as in ours, if the number of principal components is $M = 1$, then the correlation structure of responses at different times is equicorrelated, but this is not the case when $M > 1$.

As in any kind of spline-based analysis, it is useful to think about penalizing the likelihood to ensure smoothness of fit and to avoid overfitting. In model (2), we have two functions that need to be penalized: the fixed effects functions represented as $b^T(t)\theta_\mu$ and the random effects functions represented as $b^T(t)\Theta_f \alpha_i$. For the fixed effects function, for example, the usual device is to have a penalty parameter, here called λ_μ , and to penalize the likelihood by $\lambda_\mu \theta_\mu^T K \theta_\mu$, where K is a so-called penalty matrix. The typical form of the penalty matrix is the integrated square of the second derivative matrix, i.e., $K = \int b''(t)b''(t)^T dt$. Methods for constructing the orthogonal basis functions B_i and for computing the penalty matrix K are described by Zhou et al. (2008).

Enforcing smoothing of the random effects functions is somewhat more involved, because $b^T(t)\Theta_f$ has M columns, each of which needs to be penalized. Here is the device used by Zhou, et al. (2008). If ϑ_j is the j^{th} column of Θ_f , the penalty for the random functions

becomes $\lambda_f \sum_{j=1}^M \vartheta_j^T K \vartheta_j$. Overall then, to ensure smoothness, the log likelihood is penalized by $\lambda_\mu \theta_\mu^T K \theta_\mu + \lambda_f \sum_{j=1}^M \vartheta_j^T K \vartheta_j$.

Of course, penalty parameters need to be estimated, and for this purpose there are a variety of possibilities. Zhou, et al. (2008) used 10-fold crossvalidation to estimate (λ_μ, λ_f) for any given M . To choose the number of principal components, they use an informal ad hoc procedure, see their Section 5.2. James et al. (2000) used the informal %-variation explained and a likelihood ratio test, although they found that the performance of the latter was “ambiguous” and they recommended the use of the former.

Especially when the number of subjects is small, there will be ambiguity as to the choice of the number of principal components. We propose to acknowledge this ambiguity by taking a Bayesian approach to the problem.

An outline of this paper is as follows. In Section 2 we present the basic algorithm. Section 3 describes simulation studies, while Section 4 investigates empirical examples. Concluding remarks are presented in Section 5. Technical details are given in an appendix.

2. ALGORITHM

2.1 Parameters of the Basic Model and their Posterior Distributions

The model is as described at (2).

In what follows, let $\sigma_\mu^2 = \lambda_\mu^{-1}$ and $\sigma_f^2 = \lambda_f^{-1}$. It is convenient to work with the parameterization $\xi_1 = \log(\sigma_\epsilon)$, $\xi_2 = \log(\sigma_\mu)$ and $\xi_3 = \log\{\text{diag}(D_\alpha)\}$. We write the number of principal components as M .

As described in Section 2.2, because of the restriction that $\Theta_f^T \Theta_f = I$, we will parameterize Θ_f in terms of auxiliary parameters Ω_f . If $\alpha = (\alpha_1, \dots, \alpha_n)$, the main effects vector is θ_μ and $\mathbf{Y}$ denotes the response data, then except for constants of proportionality, we write the posterior distribution of $(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2, \theta_\mu, \alpha)$ given $(\mathbf{Y})$ as $f(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2, \theta_\mu, \alpha | \mathbf{Y})$, see the Appendix for details.

In formulating this problem, we face a number of challenges.

- By definition, Θ_f is constrained to be orthogonal. We achieve this orthogonality by combining the polar coordinate approach of Zhang and Davidian (2001) with a Gram-Schmidt transformation, see Section 2.2.
- Even though Θ_f is orthogonal by construction, it still needs to be smoothed.
- Because the models are of different dimensions, a basic reversible jump type algorithm is attractive (Green, 1995). However, we observed difficulty in accepting the death step of such an algorithm, and in order to explore the model space chose instead to implement the Stochastic Approximation Monte Carlo or SAMC algorithm of Liang, Liu & Carroll (2007).

2.2 Parameterizing Θ_f

In this section, we show how to write Θ_f as a function of other parameters Ω_f , in such a way that Θ_f is orthogonal, and all elements of its first row are positive, as required by the identifiability result in Zhou et al. (2008).

Because Θ_f is an orthogonal $p \times M$ matrix with $p > M$, we turn to a polar coordinate representation, handily provided by Zhang and Davidian (2001). Let $\Theta_f = (\boldsymbol{\vartheta}_1, \dots, \boldsymbol{\vartheta}_M)$ where $p > M + 1$, each $\boldsymbol{\vartheta}_j$ is $p \times 1$, and define the $p \times 1$ vector s_j as follows. Its first element

is $\sin(\gamma_{1j})$, and its k^{th} element for $k = 2, \dots, p - 1$ is $\sin(\gamma_{kj}) \prod_{\ell=1}^{k-1} \cos(\gamma_{\ell j})$. Finally, its p^{th}

element is $\prod_{\ell=1}^{p-1} \cos(\gamma_{\ell j})$. These are thus polar coordinates in $p-1$ parameters. Take $-\pi/2 \leq \gamma_{kj} \leq \pi/2$ for all (k, j) except $k=j=1$, and let $0 \leq \gamma_{11} \leq \pi/2$ to force identifiability, see Zhou et al. (2008) for details. The matrix s_j is of unit length: $s_j^T s_j = \|s_j\|^2 = 1$. Define

$\gamma_j^T = (\gamma_{1j}, \dots, \gamma_{(p-1)j})$ and $\Omega_f = (\gamma_1, \dots, \gamma_M)$, then we can generate γ_{kj} for $k=1, \dots, p-1$ and $j=1, \dots, M$ as uniform on their support.

