Biometrika (2008), **95**, 3, *pp.* 601–619
doi: 10.1093/biomet/asn035

Printed in Great Britain

Joint modelling of paired sparse functional data using principal components

BY LAN ZHOU, JIANHUA Z. HUANG AND RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
lzhou@stat.tamu.edu jianhua@stat.tamu.edu carroll@stat.tamu.edu

SUMMARY

We propose a modelling framework to study the relationship between two paired longitudinally observed variables. The data for each variable are viewed as smooth curves measured at discrete time-points plus random errors. While the curves for each variable are summarized using a few important principal components, the association of the two longitudinal variables is modelled through the association of the principal component scores. We use penalized splines to model the mean curves and the principal component curves, and cast the proposed model into a mixed-effects model framework for model fitting, prediction and inference. The proposed method can be applied in the difficult case in which the measurement times are irregular and sparse and may differ widely across individuals. Use of functional principal components enhances model interpretation and improves statistical and numerical stability of the parameter estimates.

Some key words: Functional data; Longitudinal data; Mixed-effects model; Penalized spline; Principal component; Reduced-rank model.

1. INTRODUCTION

The relationship between two paired longitudinal observed variables has been studied with regression models for longitudinal data (Liang & Zeger, 1986; Fahrmeir & Tutz, 1994; Moyeed & Diggle, 1994; Zeger & Diggle, 1994; Hoover et al., 1998; Wu et al., 1998; Huang et al., 2002). Also, Liang et al. (2003) modelled the paired longitudinal variables using a mixed-effects varying coefficient model with a measurement error in the covariates. Let X_{ij} and Y_{ij} denote longitudinal observations of a covariate and response for subject i at time occasion t_{ij}. The model of Liang et al. (2003) can be written as

$$Y_{ij} = \beta_0(t_{ij}) + \gamma_{0i}(t_{ij}) + X_{ij}\{\beta_1(t_{ij}) + \gamma_{1i}(t_{ij})\} + e_i(t_{ij}) \qquad (j = 1, \ldots, m_i, i = 1, \ldots, n),$$

where $\beta_0(t)$ and $\beta_1(t)$ are fixed functions, $\gamma_{0i}(t)$ and $\gamma_{1i}(t)$ are zero-mean subject-specific random functions and $e_i(t)$ are zero-mean error processes. In contrast to much existing work, the method effectively models the within-subject correlation in a flexible way by considering subject-specific regression coefficient functions.

However, the regression-based methods, including that of Liang et al. (2003), have several limitations. First, one needs to distinguish response and regressor variables, but sometimes such a distinction is not natural. Secondly, as in Liang et al. (2003), the regression-based methods usually focus on the contemporaneous relationship, that is, the relationship at the same time-point, between two variables. One could include lagged variables as regressors, but there are technical difficulties in the implementation of their method when the observation times for different variables differ, as often occurs in practice. Finally, it may be hard to interpret the results

from a contemporaneous regression model if we wish to consider all time-points from the past collectively. The usual interpretation of a regression slope as the average change in the response associated with a unit increase in the regressor is hardly satisfactory since the regressors from different time-points are correlated.

To overcome these shortcomings, we propose an alternative approach. The data for each variable are viewed as smooth curves sampled at discrete time-points plus random errors. The curves are decomposed as the sum of a mean curve and subject-specific deviations from the mean curve. The deviations are subsequently summarized by scores on a few important principal component curves extracted from the data. The association of the pair of curves is then modelled through the association of two low-dimensional vectors of principal component scores corresponding to the two underlying variables. By modelling the mean curves and the principal component curves as penalized splines, we cast our approach into a mixed-effects model framework for model fitting, prediction and inference.

Our method views longitudinal data as sparsely observed functional data (Rice, 2004). Ramsay & Silverman (2005) provide a comprehensive treatment of functional data analysis. The approach in this paper is most closely related to that of James et al. (2000) and Rice & Wu (2001). However, those papers considered models only for single curves, instead of paired curves as in this paper. Similarly to James et al. (2000), our approach is model-based, with the principal component curves being directly outputted from the fitted model. Yao et al. (2005a) proposed a different type of principal components analysis for sparse functional data through the eigen-decomposition of the covariance kernel estimated using two-dimensional smoothing. Yao et al. (2005b) dealt with the functional linear model for longitudinal data using regression through principal component scores. Another approach to modelling the association of paired curves is functional canonical correlation (Leurgans et al., 1993; He et al., 2003), but its adaptation to sparse functional data remains an open problem.

2. THE MIXED-EFFECTS MODEL FOR SINGLE CURVES

2·1. *The mixed-effects model*

Shi et al. (1996) and Rice & Wu (2001) suggest using a set of smooth basis functions $b_l(t)$ $(l = 1, \ldots, q)$, such as B-splines, to represent the curves, where the spline coefficients are assumed to be random to capture the individual- or curve-specific effects. Let $Y_i(t)$ be the value of the ith curve at time t and write

$$Y_i(t) = \mu(t) + h_i(t) + \epsilon_i(t), \tag{1}$$

where $\mu(t)$ is the mean curve, $h_i(t)$ represents the departure from the mean curve for subject i and $\epsilon_i(t)$ is random noise with mean zero and variance σ^2. Let $b(t) = \{b_1(t), \ldots, b_q(t)\}^{\mathrm{T}}$ be the vector of basis functions evaluated at time t. Denote by β an unknown but fixed vector of spline coefficients, and let γ_i be a random vector of spline coefficients for each curve with covariance matrix Γ. When $\mu(t)$ and $h_i(t)$ are modelled with a linear combination of B-splines, equation (1) has the mixed-effects model form

$$Y_i(t) = b(t)^{\mathrm{T}}\beta + b(t)^{\mathrm{T}}\gamma_i + \epsilon_i(t). \tag{2}$$

In practice, $Y_i(t)$ is observed only at a finite set of time-points. Let Y_i be the vector consisting of the n_i observed values, let B_i be the corresponding $n_i \times q$ spline basis matrix evaluated at these time-points and let ϵ_i be the corresponding random noise vector with covariance matrix $\sigma^2 I$. The

mixed-effects model for the observed data is

$$Y_i = B_i\beta + B_i\gamma_i + \epsilon_i. \tag{3}$$

The EM algorithm can be used to calculate the maximum likelihood estimates $\hat{\beta}$ and $\hat{\Gamma}$ (Laird & Ware, 1982). Given these estimates, the best linear unbiased predictors of the random effects γ_i are

$$\hat{\gamma}_i = \left(\hat{\sigma}^2\hat{\Gamma}^{-1} + B_i^{\mathrm{T}}B_i\right)^{-1}B_i^{\mathrm{T}}(Y_i - B_i\hat{\beta}).$$

The mean curve $\mu(t)$ can then be estimated by $\hat{\mu}(t) = b(t)^{\mathrm{T}}\hat{\beta}$ and the subject-specific curves $h_i(t)$ can be predicted as $\hat{h}_i(t) = b(t)^{\mathrm{T}}\hat{\gamma}_i$.

2·2. *The reduced-rank model*

Since Γ involves $q(q+1)/2$ different parameters, its estimator based on a sparse dataset can be highly variable, and the large number of parameters may also make the EM algorithm fail to converge to the global maximum. James et al. (2000) pointed out these problems with the mixed-effects model and instead proposed a reduced-rank model, in which the individual departure from the mean is modelled by a small number of principal component curves. The reduced-rank model is

$$Y_i(t) = \mu(t) + \sum_{j=1}^{k} f_j(t)\alpha_{ij} + \epsilon_i(t) = \mu(t) + f(t)^{\mathrm{T}}\alpha_i + \epsilon_i(t), \tag{4}$$

where $\mu(t)$ is the overall mean, f_j is the jth principal component function or curve, $f = (f_1, \ldots, f_k)^{\mathrm{T}}$ and $\epsilon_i(t)$ is the random error. The principal components are subject to the orthogonality constraint $\int f_j f_l = \delta_{jl}$, with δ_{jl} being the δ function. The components of the random vector α_i give the relative weights of the principal component functions for the ith individual and are called principal component scores. The α_is and ϵ_is are independent and are assumed to have mean zero. The α_is are taken to have a common covariance matrix and the ϵ_is are assumed temporally uncorrelated with a constant variance of σ^2.

Similarly to the mixed-effects model (2), we represent μ and f using B-splines. Let $b(t) = \{b_1(t), \ldots, b_q(t)\}^{\mathrm{T}}$ be a spline basis with dimension q. Let θ_μ and Θ_f be, respectively, a q-dimensional vector and a $q \times k$ matrix of spline coefficients. Write $\mu(t) = b(t)^{\mathrm{T}}\theta_\mu$ and $f(t)^{\mathrm{T}} = b(t)^{\mathrm{T}}\Theta_f$. The reduced-rank model then takes the form

$$Y_i(t) = b(t)^{\mathrm{T}}\theta_\mu + b(t)^{\mathrm{T}}\Theta_f\alpha_i + \epsilon_i(t), \qquad \epsilon_i(t) \sim (0, \sigma_\epsilon^2), \qquad \alpha_i \sim (0, D_\alpha), \tag{5}$$

where D_α is diagonal, subject to

$$\Theta_f^{\mathrm{T}}\Theta_f = I, \quad \int b(t)b(t)^{\mathrm{T}}\,dt = I. \tag{6}$$

The equations in (6) imply that

$$\int f(t)f(t)^{\mathrm{T}}\,dt = \Theta_f^{\mathrm{T}}\int b(t)b(t)^{\mathrm{T}}\,dt\,\Theta_f = I,$$

which are the usual orthogonality constraints on the principal component curves.