To form an orthogonal $p \times M$ matrix, we use a Gram-Schmidt transformation. Define $\text{proj}_e(s) = (s^T e)e$. Then $\vartheta_1 = s_1$, for $k=2, \dots, M$, define $\vartheta_{k*} = \left(I - \sum_{j=1}^{k-1} \vartheta_j \vartheta_j^T\right) s_k$, and $\vartheta_k = \vartheta_{k*} / \|\vartheta_{k*}\|$. Finally, take $\Theta_f = (\vartheta_1, \dots, \vartheta_M)$.

Write $e^T = (1, 0, \dots, 0)$. Note that the first element of ϑ_{k*} , $e^T \vartheta_{k*}$ is $e^T \left(I - \sum_{j=1}^{k-1} \vartheta_j \vartheta_j^T\right) s_k$. To force the first element of ϑ_k to be positive, we set $\vartheta_k = \text{sign}(e^T \vartheta_{k*}) \vartheta_{k*} / \|\vartheta_{k*}\|$.

2.3 Basic Reversible Jump Algorithm

Let M denote the number of principal components, and assume that it has a uniform prior distribution. We marginalized the posterior by integrating out θ_μ and α , resulting in the function $g(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2 | \mathbf{Y})$ given in equation (6) in the Appendix. Let $(\Omega_f^{(v)}, \xi_3^{(v)}, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)})$ denote the sample obtained at iteration v . Below we outline the algorithm assuming that there are only three possible principal components; however, this can easily be extended to include any number of principal components.

2.3.1 Initial Step—In this step, we decide which component of the model to update at iteration $v+1$. We let “MH” here denote a Metropolis-Hastings step, see Section 2.3.2. Birth and death steps are described in Sections 2.3.3 and 2.3.4, respectively.

- Draw $U = \text{Uniform}(0, 1)$.
- If $M = 1$ and $U < 0.5$ do “birth” step and set $\omega_{12} = 0.5$; if $M = 1$ and $U \geq 0.5$ do “MH” step.
- If $M = 2$ and $U < 1/3$ do “birth” step and set $\omega_{23} = 1/3$; If $M = 2$ and $1/3 \leq U < 2/3$ do “death” step and set $\omega_{21} = 1/3$; if $M = 2$ and $U \geq 2/3$ do “MH” step.
- If $M = 3$ and $U < 0.5$ do “death” step and set $\omega_{32} = 0.5$; if $M = 3$ and $U \geq 0.5$ do “MH” step.

2.3.2 Metropolis-Hastings Step—In this step, we keep the dimensions of $\Omega_f^{(v)}$ and $\xi_3^{(v)}$ unchanged while the parameters are updated. Draw $U = \text{Uniform}(0, 1)$. If $U \leq 1/5$ make Move-1; if $1/5 < U \leq 2/5$, make Move-2; if $2/5 < U \leq 3/5$, make Move-3; if $3/5 < U \leq 4/5$, make Move-4, and if $4/5 < U \leq 1$ make Move-5. In what follows, we set $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0.1$.

- (a) (Move-1) Updating $\xi_1^{(v)}$ and $\xi_2^{(v)}$. Let U_1 and U_2 be independent $\text{Uniform}(-0.5, 0.5)$ random variables. Draw $\xi_1^* = \xi_1^{(v)} + 0.2U_1$ and $\xi_2^* = \xi_2^{(v)} + 0.5U_2$. Calculate the MH ratio

$$r = \frac{f(\Omega_f^{(v)}, \xi_3^{(v)}, \xi_2^*, \xi_1^*, \sigma_f^{2(v)} | Y)}{f(\Omega_f^{(v)}, \xi_3^{(v)}, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y)}.$$

Accept (ξ_1^*, ξ_2^*) with probability $\min(1, r)$.

- (b) (Move-2) Updating $\Omega_f^{(v)}$. Draw a vector $e = \text{Normal}(0, \tau_2^2)I$ and add e to a column of $\Omega_f^{(v)}$ selected at random. Denote the new Ω_f matrix by Ω_f^* . Calculate the MH ratio

$$r = \frac{f(\Omega_f^*, \xi_3^{(v)}, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y) I(\Omega_f^* \in \mathcal{A})}{f(\Omega_f^{(v)}, \xi_3^{(v)}, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y)},$$

where $\mathcal{A}$ denotes the set $(-\pi/2, \pi/2)^{\dim(\Omega_f^*)}$. Accept Ω_f^* with probability $\min(1, r)$.

- (c) (Move-3) Updating $\xi_3^{(v)}$. Let $U_3 = \text{Uniform}(-0.5, 0.5)$. Draw the j^{th} component of $\xi_3^*, \xi_{3,j}^* = \xi_{3,j}^{(v)} + 0.1U_3$. Calculate the MH ratio

$$r = \frac{f(\Omega_f^{(v)}, \xi_3^*, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y)}{f(\Omega_f^{(v)}, \xi_3^{(v)}, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y)}.$$

Accept ξ_3^* with probability $\min(1, r)$.

- (d) (Move-4) Updating $\sigma_f^{2(v)}$. This can be done by a Gibbs step, see the appendix.
- (e) (Move-5) Put the elements of $\xi_3^{(v)}$ in descending order and exchange columns of $\Omega_f^{(v)}$ and $\Theta_f^{(v)}$ accordingly. Because identifiability requires that the elements of $D_\alpha = \text{diag}\{\exp(\xi_3)\}$ be in descending order, this step avoids the label switching problem. Accept Ω_f^* and ξ_3^* with probability $\min(1, r)$.

2.3.3 Birth Step—In this step, we try to append a new column to $\Omega_f^{(v)}$ and append an element to $\xi_3^{(v)}$.

- (a) (Angle sampling) Draw a random vector γ from $\text{Uniform}(-\frac{\pi}{2}, \frac{\pi}{2})$, and draw $\xi^* = \text{Normal}(0, \tau_4^2)$. set $\Omega_f^* = (\Omega_f^{(v)}, \gamma)$ and $\xi_3^* = \text{diag}[\xi_3^{(v)}, \xi^*]$.
- (b) (Proposal acceptance) Calculate the MH ratio

$$r = \frac{f(\Omega_f^*, \xi_3^*, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y+1)}{f(\Omega_f^{(v)}, \xi_3^{(v)}, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y)} \frac{\tau_4}{\pi^{-(p-1)} \phi(\xi^* / \tau_4)} \frac{\omega_{k+1,k}}{\omega_{k,k+1}}. \quad (3)$$

- (c) Accept Ω_f^* and ξ_3^* with probability $\min(1, r)$.

2.3.4 Death step—In this step, we try to delete a random column of $\Omega_f^{(v)}$ and the corresponding diagonal element of $\xi_3^{(v)}$.

$$r = \frac{f(\Omega_f^*, \xi_3^*, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y - 1)}{f(\Omega_f^{(v)}, \xi_3^{(v)}, \xi_2^{(v)}, \xi_1^{(v)}, \sigma_f^{2(v)} | Y)} \frac{\pi^{-(p-1)} \phi(\xi^{(v)} / \tau_4)}{\tau_4} \frac{\omega_{k-1,k}}{\omega_{k,k-1}}, \quad (4)$$

where $\xi^{(v)}$ denotes the selected diagonal element of $\xi_3^{(v)}$.

The two groups of parameters, (Ω_f, ξ_3) and $(\xi_1, \xi_2, \sigma_f^2)$, are updated in an unbalanced way in the above algorithm. This is allowed by the Gibbs sampler.

2.4 SAMC

Because of low acceptance rates in the death step of the algorithm in Section 2.3, we opt to implement the Stochastic Approximation Monte Carlo or SAMC algorithm (Liang et al., 2007, Section 5; Liang, 2009), which is a method designed to provide samples from the entire space that are then weighted to form posterior distributions. The SAMC algorithm, especially appropriate for our problem, is given in equation (13) of Liang et al. (2007). We describe the algorithm for three principal components, detailed reasoning for this is provided later.

Let v be the index in the MCMC calculations. Let $\zeta = (1/3, 1/3, 1/3)$ and let v_0 be a tuning parameter that can be adjusted. Define $\kappa_v = v_0 / \max(v_0, v)$. Also define $\theta^{(v)} = (\theta_{v,1}, \theta_{v,2}, \theta_{v,3})^T$, where $\theta_{v,i}$ is the current estimate of the i^{th} normalizing constant of the target density, which in this case equals $\log(1/\zeta_i)$. Let $\mathcal{T}$ be the sampling space of allowable values of $\theta^{(v)}$. Liang et al. (2007) make this a very large space, but because of the exponentiation involved in the adjusted Metropolis ratio, we have restricted this to the cube $[-10, 10]^3$.

First start with $v = 1$ and $\theta^{(v)} = \{\log(3), \log(3), \log(3)\}^T$. Let $M^{(v)}$ be the current number of principal components, and let M^* be a proposal for the number of principal components. If $M^{(v)} = M^*$, the Metropolis steps in Section 2.3.2 do not change. If $M^{(v)} \neq M^*$, and if r is the birth or death ratio in (3) or (4), respectively, the only thing that changes is that r is multiplied by $\exp(\theta_{v,M^{(v)}} - \theta_{v,M^*})$. Let the new model be $M^{(v+1)}$. Set $\theta^* = \theta^{(v)} + \kappa_{v+1}(e_{M^{(v+1)}} - \zeta)$, where the three component vector, $e_{M^{(v+1)}}$, is all zeros except with a 1.0 in location $M^{(v+1)}$. Then if $\theta^* \in \mathcal{T}$, we set $\theta^{(v+1)} = \theta^*$. Otherwise, we set $\theta^{(v+1)} = \theta^* + c^*$, where c^* is chosen so that $\theta^* + c^* \in \mathcal{T}$.