The requirement that the covariance matrix D_α of α_i is diagonal is for identifiability purposes. Without imposing (6), neither Θ_f nor D_α can be identified: only the covariance matrix of $\Theta_f\alpha_i$, namely $\Theta_f D_\alpha \Theta_f^{\mathrm{T}}$, can be identified. To identify Θ_f and D_α, note that $\Theta_f\alpha_i = \tilde{\Theta}_f\tilde{\alpha}_i$, where $\tilde{\Theta}_f = \Theta_f C$ and $\tilde{\alpha}_i = C^{-1}\alpha_i$ for any invertible $k \times k$ matrix C. Therefore, by requiring that D_α be diagonal and that the Θ_f have orthonormal columns, we prevent reparameterization by linear

transformation. The identifiability condition is more precisely given in the following lemma, which follows from the uniqueness of the eigen-decomposition of a covariance matrix.

LEMMA 1. *Assume that* $\Theta_f^{\mathrm{T}}\Theta_f = I$ *and that the first nonzero element of each column of* Θ_f *is positive. Let* α_i *be ordered according to their variances in decreasing order. Suppose the elements of* α_i *have different variances, that is,* $\mathrm{var}(\alpha_{i1}) > \cdots > \mathrm{var}(\alpha_{ik})$. *Then the model specified by equations* (5) *and* (6) *is identifiable.*

In Lemma 1, the first nonzero element of each column of Θ_f is used to determine the sign at the population level. With finite samples, it is best to use the element of the largest magnitude in each column of Θ_f to determine the sign, since this choice is least influenced by finite-sample random fluctuation.

The observed data usually consist of $Y_i(t)$ sampled at a finite number of observation times. For each individual i, let $t_{i1}, \ldots, t_{in_i}$ be the different time-points at which measures are available. Write

$$Y_i = \{Y_i(t_{i1}), \ldots, Y_i(t_{in_i})\}^{\mathrm{T}}, \quad B_i = \{b(t_{i1}), \ldots, b(t_{in_i})\}^{\mathrm{T}}.$$

The reduced-rank model can then be written as

$$Y_i = B_i\theta_\mu + B_i\Theta_f\alpha_i + \epsilon_i, \qquad \Theta_f^{\mathrm{T}}\Theta_f = I, \qquad \epsilon_i \sim (0, \sigma^2 I), \qquad \alpha_i \sim (0, D_\alpha). \tag{7}$$

The orthogonality constraints imposed on $b(t)$ are achieved approximately by choosing $b(t)$ such that $(L/g)B^{\mathrm{T}}B = I$, where $B = \{b(t_1), \ldots, b(t_g)\}^{\mathrm{T}}$ is the basis matrix evaluated on a fine grid of time-points $t_1, \ldots, t_g$ and L is the length of the interval in which we take these grid points; see Appendix 1 for details of implementation. Since (7) is also a mixed-effects model, an EM algorithm can be used to estimate the parameters. By focusing on a small number of leading principal components, the reduced-rank model (7) employs a much smaller set of parameters than the original model (3), and thus more reliable parameter estimates can be obtained.

2·3. *The penalized spline reduced-rank model*

The reduced-rank model of James et al. (2000) uses fixed-knot splines. For many applications, especially when the sample size is small, only a small number of knots can be used in order to fit the model to the data. An alternative, more flexible approach is to use a moderate number of knots and apply a roughness penalty to regularize the fitted curves (Eilers & Marx, 1996; Ruppert et al., 2003).

For the reduced-rank model (4)–(7), we can use a moderate q, in the range of 10–20, say, and employ the method of penalized likelihood, with roughness penalties that force the fitted functions $\mu(t)$ and $f_1(t), \ldots, f_k(t)$ to be smooth. We focus on roughness penalties of the form of integrated squared second derivatives, though other forms are also applicable. One approach is to use the penalty

$$\lambda_\mu \int \{\mu''(t)\}^2\, dt + \lambda_{f1} \int \{f_1''(t)\}^2\, dt + \cdots + \lambda_{fk} \int \{f_k''(t)\}^2\, dt, \tag{8}$$

where $\lambda_\mu, \lambda_{f1}, \ldots, \lambda_{fk}$ are tuning parameters. However, for simplicity, we shall take $\lambda_{f1} = \cdots = \lambda_{fk} = \lambda_f$. In terms of model (7), this simplified penalty can be written as

$$\lambda_\mu{}^{\mathrm{T}}\theta_\mu \int b''(t)b''(t)^{\mathrm{T}}\theta_\mu\, dt + \lambda_f \sum_{j=1}^{d} \theta_{fj}^{\mathrm{T}} \int b''(t)b''(t)^{\mathrm{T}}\, dt\, \theta_{fj}, \tag{9}$$

where θ_{fj} is the jth column of Θ_f.

Assume that the α_is and ϵ_is are normally distributed. Then,

$$Y_i \sim N(B_i\theta_\mu, \sigma^2 I + B_i\Theta_f D_\alpha \Theta_f^{\mathrm{T}} B_i^{\mathrm{T}}) \qquad (i = 1, \ldots, n),$$

and minus twice the loglikelihood based on the Y_is, with an irrelevant constant omitted, is

$$\sum_{i=1}^{n} \log|\sigma^2 I + B_i\Theta_f D_\alpha \Theta_f^{\mathrm{T}} B_i^{\mathrm{T}}| + (Y_i - B_i\theta_\mu)^{\mathrm{T}}(\sigma^2 I + B_i\Theta_f D_\alpha \Theta_f^{\mathrm{T}} B_i^{\mathrm{T}})^{-1}(Y_i - B_i\theta_\mu).$$

The method of penalized likelihood minimizes the sum of the above expression and the penalty in (9). While direct optimization is complicated, it is easier to treat the α_is as missing data and employ the EM algorithm. A modification of the algorithm by James et al. (2000) that takes into account the roughness penalty can be applied. The details are not presented here. The algorithm can also be obtained easily as a simplification of our algorithm for joint modelling of paired curves to be given in § 3.

3. The mixed-effects model for paired curves

For data consisting of paired curves, an important problem of interest is modelling the association of the two curves. We first model each curve using the reduced-rank principal components model as discussed in § 2·2, and then model the association of curves by jointly modelling the principal component scores. Roughness penalties are introduced as in § 2·3 to obtain smooth fits of the mean curve and principal components.

Let $Y_i(t)$ and $Z_i(t)$ denote the two measurements at time t for the ith individual. The reduced-rank model has the form

$$Y_i(t) = \mu(t) + \sum_{j=1}^{k_\alpha} f_j(t)\alpha_{ij} + \epsilon_i(t) = \mu(t) + f(t)^{\mathrm{T}}\alpha_i + \epsilon_i(t),$$

$$Z_i(t) = \nu(t) + \sum_{j=1}^{k_\beta} g_j(t)\beta_{ij} + \xi_i(t) = \nu(t) + g(t)^{\mathrm{T}}\beta_i + \xi_i(t),$$

where $\mu(t)$ and $\nu(t)$ are the mean curves, $f = (f_1, \ldots, f_{k_\alpha})^{\mathrm{T}}$ and $g = (g_1, \ldots, g_{k_\beta})^{\mathrm{T}}$ are vectors of principal components, $\epsilon_i(t)$ and $\xi_i(t)$ are measurement errors. The α_is, β_is, ϵ_is and ξ_is are assumed to have mean zero. The measurement errors $\epsilon_i(t)$ and $\xi_i(t)$ are assumed to be uncorrelated with constant variances σ_ϵ^2 and σ_ξ^2, respectively. It is also assumed that the α_is, ϵ_is and ξ_is are mutually independent, as are the β_is, ϵ_is and ξ_is. The principal components are subject to the orthogonality constraints $\int f_j f_l = \delta_{jl}$ and $\int g_j g_l = \delta_{jl}$, with δ_{kl} being the Kronecker delta.

For identifiability, the principal component scores α_{ij} $(j = 1, \ldots, k_\alpha)$, are independent with strictly decreasing variances; see Lemma 1. Similarly, the principal component scores β_{ij} $(j = 1, \ldots, k_\beta)$, are also independent with strictly decreasing variances. Denote the diagonal covariance matrices of α_i and β_i by D_α and D_β, respectively.

The relationship between $Y_i(t)$ and $Z_i(t)$ is assumed through the correlation between the principal component scores α_i and β_i. To be specific, we assume that $\mathrm{cov}(\alpha_i, \beta_i) = C$. Then, α_i and β_i are modelled jointly as follows:

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_\alpha & C \\ C^{\mathrm{T}} & D_\beta \end{pmatrix} \right\}.$$

This is equivalent to the regression model

$$\beta_i = \Lambda\alpha_i + \eta_i, \tag{10}$$

where $\Lambda = C^{\mathrm{T}}D_\alpha^{-1}$ or $C = D_\alpha\Lambda^{\mathrm{T}}$, from which it follows that the covariance matrix of η_i is $\Sigma_\eta = D_\beta - \Lambda D_\alpha \Lambda^{\mathrm{T}}$. We find this regression formulation to be more convenient when calculating the likelihood function.

The roles of $Y(t)$ and $Z(t)$ and therefore the roles of α_i and β_i are symmetric in our modelling framework. In the regression formulation (10), however, α_i and β_i do not appear to play symmetric roles, and the interpretation of Λ depends on what is used as the regressor and what is used as the response. However, this formulation only serves as a computational device. If we switch the roles of α_i and β_i, we still obtain the same estimates of the original parameters (D_α, D_β, C).