Having run SAMC, we are now in a position to do posterior inference. Thus for example, assuming equal prior probabilities for the three possible models, the posterior probability that the number of principal components, $M = j$, is given as

$$\text{pr}(M=j|Y) = \sum_v \exp(\theta_{v,M^{(v)}}) I(M^{(v)}=j) / \sum_v \exp(\theta_{v,M^{(v)}}),$$

where T is the number of iterations used to do posterior inference. Posterior quantities can be calculated similarly. For example, given that we have j principal components, let $\sigma_{\mu}^{(v)}$ be the value of σ_{μ} at iteration v . Then

$$\begin{aligned} E(\sigma_{\mu}|M=j, Y) &= \sum_v^T \sigma_{\mu}^{(v)} \exp(\theta_{v,M^{(v)}}) I(M^{(v)}=j) / \sum_v^T \exp(\theta_{v,M^{(v)}}) I(M^{(v)}=j); \\ E(\sigma_{\mu}|Y) &= \sum_v^T \sigma_{\mu}^{(v)} \exp(\theta_{v,M^{(v)}}) / \sum_v^T \exp(\theta_{v,M^{(v)}}). \end{aligned} \quad (5)$$

From (5), it can be seen that our approach employs model averaging to do posterior inference.

We provide details only for the case of up to 3 principal components because, in biological data, 3 principal components usually suffice, and most algorithms become unstable for higher number of principal components, see the comments of Rice and Silverman (1991) and Hall, Müller & Wang (2006) in a different context. If more than 3 principal components are used in our simulations and data analyses, the 4th and higher diagonal elements of the variance matrix D_{α} , which are in decreasing order, become crowded around zero, making the corresponding elements of Θ_f essentially unidentified.

3. SIMULATION DETAILS AND RESULTS

3.1 Simulation From the Model

We simulated 200 data sets from model (2) with $n = 20$ subjects, $m_i \equiv m = 11$ equally-spaced observations per subject, and with $\Theta_f = p \times M$, where $p = 5$ and $M = 1$ and $m = 1$ and $M = 2$. When $M = 1$, $D_{\alpha} = 1.00$, while with $M = 2$, we generated two scenarios, $D_{\alpha} = \text{diag}(1.00, 0.30)$ and $D_{\alpha} = \text{diag}(1.50, 0.75)$. We set $\sigma_{\epsilon}^2 = 0.25$ and $\theta_{\mu} = (-4.09, 7.07, -7.82, 6.59, -2.10)^T$. The values of Ω_f were generated completely at random for each of the 200 data sets.

We ran SAMC with 200,000 Gibbs steps, and repeated the process 5 times with different starting values. In each case, we started by assuming 1 principal component, with Ω_f generated completely at random, and with starting values for $\xi_1 = \log(\sigma_{\epsilon}) =$

$\log\{\text{Uniform}(0.4, 0.6)\}$, $\xi_3 = \text{Uniform}(0.5, 1.5)$, $\xi_2 = \log(\sigma_{\mu}) = \text{Uniform}[4, 6]$ and $\sigma_f^2 = 2c$, where c yields 4.5 degrees of freedom for the smoother matrix trace $\{(B^T B + K/c)^{-1} B^T B\}$. We also repeated the analysis with no smoothing: as might be expected, there was slightly more variability in this latter case.

In all cases, σ_{ϵ}^2 and θ_{μ} were estimated with little bias. In addition, the within-subject function variances $\text{diag}\{\text{cov}(B_i \Theta_f \alpha_i)\}$ were estimated with only minor bias. An example of this is given in Figure 1, for the case that $D_{\alpha} = \text{diag}(1.50, 0.75)$. The top left panel shows the true function and the 5th and 95th pointwise percentiles of the simulations: there is almost no bias so that the mean of the simulations is not displayed. The bottom right panel shows that the within-subject function variances are also essentially unbiased. The top right panel gives the posterior probabilities of the models: in this case, 2 principal components were selected for all 200 data sets. The bottom left panel gives 16 examples of the within-subject function variances.

We also compared our analysis with the method of Zhou et al. (2008). Their informal method choose 2 principal components in 90.5% of the cases, and 1 principal component in 7% of the cases. In Table 1, we compare the mean absolute error and the mean squared error

for estimating the within-subject function variances for the method of Zhou et al. (2008) versus our method. SAMC had somewhat smaller mean squared errors for estimating the mean function, and much smaller mean squared errors for estimating the variance function.

3.2 Simulation Away From the Model

In this case, we simulated with $\sigma_\epsilon^2=0.25$, and in two settings with $D_\alpha = 1.00$ and $D_\alpha = \text{diag}(1.50, 0.75)$. We used the Bspline basis functions as in Zhou et al. (2008) with 7 basis functions. In the case of one principal component, the model is

$$Y_{i\ell} = \sin(2\pi t_\ell) + \sqrt{2}\cos(2\pi t_\ell)\alpha_i + \epsilon_{i\ell},$$

while with two principal components, and $\alpha_i = (\alpha_{i1}, \alpha_{i2})^T$,

$$Y_{i\ell} = \sin(2\pi t_\ell) + \sqrt{2}\cos(2\pi t_\ell)\alpha_{i1} + \left\{11.65(x - 0.5)^2 - \cos(2\pi t_\ell)/\pi^2\right\}\alpha_{i2} + \epsilon_{i\ell}.$$

We used 7 basis functions with 3 interior knots. All functions are adequately but not exactly approximated with this number of knots. The same priors were used as above, except that we centered them at the estimates from Zhou et al. (2008). We also experimented with fixed priors as in Section 3.2, with little change in results.

The results are again reported in Table 1, with graphical representation in the two principal components case given in Figure 2. In the former, we see that SAMC very similar to the method of Zhou et al. (2008) for estimating the mean function $\sin(2\pi t_j)$, and much better in terms of estimating the variance function. Figure 2 shows that there is more ambiguity in the posterior model probabilities across the simulated data sets, although in 97% of the cases two principal components have posterior probability > 0.50 . We actually think that this ambiguity is a strength of our approach. If the model is not correct, then there should be ambiguity in the choice of the number of principal components, with some random data sets coming close to being fit by 2 principal components, and others needing 3.

4. EMPIRICAL EXAMPLES

4.1 Outline

The simulations presented in Section 3 suggest that the method of Zhou et al. (2008) and our SAMC method will have roughly equal performance in terms of estimating the mean function, but will tend to have different performance when estimating the variance function. The SAMC method of course comes endowed with Bayesian posterior credible intervals. In this section, we will describe two data sets, each with two responses, and then display the analysis of all four. All analyses used 7 basis functions: the estimates of σ_ϵ^2 , $\sigma_\mu^2=1/\lambda_\mu$, $\sigma_f^2=1/\lambda_f$ and the first component of D_α from Zhou et al. (2008) were used as starting values. The starting value of Θ_f was random. SAMC was run with 200,000 iterations twice, and the results pooled.

4.2 RNA and CD4 Data

Our first example concerns data from an AIDS clinical trial. Both virologic and immunologic surrogate markers such as plasma HIV RNA copies (viral load) and CD4+ cell counts currently play important roles in evaluating antiviral therapies in AIDS clinical research. Both CD4+ cell counts and plasma HIV RNA copies have each been used as sole surrogate markers in AIDS clinical trials. Here we study the time course of plasma HIV

RNA copies and CD4+ cell counts obtained from an AIDS clinical study conducted by the AIDS Clinical Trials Group (ACTG 315). In this study, forty-six evaluable HIV-1 infected patients were treated with potent antiviral therapy consisting of zidovudine, zalcitabine and zalcitabine, see Lederman, Connick, Landay, Kuritzkes, Spritzler, St Clair, Kotzin, Fox, Chiozzi, Leonard, Rousseau, Wade, Roe, Martinez & Kessler (1998) and Wu & Ding (1999) for more details on this study. After initiation of potent antiviral treatment at day 0, patients were followed for up to 10 visits and at each visit both viral load and CD4+ cell counts were monitored simultaneously. The actual visit times are irregularly spaced and different for different patients. The visit time varies from day 0 (first visit) to day 196. We took logarithms for both CD4 data and RNA data and then standardized them to have mean zero and variance one. We also standardized follow up times to the unit interval. In this case, what we have been calling “time”, $t_{i\ell}$, is indeed time.

4.3 Phenotype Data

We apply our method to phenotypic longitudinal data of colonic crypts in the rat model. Colonic crypts are tube like structures in the colon formed by cells which undergo development from stem cells to fully differentiated mature colonic cells. This development begins at the bottom of the colonic crypt where the mother stem cells divide and produce new colon cells which will mature as they travel up the crypt to the colon surface. Colonic DNA damage can change the behavior of any cell by fundamentally altering its normal function and potentially inducing uncontrollable growth. This can contribute to carcinogenesis, see Hong, Bancroft, Turner, Davidson, Murphy, Carroll, Chapkin & Lupton (2005).

In this model, the times $t_{i\ell}$ correspond to the cell locations, normalized to the unit interval. The responses are the DNA adduct levels, a measure of DNA damage, in the proximal and in the distal region of the colon. In this analysis, there were 30 subjects. The DNA adduct levels in the proximal and distal regions were log-transformed, and then standardized to have mean zero and variance one.

In this case, what we call “time” is not time, but actually location of the cell. Since stem cells, or colloquially “baby” cells are at the bottom of the crypts, with t near zero, and since mature cells at the top of the crypts, t near one, the functions may be thought to be functions of the age of the cells.

4.4 Results

The method of Zhou et al. (2008) picked 1 principal component in all cases. The SAMC method had posterior probabilities of 1 principal component equal to 1.0 for the CD4 and RNA data. However, for the distal adducts, the posterior probability of two principal components was 0.92, while for proximal data 1 principal component has a posterior probability of 0.60. In all cases, the posterior probability of 3 principal components equalled 0.00.

As expected by the simulations, the estimated posterior mean function $b^T(t)\theta_\mu$ was almost identical to that found by Zhou et al. (2008) (not displayed). In Figure 3, we display the posterior mean and medians of $b^T(t)\theta_\mu$, along with 90% pointwise credible intervals. In the adduct level, the distribution of $t_{i\ell}$ was almost uniform, and hence the lengths of the credible intervals might be expected to be nearly constant, a conjecture seen in the plots. On the other hand, in the CD4 and RNA data, the distribution of $t_{i\ell}$ is highly skewed right, with most of the observations being between 0.0 and 0.4. In this case, one would expect the credible intervals to be much wider for larger values of $t_{i\ell}$ than for smaller values, an expectation born out in the plot.

Figure 4 displays results for estimating the marginal variances $b^T(t) \Theta_f D_a \Theta_f^T b(t)$ (magenta dot-dashed lines), the estimate from Zhou et al. (2008) (black solid lines) and posterior pointwise 90% credible intervals. Variance functions are of course much harder to estimate than mean functions, and as seen in the simulations here we do not expect SAMC and the method of Zhou et al. (2008) to be nearly identical, and they are not, although the differences are fairly small given the uncertainties.