Let $R = D_\alpha^{-1/2} C D_\beta^{-1/2}$ be the matrix of correlation coefficients, which provides a scale-free measure of the association between α_i and β_i. We call diagonal entries of D_α and D_β, together with σ_ϵ^2 and σ_ξ^2, the variance parameters and we refer to the off-diagonal entries of R as the correlation parameters.

We represent μ, ν, f and g as a member of the same space of spline functions with dimension q. The basis of the spline space, denoted by $b(t)$, is chosen to be orthonormal, that is, the components of $b(t) = \{b_1(t), \ldots, b_q(t)\}^{\mathrm{T}}$ satisfy $\int b_j(t) b_l(t)\, dt = \delta_{jl}$. Let θ_μ and θ_ν be q-dimensional vectors of spline coefficients such that

$$\mu(t) = b(t)^{\mathrm{T}}\theta_\mu, \qquad \nu(t) = b(t)^{\mathrm{T}}\theta_\nu. \tag{11}$$

Let Θ_f and Θ_g be, respectively, $q \times k_\alpha$ and $q \times k_\beta$ matrices of spline coefficients such that

$$f(t)^{\mathrm{T}} = b(t)^{\mathrm{T}}\Theta_f, \qquad g(t)^{\mathrm{T}} = b(t)^{\mathrm{T}}\Theta_g. \tag{12}$$

For each individual i, the two variables may have different observation times. However, for simplicity in presentation, we assume that there is a common set of observation times, $t_{i1}, \ldots, t_{in_i}$. Write $Y_i = \{Y_i(t_{i1}), \ldots, Y_i(t_{in_i})\}^{\mathrm{T}}$ and similarly for Z_i. Let $B_i = \{b(t_{i1}), \ldots, b(t_{in_i})\}^{\mathrm{T}}$. The model for the observed data can be written as

$$\begin{aligned} Y_i &= B_i\theta_\mu + B_i\Theta_f\alpha_i + \epsilon_i, \\ Z_i &= B_i\theta_\nu + B_i\Theta_g\beta_i + \xi_i, \\ \beta_i &= \Lambda\alpha_i + \eta_i, \\ \epsilon_i \sim (0, \sigma_\epsilon^2 I_{n_i}), \quad \xi_i &\sim (0, \sigma_\xi^2 I_{n_i}), \quad \alpha_i \sim (0, D_\alpha), \quad \beta_i \sim (0, D_\beta). \end{aligned} \tag{13}$$

To make this model identifiable, we require that $\Theta_f^{\mathrm{T}}\Theta_f = I$ and $\Theta_g^{\mathrm{T}}\Theta_g = I$, and that the first nonzero element of each column of Θ_f and Θ_g be positive. In addition, the elements of α_i and β_i are ordered according to their variances in decreasing order.

Parameter estimation using the penalized normal likelihood is discussed in detail in § 4. Given the estimated parameters, the mean curves of Y and Z and the principal component curves are estimated by plugging relevant parameter estimates into (11) and (12). Predictions of the principal component scores α_i and β_i are obtained using the best linear unbiased predictors,

$$\hat{\alpha}_i = E(\alpha_i \mid Y_i, Z_i, \Xi), \qquad \hat{\beta}_i = E(\beta_i \mid Y_i, Z_i, \Xi),$$

where Ξ denotes collectively all the estimated parameters, and the conditional means can be calculated using the formulae given in Appendix 2. The predictors of the α_is and β_is, combined with the estimates of $\mu(t)$, $\nu(t)$, $f(t)$ and $g(t)$, give predictors of the individual curves.

4. Fitting the bivariate reduced-rank model

4·1. *Penalized likelihood*

If we assume normality, the joint distribution of Y_i and Z_i is determined by the mean vector and variance-covariance matrix, which are given by

$$E(Y_i) = B_i\theta_\mu, \qquad E(Z_i) = B_i\theta_\nu,$$
$$\mathrm{var}(Y_i) = B_i\Theta_f D_\alpha \Theta_f^{\mathrm{T}} B_i^{\mathrm{T}} + \sigma_\epsilon^2 I_{n_i}, \qquad \mathrm{var}(Z_i) = B_i\Theta_g D_\beta \Theta_g^{\mathrm{T}} B_i^{\mathrm{T}} + \sigma_\xi^2 I_{n_i},$$
$$\mathrm{cov}(Y_i, Z_i) = B_i\Theta_f D_\alpha \Lambda^{\mathrm{T}} \Theta_g^{\mathrm{T}} B_i^{\mathrm{T}}.$$

Let $L(Y_i, Z_i)$ denote the contribution to the likelihood from subject i. The joint likelihood for the whole dataset is $\prod_{i=1}^n L(Y_i, Z_i)$. The method of penalized likelihood minimizes the criterion

$$\begin{aligned} &-2\sum_{i=1}^n \log L(Y_i, Z_i) \\ &+\lambda_\mu \theta_\mu^{\mathrm{T}} \int b''(t)b''(t)^{\mathrm{T}}\,dt\,\theta_\mu + \lambda_f \sum_{j=1}^{k_\alpha} \theta_{fj}^{\mathrm{T}} \int b''(t)b''(t)^{\mathrm{T}}\,dt\,\theta_{fj} \\ &+\lambda_\nu \theta_\nu^{\mathrm{T}} \int b''(t)b''(t)^{\mathrm{T}}\,dt\,\theta_\nu + \lambda_g \sum_{j=1}^{k_\beta} \theta_{gj}^{\mathrm{T}} \int b''(t)b''(t)^{\mathrm{T}}\,dt\,\theta_{gj}, \end{aligned} \tag{14}$$

where $d_{\alpha j}$ and $d_{\beta j}$ are, respectively, the jth diagonal elements of D_α and D_β, while θ_{fj} and θ_{gj} are, respectively, the jth columns of Θ_f and Θ_g. There are four regularization parameters, and this gives the flexibility of allowing different amounts of smoothing for the mean curves and principal components.

Direct minimization of (14) is complicated. If the α_is and β_is were observable, then the joint likelihood for $(Y_i, Z_i, \alpha_i, \beta_i)$ could be factorized as

$$L(Y_i, Z_i, \alpha_i, \beta_i) = f(Y_i \mid \alpha_i) f(Z_i \mid \beta_i) f(\beta_i \mid \alpha_i) f(\alpha_i).$$

With an irrelevant constant ignored, it follows that

$$\begin{aligned} &-2\log L(Y_i, Z_i, \alpha_i, \beta_i) \\ &\quad = n_i \log(\sigma_\epsilon^2) + \frac{1}{\sigma_\epsilon^2}(Y_i - B_i\theta_\mu - B_i\Theta_f\alpha_i)^{\mathrm{T}}(Y_i - B_i\theta_\mu - B_i\Theta_f\alpha_i) \\ &\qquad + n_i \log(\sigma_\xi^2) + \frac{1}{\sigma_\xi^2}(Z_i - B_i\theta_\nu - B_i\Theta_g\beta_i)^{\mathrm{T}}(Z_i - B_i\theta_\nu - B_i\Theta_g\beta_i) \\ &\qquad + \log(|\Sigma_\eta|) + (\beta_i - \Lambda\alpha_i)^{\mathrm{T}}\Sigma_\eta^{-1}(\beta_i - \Lambda\alpha_i) + \log(|D_\alpha|) + \alpha_i^{\mathrm{T}} D_\alpha^{-1}\alpha_i. \end{aligned} \tag{15}$$

Clearly, the unknown parameters are separated in the loglikelihood and therefore separate optimization is feasible. We thus treat α_i and β_i as missing values and use the EM algorithm (Dempster et al., 1977) to estimate the parameters.

4·2. *Conditional distributions*

The E-step of the EM algorithm consists of finding the prediction of the random effects α_i and β_i and their moments based on (Y_i, Z_i) and the current parameter values. In this section, all calculation is done given the current parameter values, although the dependence is suppressed in

the notation throughout. The conditional distribution of (α_i, β_i) given (Y_i, Z_i) is normal,

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_{i,\alpha} \\ \mu_{i,\beta} \end{pmatrix}, \Sigma_i = \begin{pmatrix} \Sigma_{i,\alpha\alpha}, & \Sigma_{i,\alpha\beta} \\ \Sigma_{i,\beta\alpha}, & \Sigma_{i,\beta\beta} \end{pmatrix} \right\}. \tag{16}$$

The predictions required by the EM algorithm are

$$\begin{gathered} \hat{\alpha}_i = E(\alpha_i \mid Y_i, Z_i) = \mu_{i,\alpha}, \qquad \hat{\beta}_i = E(\beta_i \mid Y_i, Z_i) = \mu_{i,\beta}, \\ M_{i,\alpha\alpha} = E\left(\alpha_i \alpha_i^{\mathrm{T}} \mid Y_i, Z_i\right) = \hat{\alpha}_i \hat{\alpha}_i^{\mathrm{T}} + \Sigma_{i,\alpha\alpha}, \qquad M_{i,\beta\beta} = E\left(\beta_i \beta_i^{\mathrm{T}} \mid Y_i, Z_i\right) = \hat{\beta}_i \hat{\beta}_i^{\mathrm{T}} + \Sigma_{i,\beta\beta}, \\ M_{i,\alpha\beta} = E\left(\alpha_i \beta_i^{\mathrm{T}} \mid Y_i, Z_i\right) = \hat{\alpha}_i \hat{\beta}_i^{\mathrm{T}} + \Sigma_{i,\alpha\beta}. \end{gathered} \tag{17}$$

Calculation of the conditional moments of the multivariate normal distribution (16) is given in Appendix 2.