5. DISCUSSION

We have shown how to use Stochastic Approximation Monte Carlo (SAMC) for Bayesian principal component selection and model estimation for longitudinal functional data. Our simulations suggest that our method is comparable to the method of Zhou et al. (2008) in terms of frequentist mean squared errors when estimating the mean function, and generally considerably more accurate when estimating the random variation about the mean function. While there are statistical gains in our approach, it is still MCMC based, and hence in terms of computing time generally much slower than the EM-based algorithm. Both methods scale up similarly when the number of subjects increases.

An interesting case to consider would be where the random functions $g_i(\cdot)$ in (1) or their spline equivalents $b^T(\cdot)\theta_i a_i$ in (2) are spatially correlated. Problems with such spatial correlation structure have been investigated by Baladandayuthapani, Hong, Mallick, Lupton, Turner & Carroll (2008) in a Bayesian manner and by Li, Wang, Hong, Turner, Lupton & Carroll (2007) in a frequentist manner. In unpublished work, we have developed a direct generalization of model (2) for this context, which again involves estimating the number of principal components. We believe that SAMC can be used successfully in this context.

Another interesting question is how to deal with correlation when, for example, there are pairs of responses such as might occur for familial data where different family members are followed in time (Sneddon & Sutradhar, 2004). Zhou, et al. (2008) discuss a generalization of model (2), this may be one way of attacking the problem from the principal components framework: they use an EM-algorithm. It would not be a great leap to generalize the SAMC algorithm to handle this case.

Acknowledgments

We wish to thank the editor and a referee for their detailed reading of the original manuscript and their many suggestions that led to a significant improvement in the paper. We also thank the participants of the International Symposium in Statistics on Inferences in Generalized Linear Longitudinal Mixed Models, whose comments led us to rethink important parts of this work.

Martinez was supported by a post-doctoral training grant from the National Cancer Institute (R25T-CA90301), Carroll's research was supported by a grant from the National Cancer Institute (CA57030), Liang's research was supported by a grant from the National Science Foundation (DMS-0706755) and Zhou was supported by a grant from the National Science Foundation (DMS-0907170). Both Carroll and Liang were also supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

APPENDIX

A.1 Marginal Posterior Density Computation

We provide some detail of integrations which lead to the marginal posterior density

$$f(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2 | \mathbf{Y}).$$

Recall that $\xi_1 = \log(\sigma_\epsilon)$, $\xi_2 = \log(\sigma_\mu)$ and $\xi_3 = \log\{\text{diag}(D_\alpha)\}$. Our prior distributions are in terms of (ξ_1, ξ_2, ξ_3) , but since they are not being integrated our displayed equations will use $(\sigma_\epsilon, \sigma_\mu, D_\alpha)$.

Our prior distributions were $\xi_1 = \text{Normal}(\mu_1, \sigma_1^2)$, $\xi_2 = \text{Normal}(\mu_2, \sigma_2^2)$, and

$\xi_3 = \text{Normal}(\mu_3, \sigma_f^2 I)$. It is more tricky to penalize Θ_f as done by Zhou et al. (2008), because

Θ_f must be orthogonal. We wrote the prior for $[\sigma_f^2 | \Omega_f]$ to be $\text{IG}\left(A + pM/2, B + \sum_{\ell=1}^M \vartheta_j^T K \vartheta_j\right)$, with $A = B = 3$, where ϑ_j is the j^{th} column of Θ_f . This has the convenient behavior that samples from this distribution are of the same form as samples for a smoothing parameter in an ordinary penalized spline regression without constraints, see Carroll, Ruppert, Tosteson, Crainiceanu & Karagas (2004). The prior distribution for Ω_f is independent uniforms on $[-\pi/2, \pi/2]$, and is written as $p(\Omega_f) = \pi^{-M(p-1)} I(-\pi/2 < \Omega_f < \pi/2)$.

Let the prior density for $(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2) = p(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2)$. Then except for a normalizing constant, the posterior density is

$$f(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2, \theta_\mu, \alpha | Y) = \prod_{i=1}^n \left\{ p(Y_i | \Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2, \theta_\mu, \alpha) p(\alpha_i | \xi_3) \right\} p(\theta_\mu) p(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2).$$

For convenience in displaying the equations, where it causes no confusion we will write this posterior in terms of $(\sigma_\epsilon, \sigma_\mu, D_\alpha)$, which is given as

$$\begin{aligned} & (\sigma_\epsilon^2)^{-N/2} \exp \left\{ - (2\sigma_\epsilon^2)^{-1} \sum_{i=1}^n (Y_i - B_i \theta_\mu - B_i \Theta_f \alpha_i)^T (Y_i - B_i \theta_\mu - B_i \Theta_f \alpha_i) \right\} \\ & \times |2\pi D_\alpha|^{-n/2} \exp \left\{ - (1/2) \sum_{i=1}^n \alpha_i^T D_\alpha^{-1} \alpha_i \right\} \times |\sigma_\mu^2 K^{-1}|^{-1/2} \exp \left\{ - (1/2) \theta_\mu^T (K/\sigma_\mu^2) \theta_\mu \right\} \\ & \times p(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2), \end{aligned}$$

where $N = \sum_{i=1}^n n_i$. Integrate with respect to $\alpha = (\alpha_1, \dots, \alpha_n)$