4·3. *Optimization*

The M-step of the EM algorithm updates the parameter estimates by minimizing

$$\begin{aligned} &-2E\{\log L(Y_i, Z_i, \alpha_i, \beta_i \mid Y_i, Z_i)\} \\ &\quad + \lambda_\mu \theta_\mu{}^{\mathrm{T}} \int b''(t) b''(t)^{\mathrm{T}} \, dt \, \theta_\mu + \lambda_f \sum_{j=1}^{k} \theta_{fj}^{\mathrm{T}} \int b''(t) b''(t)^{\mathrm{T}} \, dt \, \theta_{fj} \\ &\quad + \lambda_\nu \theta_\nu{}^{\mathrm{T}} \int b''(t) b''(t)^{\mathrm{T}} \, dt \, \theta_\nu + \lambda_g \sum_{j=1}^{k} \theta_{gj}^{\mathrm{T}} \int b''(t) b''(t)^{\mathrm{T}} \, dt \, \theta_{gj}, \end{aligned}$$

or by reducing the value of this objective function as an application of the generalized EM algorithm. Since the parameters are well separated in the expression for the conditional loglikelihood, see (15), we can update the parameter estimates sequentially given their current values. We first update σ_ϵ^2 and σ_ξ^2, then θ_μ and θ_ν, and finally D_α, D_β and Λ. Details of the updating formulae are given in Appendix 3. In the last step, some care is needed to enforce the orthonormality constraints on the principal components.

5. MODEL SELECTION AND INFERENCE

5·1. *Specification of splines and penalty parameters*

Given the nature of sparse functional data and the usual low signal-to-noise ratio typical in such datasets, we expect that only the major smooth features in the data can be extracted by statistical methods. Placement of knot positions is therefore not critical for our method, and reasonable ways of doing this include spacing the knots equally over the data range or using sample quantiles of observation times. In our analysis of the AIDS data in § 7, for example, the knots were placed at the common scheduled visit times. Neither is the choice of the number of knots critical, as long as it is moderately large, since the smoothness of the fitted curves is mainly controlled by the roughness penalty. For typical sparse functional datasets, 10–20 knots is often sufficient.

To choose penalty parameters, a subjective choice is often satisfactory. A natural approach for automatic choice of penalty parameters is to maximize the crossvalidated loglikelihood. All examples in this paper use ten-fold crossvalidation. The criterion used for model selection is the sum of the ten calculated testset loglikelihoods.

There are four penalty parameters, so we need to search over a four-dimensional space for a good choice of these parameters. Although the simplex method of Nelder and Mead (1965) could

be used, a crude grid search worked well for all examples we considered. With five grid-points on each dimension, there are in total 625 possible combinations for four parameters. Implemented in Fortran, this strategy is computationally feasible and has been used for the data example in § 7. One possible simplification is to let $\lambda_\mu = \lambda_\nu$ and $\lambda_f = \lambda_g$ and thus reduce the dimension to two. This simplification, with five grid-points for each of the two dimensions, has been used for our simulation study in § 6.

5·2. *Selection of the number of significant principal components*

It is important to identify the number of important principal components in functional principal component analysis. For the single-curve model, choosing to fit too many principal components can degrade the fit of them all (James et al., 2000). Fitting too many principal components in the joint modelling is even more harmful, since instability can result if we try to estimate correlation coefficients among a set of latent random variables with big differences in variances.

In our method, we first apply the penalized spline reduced-rank model in § 2·3 to each variable separately and use these single-curve models to select the number of significant principal components for each variable. We then fit the joint model using the chosen numbers of significant principal components from fitting single-curve models; the numbers are refined if necessary. For the single-curve models, we use a stepwise addition approach, starting with one principal component and then adding one principal component at a time to the model. The process stops if the variances of the scores of the principal components already in the model do not change much after the addition of one more principal component, and the variance of the scores of the newly added principal component is much smaller than variances for those already in the model.

A more detailed description of the procedure is as follows. Let k_a and k_b denote the number of important principal components used in a single-curve model for Y and Z, respectively. Let $D_{\alpha,l}^{(k)}$ $(l = 1, \ldots, k)$, denote the variances of the principal component scores for an order-k model for Y. Similarly define $D_{\beta,l}^{(k)}$s for Z. To choose k_a we start with $k = 1$ and increase k by 1 at a time until we decide to stop according to the criterion described now. For each k, we fit an order-k and an order-$(k+1)$ single-curve model for Y. If $D_{\alpha,l}^{(k+1)} \doteq D_{\alpha,l}^{(k)}$ for all $l = 1, \ldots, k$ and $D_{\alpha,k+1}^{(k+1)} < cD_{\alpha,k}^{(k+1)}$ for some prespecified small constant c, we stop at that k and set $k_a = k$.

We select k_b similarly. We have used c in the range $1/25$ to $1/9$ in the above procedure. The joint model is then fitted with the selected k_a and k_b. The variances of the principal component scores from fitting the joint model need not be the same as those from the single-curve models. A refinement using the joint model takes the form of a stepwise deletion procedure. If the variance of the scores of the last principal component is much smaller than the variance of the scores of the previous principal component, delete that principal component from the model. This can be done sequentially if necessary.

We tested this procedure on the simulated datasets from § 6 where, in the true model, the variable Y has one significant principal component and the variable Z has two significant principal components. The results of applying the procedure without the second-stage refinement are as follows. When $c = 1/25$ was used, among 200 simulation runs, for variable Y, 97% picked one important principal component and 3% picked two important principal components; for variable Z, 98% picked two important principal components and 2% picked three important principal components. When $c = 1/9$ was used, in all simulations, one principal component was picked for variable Y; for variable Z, in 99% of simulations, two principal components were picked, and in 1% of simulations, one principal component was picked. The first-stage stepwise addition process has thus already provided quite an accurate choice of the number of important principal

components, and the second-stage stepwise deletion refinement is not necessary for this example. The use of the second-stage refinement will be illustrated using the data analysis in § 7.

5·3. *Confidence intervals*

The bootstrap can be applied to produce pointwise confidence intervals of the overall mean functions for both variables and the principal components curves, and of the variance and correlation coefficient parameters. The confidence intervals are based on appropriate sample quantiles of relevant estimates from the bootstrap samples. Here the bootstrap samples are obtained by resampling the subjects, in order to preserve the correlation of observations within subject. When applying the penalized likelihood to the bootstrap samples, the same specification of splines and penalty parameters may be used.

6. SIMULATION

In this section, we illustrate the performance of penalized likelihood in fitting the bivariate reduced-rank model. In each simulation run, we have $n = 50$ subjects and each subject has up to four visits between times 0 and 100. We generated the visit times by mimicking a typical clinical setting. The visit times for each subject were generated sequentially with the spacings between the visits normally distributed. In the actual generating procedure, each subject has a baseline visit, so that $t_{i1} = 0$ for $i = 1, \ldots, 50$. Then, for subject i ($i = 1, \ldots, 50$), for $k = 1, \ldots, 4$, we generate $t_{i,k+1}$ such that $t_{i,k+1} - t_{i,k} \sim N(30, 10^2)$. Let k_i be the first k such that $k \leqslant 4$ and $t_{i,k+1} > 100$. Then, the visit times for subject i are $t_{i,1}, \ldots, t_{i,k_i}$.

At visit time t, subject i has two observations (Y_{it}, Z_{it}) where Y_{it} and Z_{it} are generated according to

$$Y_{it} = \mu(t) + f_y(t)\alpha_i + \epsilon_{it}, \qquad Z_{it} = \nu(t) + f_{z1}(t)\beta_{i1} + f_{z2}(t)\beta_{i2} + \xi_{it}.$$

Here, the mean curves have the form $\mu(t) = 1 + t/100 + \exp\{-(t-60)^2/500\}$ and $\nu(t) = 1 - t/100 - \exp\{-(t-30)^2/500\}$. The principal component curves are $f_y(t) = \sin(2\pi t/100)/\sqrt{50}$, $f_{z1}(t) = f_y(t)$ and $f_{z2}(t) = \cos(2\pi t/100)/\sqrt{50}$. The principal component functions are normalized such that $\int_0^{100} f_y^2(t)\,dt = 1$, $\int_0^{100} f_{z1}^2(t)\,dt = 1$ and $\int_0^{100} f_{z2}^2(t)\,dt = 1$. The variable Z's two principal component curves are orthogonal: $\int_0^{100} f_{z1}(t) f_{z2}(t)\,dt = 0$. The principal component scores α_i, β_{i1} and β_{i2} are independent between subjects, and their distributions are normal with mean 0 and variances $D_\alpha = 36$, $D_{\beta 1} = 36$ and $D_{\beta 2} = 16$, respectively. The variable Z's two principal component scores, β_{i1} and β_{i2}, are independent. In addition, the correlation coefficient between α_i and β_{i1} is $\rho_1 = -0{\cdot}8$ and that between α_i and β_{i2} is $\rho_2 = -0{\cdot}45$. The measurement errors ϵ_{it} and ξ_{it} are independent and normally distributed with mean 0 and variance 0·5.

The penalized likelihood method was applied to fit the joint model with $k_a = 1$ and $k_b = 2$. Penalty parameters were picked using ten-fold crossvalidation on a grid defined by $\lambda_\mu = \lambda_\nu$ in $\{k \times 10^4\}$ and $\lambda_f = \lambda_g$ in $\{2k \times 10^5\}$, for $k = 1, \ldots, 5$. Figure 1 shows fitted mean curves and the principal component curves for five simulated datasets, along with the true curves used in generating the data. Table 1 presents the sample means and mean squared errors of the variance and correlation parameters, based on 200 simulation runs. Our joint modelling approach was compared with a separate modelling approach that fits Y and Z separately using the single-curve method described in § 2·3. In terms of mean squared error, the single-curve method gives similar, but slightly worse estimates of the variance parameters. However, unlike the joint modelling approach, the single-curve method does not provide estimates of the correlation coefficients of the principal component scores. A naive approach is to use the sample correlation coefficients of the best linear unbiased predictors of the principal component scores from the single-curve model.

Table 1. *Sample mean and mean squared error (*MSE*) of variance and correlation parameters in the simulation of* § 6. *'Joint' and 'Separate' refer to, respectively, the joint modelling and separate modelling approach. A number marked with an asterisk equals the actual number multiplied by* 100

	Parameter	ρ_1	ρ_2	D_α	$D_{\beta 1}$	$D_{\beta 2}$	σ_ϵ^2	σ_ξ^2
	True	−0·80	−0·45	36·00	36·00	16·00	0·25	0·25
Joint	Mean	−0·74	−0·49	35·03	35·38	13·08	0·22	0·21
	MSE	2·71*	3·91*	72·11	93·52	25·05	0·15*	0·27*
Separate	Mean	−0·58	−0·37	35·24	36·75	12·88	0·22	0·19
	MSE	6·65*	3·27*	75·07	107·70	30·39	0·19*	0·43*

Since the best linear unbiased predictors are shrinkage estimators, such calculated correlation coefficients can be seriously biased, as shown in Figure 1. Mean integrated squared errors for estimating the mean functions were also computed for the two approaches. The joint modelling approach reduced the mean integrated squared error compared to the separate modelling approach by 23% and 33% for estimating $\mu(\cdot)$ and $\nu(\cdot)$, respectively. It is not surprising that the joint modelling approach is more efficient than separate modelling, as is well known in seemingly unrelated regressions (Zellner, 1962).

7. AIDS STUDY EXAMPLE

In this section we illustrate our model and the proposed estimation method using a dataset from a study conducted by the AIDS Clinical Trials Group, ACTG 315 (Lederman et al., 1998; Wu & Ding, 1999). In this study, 46 HIV 1 infected patients were treated with potent antiviral therapy consisting of ritonavir, 3TC and AZT. After initiation of the treatment on day 0, patients were followed for up to 10 visits. Scheduled visit times common for all patients are 7, 14, 21, 28, 35, 42, 56, 70, 84 and 168 days. Since the patients did not follow exactly the scheduled times and/or missed some visits, the actual visit times are irregularly spaced and different for different patients. The visit time varies from day 0 to day 196. The purpose of our statistical analysis is to understand the relationship between virological and immunological surrogate markers such as plasma HIV RNA copies, called the viral load, and CD4+ cell counts during HIV/AIDS treatments.

In the notation of our joint model for paired functional data in § 3, denote by Y the CD4+ cell counts divided by 100 and by Z the base-10 logarithm of plasma HIV RNA copies. As in Liang et al. (2003), the viral load data below the limit of quantification, i.e. 100 copies per ml of plasma, are imputed by the mid-value of the quantification limit, i.e. 50 copies per ml of plasma. To model the curves on the time interval [0, 196], we used cubic B-splines with 10 interior knots placed on scheduled visit days. The penalty parameters were selected by ten-fold crossvalidation. The resampling-subject bootstrap with 1000 repetitions was used to obtain confidence intervals.

Following the method described in § 5·2, we selected the number of important principal components in two stages. In the first stage, the two variables were modelled separately using the single-curve method in § 2·3. A sequence of models with different numbers of principal component functions were considered, and the corresponding variances of principal component scores for these models are given in Table 2. We decided to use two principal components for both Y and Z. In the second step, the model was fitted jointly with $k_a = 2$ and $k_b = 2$. The estimates of the variances are $D_{\alpha 1} = 110{\cdot}1$, $D_{\alpha 2} = 1{\cdot}147$, $D_{\beta 1} = 169{\cdot}8$ and $D_{\beta 2} = 11{\cdot}8$. Given that the ratio between $D_{\alpha 2}$ and $D_{\alpha 1}$ is about 1%, we decided to drop the second principal component for CD4+ counts and to use $k_a = 1$ and $k_b = 2$ in our final model. The ratio of $D_{\beta 2}$ to $D_{\beta 1}$ is about

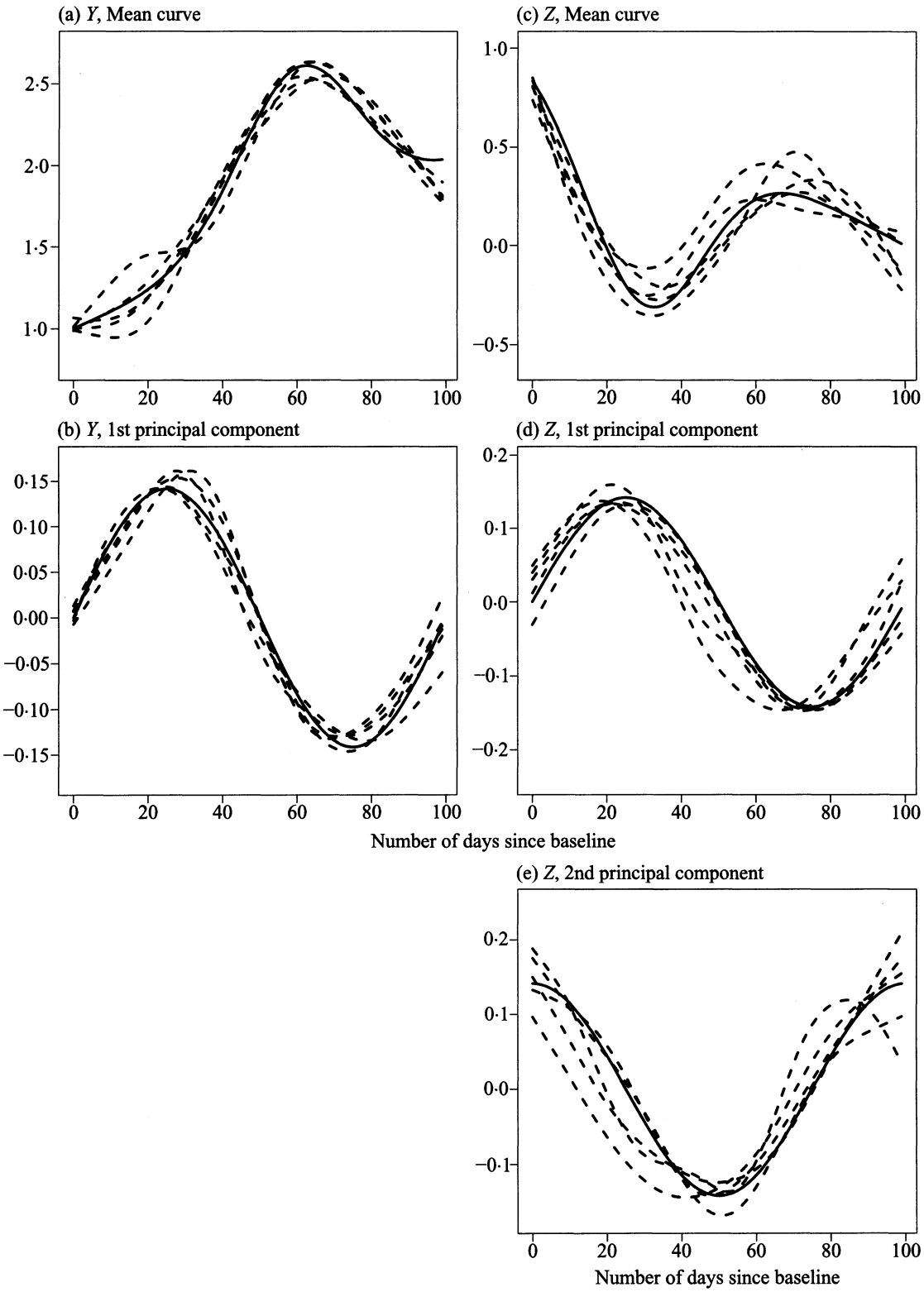

Fig. 1. Fitted mean curves and principal component curves for five simulated datasets: (a) mean curve of Y, (b) first principal component curve, for Y, (c) mean curve for Z, (d) and (e) first and second principal components for Z. Solid lines represent true curves and dashed lines represent the fitted curves for the five simulated datasets.

Table 2. *Estimated variances of principal component scores for models with different numbers of principal components in the* AIDS *example of* § 7. *The variances are ordered in decreasing order for each model*

Number of principal comp.	1	2		3		
Principal comp.	1	1	2	1	2	3
D_α	99·6	122·1	7·8	128·7	9·7	$(<10^{-4})$
D_β	93·1	172·9	11·5	174·4	11·5	$(<10^{-4})$

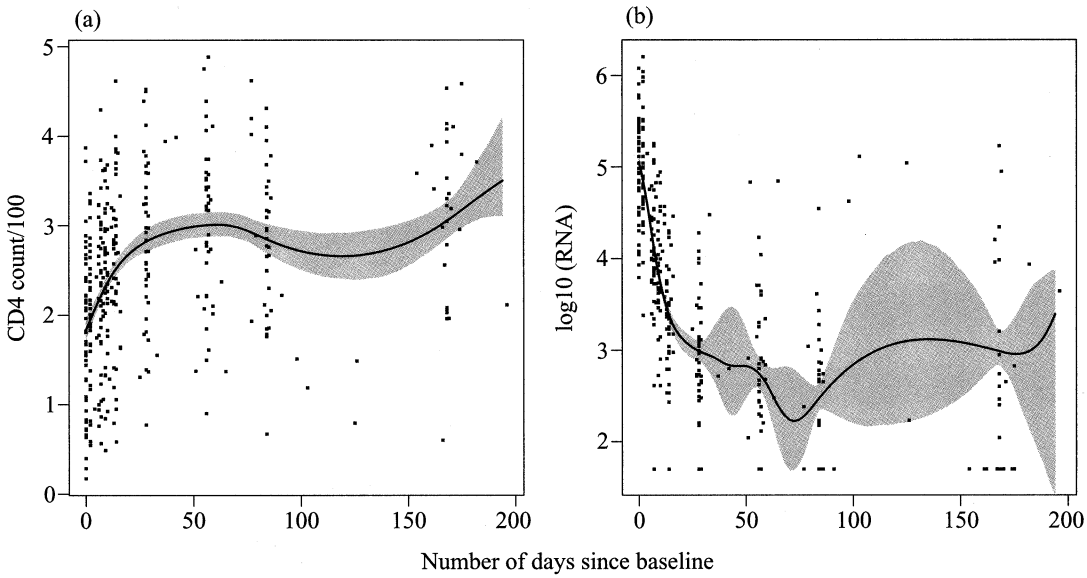

Fig. 2. AIDS study. (a) CD4+ cell counts and (b) viral load over time as a function of days, overlaid by estimated mean curves and corresponding 95% bootstrap pointwise confidence intervals.