$$\begin{aligned} f(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2, \theta_\mu | Y) = & (\sigma_\epsilon^2)^{-N/2} \times |2\pi D_\alpha|^{-n/2} \times |\sigma_\mu^2 K^{-1}|^{-1/2} \times \left(\prod_{i=1}^n |C_{2i}|^{1/2} \right) \\ & \times \exp \left\{ - (2\sigma_\epsilon^2)^{-1} \sum_{i=1}^n R_i^T R_i - (1/2) \theta_\mu^T (K/\sigma_\mu^2) \theta_\mu \right. \\ & \left. + (1/2) \sum_{i=1}^n C_{1i}^T C_{2i} C_{1i} \right\} \times p(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2), \end{aligned}$$

where $C_{2i} = (\Theta_f^T B_i^T B_i \Theta_f / \sigma_\epsilon^2 + D_\alpha^{-1})^{-1}$, $C_{1i} = (R_i^T B_i \Theta_f / \sigma_\epsilon^2)$, and $R_i = Y_i - B_i \theta_\mu$. Now we integrate with respect to θ_μ and obtain

$$f(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2 | \mathbf{Y}) = (\sigma_\epsilon^2)^{-N/2} |2\pi D_\alpha|^{-n/2} |\sigma_\mu^2 K^{-1}|^{-1/2} \left(\prod_{i=1}^n |C_{2i}|^{1/2} \right) \times |C_4|^{1/2} \\ \exp \left\{ - (1/2) \sum_{i=1}^n Y_i^T (\sigma_\epsilon^{-2} \mathbf{I}_{n_i} - \Psi_i) Y_i + (1/2) C_3^T C_4 C_3 \right\} \\ \times p(\Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2), \quad (6)$$

$$\text{where } C_4 = \left(\sigma_\mu^{-2} K + \sigma_\epsilon^{-2} \sum_{i=1}^n B_i^T B_i - \sum_{i=1}^n B_i^T \Psi_i B_i \right)^{-1}, \quad C_3^T = \left(\sigma_\epsilon^{-2} \sum_{i=1}^n Y_i^T B_i - \sum_{i=1}^n Y_i^T \Psi_i B_i \right) \\ \text{and } \Psi_i = \sigma_\epsilon^{-4} B_i \Theta_f C_{2i} \Theta_f^T B_i^T.$$

A.2 Conditional Posterior Density of θ_μ and α_i

Easy calculations show that samples from θ_μ and α_i given the data and the parameters can be obtained as

$$\begin{aligned} \left[\theta_\mu | \Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2, \mathbf{Y} \right] &= \text{Normal}(C_4 C_3, C_4); \\ \left[\alpha_i | \Omega_f, \xi_3, \xi_2, \xi_1, \sigma_f^2, \theta_\mu, \mathbf{Y} \right] &= \text{Normal}(C_{2i} C_{1i}, C_{2i}). \end{aligned}$$

REFERENCES

- Baladandayuthapani V, Hong MY, Mallick BK, Lupton JR, Turner ND, Carroll RJ. Bayesian hierarchical spatially correlated functional data analysis with application to colon carcino-genesis. *Biometrics* 2008;64:64–73. [PubMed: 17608780]
- Besse P, Ramsay JO. Principal components-analysis of sampled functions. *Psychometrika* 1986;51:285–311.
- Boente G, Fraiman R. Kernel-based functional principal components. *Statistics & Probability Letters* 2000;48:335–345.
- Cardot H. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics* 2000;12:503–538.
- Carroll RJ, Ruppert D, Tosteson TD, Crainiceanu C, Karagas MR. Nonparametric regression and instrumental variables. *Journal of the American Statistical Association* 2004;99:736–750.
- Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995;82:711–732.
- Hall P, Hosseini-Nasab M. On properties of functional principal components analysis. *Journal of the Royal Statistical Society, Series B* 2006;68:109–126.
- Hall P, Müller HG, Wang JL. Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics* 2006;34:1493–1517.
- Huang JZ, Shen H, Buja A. Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics* 2008;2:678–695.
- Hong MY, Bancroft LK, Turner ND, Davidson LA, Murphy ME, Carroll RJ, Chapkin RS, Lupton JR. Fish oil decreases oxidative DNA damage by enhancing apoptosis in rat colon. *Nutrition and Cancer* 2005;52(2):166–175. [PubMed: 16201848]
- James GM, Hastie TJ, Sugar CA. Principal component models for sparse functional data. *Biometrika* 2000;87:587–602.
- Jank W, Shmueli G. Functional data analysis in electronic commerce research. *Statistical Science* 2006;21:155–166.
- Lederman MM, Connick E, Landay A, Kuritzkes DR, Spritzler J, Clair M, Kotzin BL, Fox L, Chiozzi MH, Leonard JM, Rousseau F, Wade M, Roe JD, Martinez A, Kessler H. Immunological responses associated with 12 weeks of combination antiretroviral therapy consisting of