7%, so that, for the viral load, the second principal component, even though included in the final model, is much less important than the first.

Figure 2 presents CD4+ cell counts and viral load over time, overlaid by their estimated mean curves and 95% bootstrap pointwise confidence intervals for the means. The plots show that on average, CD4+ cell counts increase while viral load decreases dramatically until day 28. After CD4+ counts plateau out at 28 days, but the viral load still drops until about 50 days. The feature after 50 days in the viral-load plot is an artifact of few observations and an outlier affecting crossvalidation; the feature disappears with a larger smoothing parameter.

Figure 3 shows the estimated principal component curves of CD4+ counts and the viral load, along with the corresponding 95% bootstrap pointwise confidence intervals. The effect on the mean curves of adding and subtracting a multiple of each of the principal component curves is also given in Fig. 3, in which the standard deviations of the corresponding principal component scores are used as the multiplicative factors. The principal component curve for the CD4+ counts is almost constant over the time range and corresponds to an effect of a level shift from the overall mean curve. The first principal component curve for the viral load corresponds to a level shift from the overall mean with the magnitude of the shift increasing with time. The second principal component curve for the viral load changes sign during the time period and corresponds to opposite departures from the mean at the beginning and the end of the time period. Compared with the first principal component, it explains much less variability in the data and can be viewed

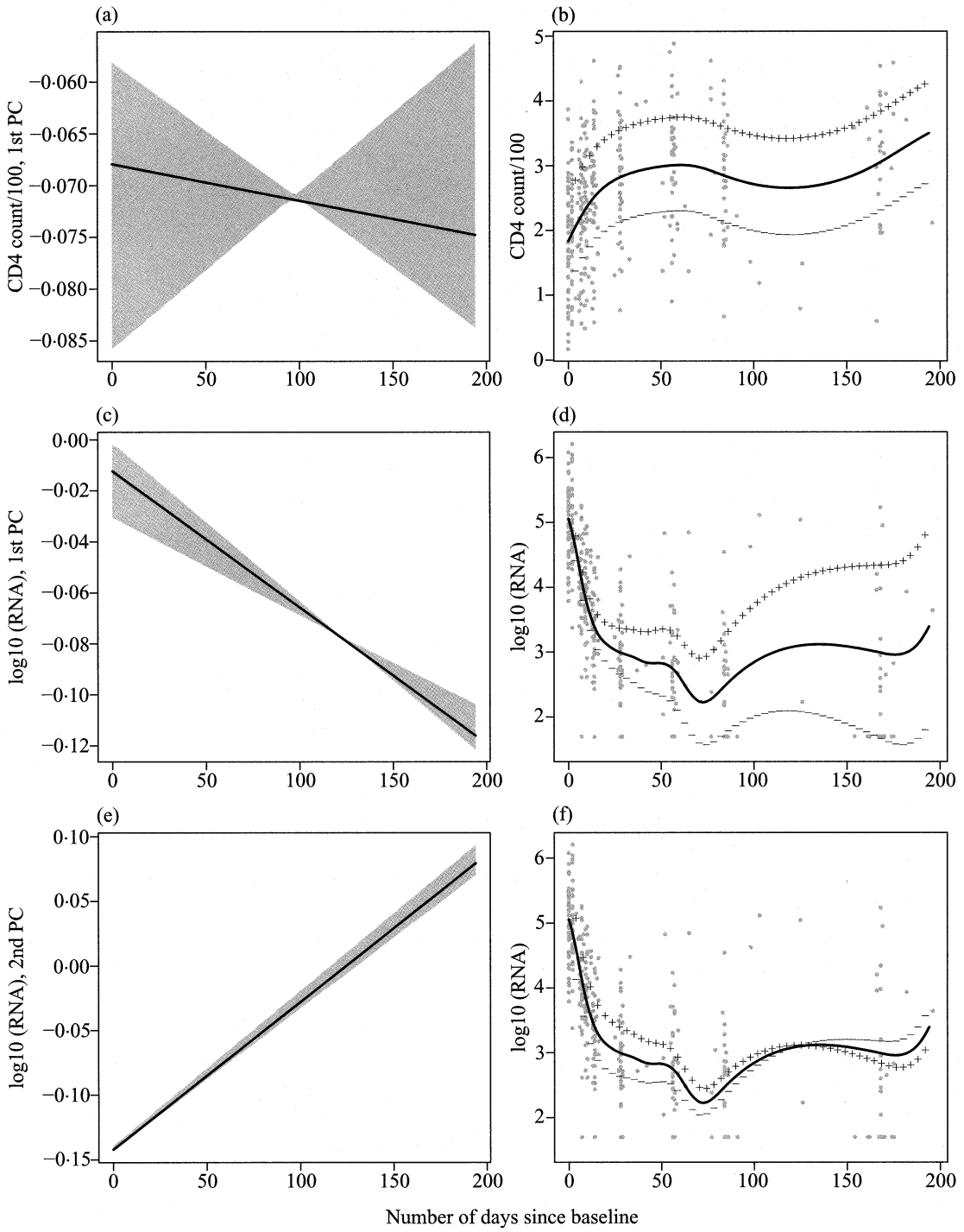

Fig. 3. AIDS study. Estimated principal component curves for (a) CD4+ cell counts and (c) and (e) viral load with corresponding 95% pointwise confidence intervals. (b), (d) and (f) Effect on the mean curves of adding (plus signs) and subtracting (minus signs) a multiple of each of the principal components, shown on the panels (a), (c) and (e).

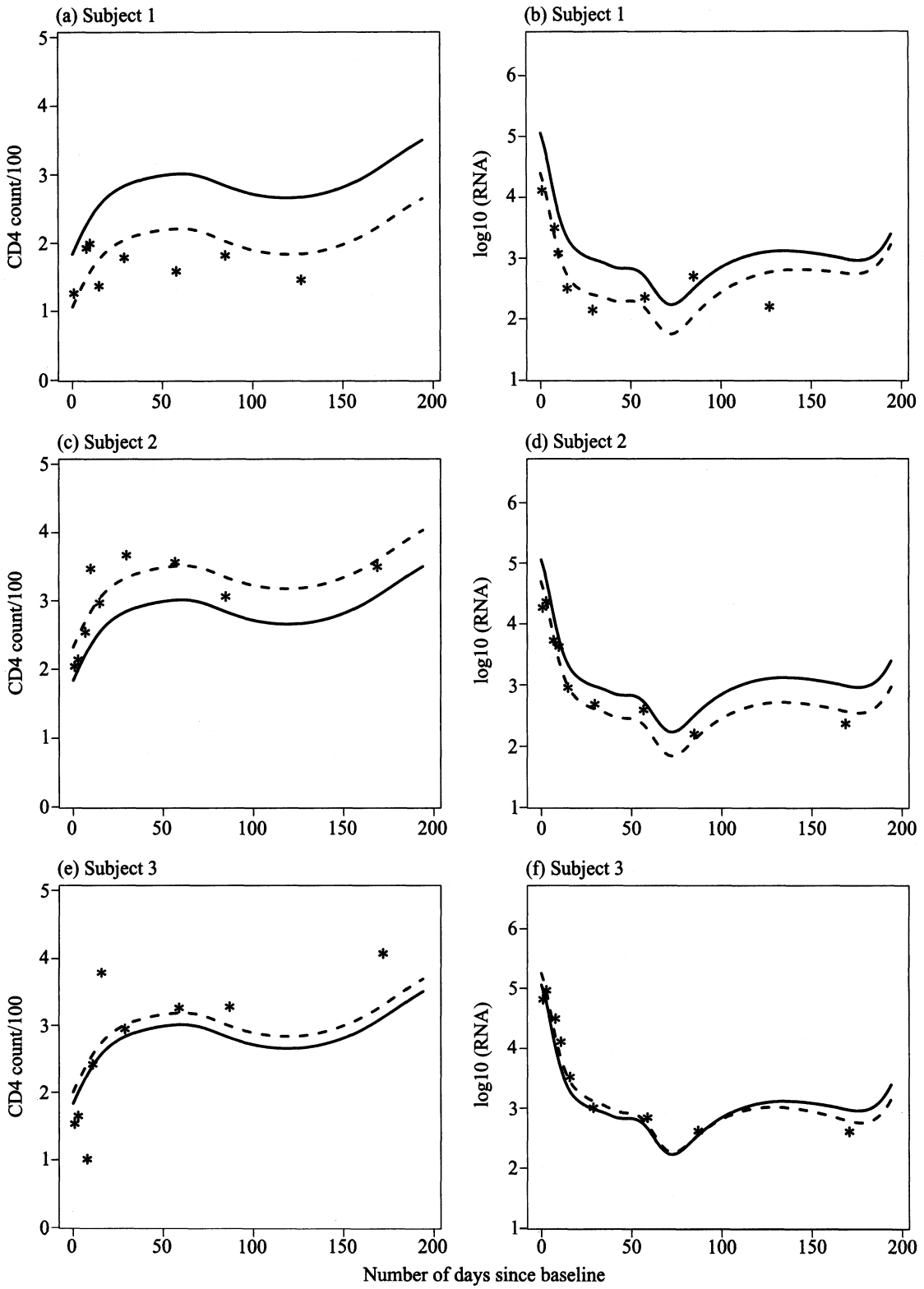

Fig. 4. Data and predictions for three selected subjects. Asterisks denote observed values of (CD4 + count)/100 and $\log_{10}$(RNA), solid lines denote estimated mean curves and dashed lines denote best linear unbiased predictions of the subject-specific curves.

Table 3. *Estimates of variance and correlation parameters and their 95% bootstrap confidence intervals in the* AIDS *example of* § 7. *'Lower' and 'upper' represent the lower and upper end points of the confidence intervals* (CI)

Parameter	ρ_1	ρ_2	D_α	$D_{\beta 1}$	$D_{\beta 2}$	σ_ϵ^2	σ_ξ^2
Estimate	−0·35	0·04	106·30	170·50	11·66	0·25	0·13
95% CI, lower	−0·92	−0·08	54·68	96·20	5·74	0·20	0·09
95% CI, upper	−0·05	0·08	163·60	302·52	17·43	0·30	0·16

as a correction factor to the prediction made by the first principal component. We did not know the shape of the principal component curves prior to the analysis, but it turns out that all estimated principal component curves are rather smooth and close to linear. This may be caused by the high level of noise in the data that prevents the identification of more subtle features; the data-driven crossvalidation does not support the use of smaller penalties. Given that the two principal components are obtained from a high-dimensional function space, dimension reduction is quite effective in this example.

In Fig. 4, we plot observed data for three typical subjects and corresponding mean curves and best linear unbiased predictions of the underlying subject-specific curves. The predicted values of the scores corresponding to the first principal component of CD4+ counts are 11·43, −7·15 and −2·49 for the three subjects, respectively. The predicted values of the scores of the first principal component of viral load are 4·43, 5·11 and 1·05, and those of the second principal component are 4·26, 2·08 and −1·50. These predicted scores and the graphs in Fig. 4 agree with the interpretation of the principal components given in the previous paragraph. For example, the first subject has a positive score while the second and third subjects have negative scores on the first principal component of CD4+ counts, corresponding to a downward and upward shift of the predicted curves from the mean curve, respectively. The crossover effect of the second principal component of viral load is clearly seen in the third subject.

Estimates of variance and correlation parameters are given in Table 3 together with the corresponding 95% bootstrap confidence intervals. Of particular interest is the parameter ρ_1, the correlation coefficient between α_{i1} and β_{i1}, which are the scores corresponding to the first principal component of CD4+ counts and viral load, respectively. The estimated ρ_1 is statistically significantly negative, which suggests that a positive score on the first principal component of CD4+ counts tends to be associated with a negative score on the first principal component of viral load. In other words, for a subject with CD4+ count lower, respectively higher, than the mean, the viral load tends to be higher, respectively lower, than the mean.

ACKNOWLEDGEMENT

Lan Zhou was supported by a post-doctoral training grant from the U.S. National Cancer Institute. Jianhua Z. Huang was partially supported by grants from the U.S. National Science Foundation and the U.S. National Cancer Institute. Raymond J. Carroll was supported by grants from the U.S. National Cancer Institute.

APPENDIX 1

Creation of a basis $b(t)$ that satisfies the orthonormal constraints

Let $\tilde{b}(t) = (\tilde{b}_1(t), \ldots, \tilde{b}_q(t))^{\mathrm{T}}$ be an initially chosen, not necessarily orthonormal, basis such as the B-spline basis. A transformation matrix T such that $b(t) = T\tilde{b}(t)$ can be constructed as follows. Write

$\tilde{B} = \{\tilde{b}(t_1), \ldots, \tilde{b}(t_g)\}^{\mathrm{T}}$. Let $\tilde{B} = \mathrm{QR}$ be the QR decomposition of $\tilde{B}$, where Q has orthonormal columns and R is an upper triangular matrix. Then, $T = (g/L)^{1/2} R^{-\mathrm{T}}$ will be a desirable transformation matrix since

$$\frac{L}{g} B^{\mathrm{T}} B = \frac{L}{g} T \tilde{B}^{\mathrm{T}} \tilde{B} T^{\mathrm{T}} = \frac{L}{g} T R^{\mathrm{T}} Q^{\mathrm{T}} Q R T^{\mathrm{T}} = I.$$

Appendix 2

Conditional moments of the multivariate normal distribution (16)

Write Σ_i^{-1} as

$$\Sigma_i^{-1} = \begin{pmatrix} \Sigma_i^{\alpha\alpha}, \Sigma_i^{\alpha\beta} \\ \Sigma_i^{\beta\alpha}, \Sigma_i^{\beta\beta} \end{pmatrix}.$$

Then, the conditional distribution satisfies

$$f(\alpha_i, \beta_i \mid Y_i, Z_i) \propto \exp\left[-\frac{1}{2}\{(\alpha_i - \mu_{i,\alpha})^{\mathrm{T}}, (\beta_i - \mu_{i,\beta})^{\mathrm{T}}\} \begin{pmatrix} \Sigma_i^{\alpha\alpha}, \Sigma_i^{\alpha\beta} \\ \Sigma_i^{\beta\alpha}, \Sigma_i^{\beta\beta} \end{pmatrix} \begin{pmatrix} \alpha_i - \mu_{i,\alpha} \\ \beta_i - \mu_{i,\beta} \end{pmatrix}\right].$$

On the other hand, $f(\alpha_i, \beta_i \mid Y_i, Z_i) \propto f(\alpha_i, \beta_i, Y_i, Z_i)$. Comparing the coefficients of the quadratic forms $\alpha_i^{\mathrm{T}} \Sigma_i^{\alpha\alpha} \alpha_i$, $\alpha_i^{\mathrm{T}} \Sigma_i^{\alpha\beta} \beta_i$ and $\beta_i^{\mathrm{T}} \Sigma_i^{\beta\beta} \beta_i$ in the two expressions of the conditional distribution, we obtain

$$\begin{aligned} \Sigma_i^{\alpha\alpha} &= D_\alpha^{-1} + \Lambda^{\mathrm{T}} \Sigma_\eta^{-1} \Lambda + \sigma_\epsilon^{-2} \Theta_f^{\mathrm{T}} B_i^{\mathrm{T}} B_i \Theta_f, \\ \Sigma_i^{\alpha\beta} &= -\Lambda^{\mathrm{T}} \Sigma_\eta^{-1}, \\ \Sigma_i^{\beta\beta} &= \Sigma_\eta^{-1} + \sigma_\xi^{-2} \Theta_g^{\mathrm{T}} B_i^{\mathrm{T}} B_i \Theta_g. \end{aligned}$$

These can be used to calculate $\Sigma_{i,\alpha\alpha}$, $\Sigma_{i,\alpha\beta}$ and $\Sigma_{i,\beta\beta}$ through the formulae

$$\begin{aligned} \Sigma_{i,\alpha\alpha} &= \left\{\Sigma_i^{\alpha\alpha} - \Sigma_i^{\alpha\beta} \left(\Sigma_i^{\beta\beta}\right)^{-1} \Sigma_i^{\beta\alpha}\right\}^{-1}, \\ \Sigma_{i,\alpha\beta} &= -\Sigma_{i,\alpha\alpha} \Sigma_i^{\alpha\beta} \left(\Sigma_i^{\beta\beta}\right)^{-1}, \\ \Sigma_{i,\beta\beta} &= \left\{\Sigma_i^{\beta\beta} - \Sigma_i^{\beta\alpha} \left(\Sigma_i^{\alpha\alpha}\right)^{-1} \Sigma_i^{\alpha\beta}\right\}^{-1}, \end{aligned}$$

or direct matrix inversion of Σ_i^{-1}.

Similarly, comparing the coefficients of the first-order terms, we obtain

$$\begin{aligned} \Sigma_i^{\alpha\alpha} \mu_{i,\alpha} + \Sigma_i^{\alpha\beta} \mu_{i,\beta} &= \sigma_\epsilon^{-2} \Theta_f^{\mathrm{T}} B_i^{\mathrm{T}} (Y_i - B_i \theta_\mu), \\ \Sigma_i^{\beta\alpha} \mu_{i,\alpha} + \Sigma_i^{\beta\beta} \mu_{i,\beta} &= \sigma_\xi^{-2} \Theta_g^{\mathrm{T}} B_i^{\mathrm{T}} (Z_i - B_i \theta_\nu), \end{aligned}$$

which implies that

$$\begin{aligned} \mu_{i,\alpha} &= \sigma_\epsilon^{-2} \Sigma_{i,\alpha\alpha} \Theta_f^{\mathrm{T}} B_i^{\mathrm{T}} (Y_i - B_i \theta_\mu) + \sigma_\xi^{-2} \Sigma_{i,\alpha\beta} \Theta_g B_i^{\mathrm{T}} (Z_i - B_i \theta_\nu), \\ \mu_{i,\beta} &= \sigma_\epsilon^{-2} \Sigma_{i,\beta\alpha} \Theta_f^{\mathrm{T}} B_i^{\mathrm{T}} (Y_i - B_i \theta_\mu) + \sigma_\xi^{-2} \Sigma_{i,\beta\beta} \Theta_g^{\mathrm{T}} B_i^{\mathrm{T}} (Z_i - B_i \theta_\nu). \end{aligned}$$

APPENDIX 3

Updating formulae for the M*-step of the* EM *algorithm*

In the updating formulae given below, the parameters that appear on the right-hand side of equations are all fixed at their current estimates. The involved conditional moments are as defined in (17) of § 4·2.