- Zidovudine, Lamivudine & Ritonavir: Results of AIDS Clinical Trials Group Protocol 315. *J. Infectious Diseases* 1998;178:70–79. [PubMed: 9652425]
- Li Y, Wang N, Hong MY, Turner ND, Lupton JR, Carroll RJ. Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *Annals of Statistics* 2007;35:1608–1643.
- Liang F. On the use of stochastic approximation Monte Carlo for Monte Carlo integration. *Statistics and Probability Letters* 2009;79:581–587.
- Liang F, Liu C, Carroll RJ. Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association* 2007;102:305–320.
- Ocaña FA, Aguilera AM, Escabias M. Computational considerations in functional principal component analysis. *Computational Statistics* 2007;22:449–465.
- Sneddon G, Sutradhar BC. On semiparametric familial longitudinal models. *Statistics & Probability Letters*. 2004;69(3):369–379.
- Ramsay JO, Dalzell CJ. Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B* 1991;53:539–572. With discussion.
- Ramsay, JO.; Silverman, BW. *Functional Data Analysis*. 2nd Edition. Springer; New York: 2005.
- Reiss PT, Ogden T. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association* 2007;102:984–996.
- Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* 1991;53:233–243.
- Rice JA, Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* 2001;57:253–59. [PubMed: 11252607]
- Rice JA. Functional and longitudinal data analysis: Perspectives on smoothing. *Statistical Sinica* 2004;14:613–29.
- Silverman BW. Smoothed functional principal components analysis by choice of norm. *Annals of Statistics* 1996;24:1–24.
- Wu H, Ding A. Population HIV-1 dynamics in vivo: application models and inference tools for virological data from AIDS clinical trials. *Biometrics* 1999;55:410–18. [PubMed: 11318194]
- Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005a;100:577–90.
- Yao F, Müller HG, Wang JL. Functional linear regression analysis for longitudinal data. *Annals of Statistics* 2005b;33:2873–903.
- Zhang D, Davidian M. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* 2001;57:795–802. [PubMed: 11550930]
- Zhou L, Huang JZ, Carroll RJ. Joint modeling of paired sparse functional data using principal components. *Biometrika* 2008;95:601–619. [PubMed: 19396364]

Figure 1.

Results of the simulation study in the case of two principal components with $D\alpha = \text{diag}(1.50; 0.75)$, as described in Section 3.1. Top left: all 200 posterior means for $B_i\theta_u$. The true value is within the 200 lines. Top right: the estimated model probabilities over the 200 runs. Bottom left: For sixteen of the data sets, this gives the true (solid line) and estimated (dashed line) within-subject functional variance $\text{diag}\{\text{cov}(B_i\Theta_f\alpha_i)\}$. Bottom right: the true and mean of the estimated within-subject functional variances.

Figure 2.

Results of the simulation study in the case of two principal components with $D\alpha = \text{diag}(1.50, 0.75)$, as described in Section 3.2. Top left: the true mean function and the pointwise 5th and 95th percentiles of the estimated mean function. The true value is within the 200 lines. Top right: the estimated model probabilities over the 200 runs. Bottom left: For sixteen of the data sets, this gives the true (solid line) and estimated (dashed line) within-subject functional variance $\text{diag}\{\text{cov}(B_i\Theta_f\alpha_i)\}$. Bottom right: the true and mean of the estimated within-subject functional variance.

Figure 3. Results of the data analysis of Section 4. Displayed are the posterior mean function (dash-dotted magenta line), the posterior median (black solid line) and 90% pointwise posterior credible intervals (blue and red dashed lines). The labels for each graph refer to the data set.

Figure 4. Results of the data analysis of Section 4. Displayed are the posterior mean variance function (dash-dotted magenta line), the estimated variance function from the method of Zhou et al. (2008) and labeled “pdfa” (black solid line) and 90% pointwise posterior credible intervals (blue and red dashed lines) for SAMC. The labels for each graph refer to the data set.

Results of the simulation study for either 1 principal component, “1PC”, or 2 principal components, “2PC”. The simulations labeled “Model” are described in Section 3.1, while those labeled “Function” are described in Section 3.2. The mean absolute errors (MAE) and the mean squared error (MSE) for estimating the “Function” $E(Y|X) = B^T\theta_\mu$ and those for estimating the “Variance” $\text{diag}\{\text{cov}(B_i\theta_i\phi_i)\}$ are listed. For the method of Zhou et al. (2008), the “Model Probabilities” are the percentage of the times their method picks various numbers of principal components. For the SAMC methods, we list the average posterior model probabilities; the numbers in parentheses are the fraction of the time that the average model probability exceeded 50%.

Table 1

Simulation	Method	Function			Variance			Model Probabilities		
		MAE	MSE	MAE	MSE	MAE	MSE	1PC	2PC	3PC
Model, 1PC	Zhou	0.154	0.052	0.281	0.301	0.995	0.000	0.005		
	SAMC	0.141	0.044	0.222	0.189	0.999	0.001	0.000		
Model, 2PC $D_u = (1.00, 0.30)$	Zhou	0.192	0.073	0.383	0.480	0.055	0.915	0.030		
	SAMC	0.181	0.066	0.421	0.432	0.014	0.975	0.001		
Model, 2PC $D_u = (1.50, 0.75)$	Zhou	0.248	0.121	0.655	1.300	0.070	0.905	0.025		
	SAMC	0.230	0.106	0.461	0.575	0.000	0.972	0.018		
Function, 1PC	Zhou	0.117	0.026	0.323	0.236	1.000	0.000	0.000		
	SAMC	0.129	0.029	0.274	0.177	0.996	0.004	0.000		
Function, 2PC	Zhou	0.210	0.089	0.897	2.188	0.030	0.970	0.000		
	SAMC	0.216	0.092	0.728	1.336	0.091	0.892	0.007		