Step 1. Update the estimates of σ_ϵ^2 and σ_ξ^2. We update σ_ϵ^2 using $\hat{\sigma}_\epsilon^2 = \sum_{i=1}^n E(\epsilon_i^{\mathrm{T}}\epsilon_i \mid Y_i)/\sum_{i=1}^n n_i$ and σ_ξ^2 similarly. The updating formulae are

$$\hat{\sigma}_\epsilon^2 = \frac{1}{\sum_{i=1}^n n_i}\sum_{i=1}^n \{(Y_i - B_i\theta_\mu - B_i\Theta_f\hat{\alpha}_i)^{\mathrm{T}}(Y_i - B_i\theta_\mu - B_i\Theta_f\hat{\alpha}_i) + \mathrm{tr}\big(B_i\Theta_f\Sigma_{i,\alpha\alpha}\Theta_f^{\mathrm{T}}B_i^{\mathrm{T}}\big)\},$$

$$\hat{\sigma}_\xi^2 = \frac{1}{\sum_{i=1}^n n_i}\sum_{i=1}^n \{(Z_i - B_i\theta_\nu - B_i\Theta_g\hat{\beta}_i)^{\mathrm{T}}(Z_i - B_i\theta_\nu - B_i\Theta_g\hat{\beta}_i) + \mathrm{tr}\big(B_i\Theta_g\Sigma_{i,\beta\beta}\Theta_g^{\mathrm{T}}B_i^{\mathrm{T}}\big)\}.$$

Step 2. Update the estimates of θ_μ and θ_ν. The updating formulae are

$$\hat{\theta}_\mu = \left\{\sum_{i=1}^n B_i^{\mathrm{T}}B_i + \sigma_\epsilon^2\lambda_\mu\int b''(t)b''(t)^{\mathrm{T}}\,dt\right\}^{-1}\sum_{i=1}^n B_i^{\mathrm{T}}(Y_i - B_i\Theta_f\hat{\alpha}_i),$$

$$\hat{\theta}_\nu = \left\{\sum_{i=1}^n B_i^{\mathrm{T}}B_i + \sigma_\xi^2\lambda_\nu\int b''(t)b''(t)^{\mathrm{T}}\,dt\right\}^{-1}\sum_{i=1}^n B_i^{\mathrm{T}}(Z_i - B_i\Theta_g\hat{\beta}_i).$$

Step 3. Update the estimates of Θ_f and Θ_g. We update the columns of Θ_f and Θ_g sequentially. Write $\Theta_f = (\theta_{\alpha 1}, \theta_{\alpha 2}, \ldots, \theta_{\alpha k_\alpha})$ and $\Theta_g = (\theta_{\beta 1}, \ldots, \theta_{\beta k_\beta})$. For $j = 1, \ldots, k_\alpha$, we minimize

$$\sum_{i=1}^n E\left(\left\|Y_i - B_i\theta_\mu - \sum_{l\neq j} B_i\theta_{\alpha l}\alpha_{il} - B_i\theta_{\alpha j}\alpha_{ij}\right\|^2 \middle| Y_i, Z_i\right) + \sigma_\epsilon^2\lambda_f\theta_{fj}^{\mathrm{T}}\int b''(t)b''(t)^{\mathrm{T}}\,dt\,\theta_{fj}$$

with respect to $\theta_{\alpha j}$. The solution gives the update for $\theta_{\alpha j}$,

$$\hat{\theta}_{\alpha j} = \left\{\sum_{i=1}^n \hat{\alpha}_{ij}^2 B_i^{\mathrm{T}}B_i + \sigma_\epsilon^2\lambda_f\int b''(t)b''(t)^{\mathrm{T}}\,dt\right\}^{-1}\sum_{i=1}^n B_i^{\mathrm{T}}\left\{(Y_i - B_i\theta_\mu)\hat{\alpha}_{ij} - \sum_{l\neq j} B_i\theta_{\alpha l}M_{i,\alpha\alpha}(l,j)\right\}.$$

Similarly, for $j = 1, \ldots, k_\beta$,

$$\hat{\theta}_{\beta j} = \left\{\sum_{i=1}^n \hat{\beta}_{ij}^2 B_i^{\mathrm{T}}B_i + \sigma_\xi^2\lambda_g\int b''(t)b''(t)^{\mathrm{T}}\,dt\right\}^{-1}\sum_{i=1}^n B_i^{\mathrm{T}}\left\{(Z_i - B_i\theta_\nu)\hat{\beta}_{ij} - \sum_{l\neq j} B_i\theta_{\beta l}M_{i,\beta\beta}(l,j)\right\}.$$

Step 4. Update the estimate of Λ. The updating formula is

$$\hat{\Lambda} = \left(\sum_{i=1}^n M_{i,\beta\alpha}\right)\left(\sum_{i=1}^n M_{i,\alpha\alpha}\right)^{-1}.$$

Step 5. Orthogonalization. The matrices Θ_f and Θ_g obtained in Step 3 need not have orthonormal columns. We orthogonalize them in this step and also provide an updated estimate of D_α, D_β and Λ. Compute

$$\hat{\Sigma}_\alpha = \frac{1}{n}\sum_{i=1}^n M_{i,\alpha\alpha} \quad \text{and} \quad \hat{\Sigma}_\beta = \frac{1}{n}\sum_{i=1}^n M_{i,\beta\beta}.$$

Let $\hat{\Theta}_f\hat{\Sigma}_\alpha\hat{\Theta}_f^{\mathrm{T}} = Q_f S_\alpha Q_f^{\mathrm{T}}$ be the eigenvalue decomposition in which Q_f has orthogonal columns and S_α is diagonal with diagonal elements arranged in decreasing order. The updated $\hat{\Theta}_f$ is Q_f and the

updated $\hat{D}_\alpha$ is S_α. Similarly, let $\hat{\Theta}_g \hat{\Sigma}_\beta \hat{\Theta}_g^{\mathrm{T}} = Q_g S_\beta Q_g^{\mathrm{T}}$ be the eigenvalue decomposition in which Q_g has orthogonal columns and S_β is diagonal with diagonal elements arranged in decreasing order. The updated $\hat{\Theta}_g$ is Q_g and the updated $\hat{D}_\beta$ is S_β. The orthogonalization process corresponds to transformations $\alpha_i \leftarrow Q_f^{\mathrm{T}} \hat{\Theta}_f \alpha_i$ and $\beta_i \leftarrow Q_g^{\mathrm{T}} \hat{\Theta}_g \beta_i$. Thus, the corresponding transformation for $\hat{\Lambda}$ obtained from Step 4 is $\hat{\Lambda} \leftarrow (Q_g^{\mathrm{T}} \Theta_g) \hat{\Lambda} (Q_f^{\mathrm{T}} \hat{\Theta}_f)^{-1}$.

References

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

Eilers, P. & Marx, B. (1996). Flexible smoothing with *B*-splines and penalties (with Discussion). *Statist. Sci.* **89**, 89–121.

Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* New York: Springer.

He, G., Müller, H.-G. & Wang, J.-L. (2003). Functional canonical analysis for square integrable stochastic processes. *J. Mult. Anal.* **85**, 54–77.

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–22.

Huang, J. Z., Wu, C. O. & Zhou, L. (2002). Varying coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika* **89**, 111–28.

James, G. M., Hastie, T. J. & Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

Laird, N. & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.

Lederman, M. M., Connick, E., Landay, A., Kuritzkes, D. R., Spritzler, J., Clair, M. S., Kotzin, B. L., Fox, L., Chiozzi, M. H., Leonard, J. M., Rousseau, F., Wade, M., D'arc Roe, J., Martinez, A. & Kessler, H. (1998). Immunological responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine & ritonavir: results of AIDS Clinical Trials Group Protocol 315. *J. Inf. Dis.* **178**, 70–9.

Leurgans, S. E., Moyeed, R. A. & Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *J. R. Statist. Soc.* B **55**, 725–40.

Liang, H., Wu, H. & Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics* **4**, 297–312.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Moyeed, R. A. & Diggle, P. J. (1994). Rates of convergence in semi-parametric modelling of longitudinal data. *Aust. J. Statist.* **36**, 75–93.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Comp. J.*, **7**, 308–13.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis,* 2nd ed. New York: Springer.

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statist. Sinica* **14**, 613–29.

Rice, J. A. & Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–59.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

Shi, M., Weiss, R. E. & Taylor, J. M. G. (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Appl. Statist.* **45**, 151–63.

Wu, C. O., Chiang, C.-T. & Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Am. Statist. Assoc.* **93**, 1388–402.

Wu, H. & Ding, A. (1999). Population HIV-1 dynamics in vivo: applicable models and inference tools for virological data from AIDS clinical trials. *Biometrics* **55**, 410–8.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.* **100**, 577–90.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005b) Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873–903.

Zeger, S. L. & Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–99.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions, and tests for aggregation bias. *J. Am. Statist. Assoc.* **57**, 348–68.

[*Received December* 2006. *Revised March* 2008]