Principal component models for sparse functional data

GARETH M. JAMES
Marshall School of Business, University of Southern California, Los Angeles, California 90089-0809
gareth@usc.edu
TREVOR J. HASTIE
Department of Statistics, Stanford University, California 94305-4065
trevor@stat.stanford.edu
and CATHERINE A. SUGAR
Marshall School of Business, University of Southern California, Los Angeles, California 90089-0809
sugar@usc.edu

March 1, 2001

SUMMARY

The elements of a multivariate data set are often curves rather than single points. Functional principal components can be used to describe the modes of variation of such curves. If one has complete measurements for each individual curve or, as is more common, one has measurements on a fine grid taken at the same time points for all curves, then many standard techniques may be applied. However, curves are often measured at an irregular and sparse set of time points which can differ widely across individuals. We present a technique for handling this more difficult case using a reduced rank mixed effects framework.

Some key words: Functional data analysis; Principal components; Mixed effects model; Reduced rank estimation; Growth curve.

1. INTRODUCTION

1·1. *The problem*

We present a technique for fitting principal component functions to data such as the growth curves illustrated in Fig. 1(a). These data consist of measurements of spinal bone mineral density for forty-eight females taken at various ages. They are a subset of the data presented in Bachrach et al. (1999). Even though only partial curves are available for each individual, there is a clear trend in the data. The solid curve gives an estimate for the mean function. It highlights the rapid growth that occurs during puberty. However, the mean function does not explain all the variability in the data.

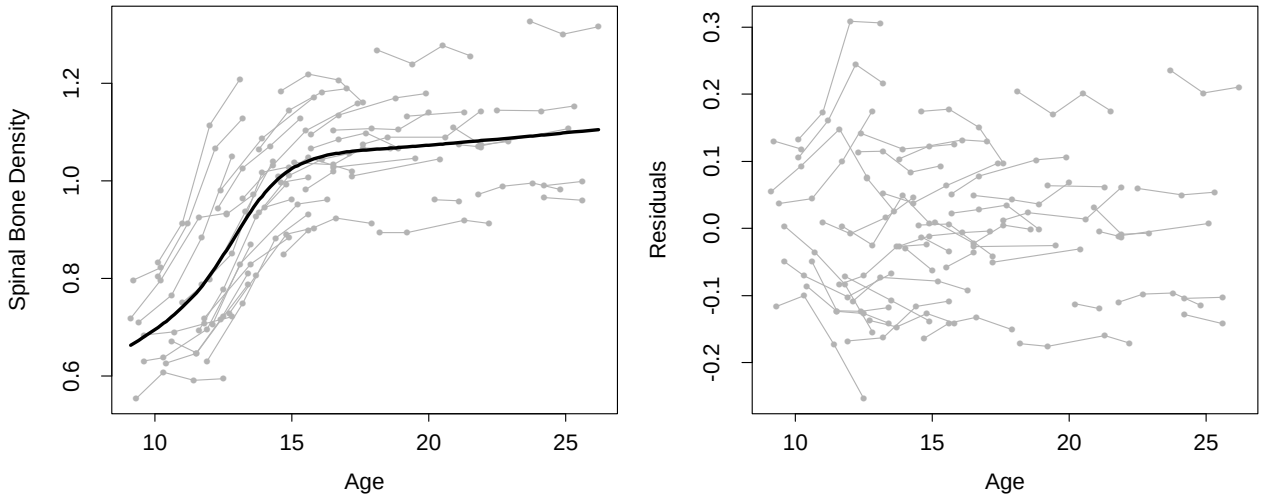

Figure 1: The data are measurements of spinal bone mineral density for forty-eight white females. There are between 2 and 4 measurements per subject (160 in all) indicated by the growth curve fragments in the plots. The solid line in (a) is an estimate for the population mean growth curve. The residuals are shown in (b). The variability of the residuals is smallest in childhood and increases slightly during the period associated with the adolescent growth spurt.

The residual plot, Fig. 1(b), is narrower during early childhood, thickens during puberty, and then narrows again as adulthood approaches. It would be useful to be able to estimate both the entire growth curve for each individual and the principal component curve or curves for the population as a whole. There is an extensive literature on such problems when individuals are measured at the same time points; for an early example involving growth curve data see Gasser et al. (1984) and for a summary of more recent work see Ramsay & Silverman (1997). However, it is not clear what is the best procedure when the time points vary among individuals. We present an estimation technique that is particularly useful when the data are sparse with measurements for individuals occurring at possibly differing time points.

1·2. *A direct approach*

When a set of N curves is measured on a fine grid of n equally spaced points the functional principal components problem can be solved by applying standard principal components analysis to the N by n matrix of observed data. Often the grid is sparse or the time-points are unequally spaced, although still common to all curves. In this case, one can impose smoothness constraints on the principal components in several ways. One simple approach is to represent them using a set of smooth basis functions. This amounts to projecting the individual rows of the data matrix on to the basis and then performing principal component analysis on the basis coefficients. Alternatively one can use the basis coefficients to estimate the individual curves, sample the curves on a fine grid and perform principal component analysis on the resulting 'data.'

When the curves are not measured at common time points one can still project each curve on

to a common basis and then perform principal component analysis on the estimated coefficients or curves. We call this procedure the direct method. It has two major drawbacks. First, if there are individuals with few measurements it may not be possible to produce a unique representation for every curve so the direct approach can't be used. Secondly, the direct method does not make optimal use of the available information because it treats estimated curves as if they were observed. All estimated values receive equal weight despite the irregular spacing of the observed data. Intuitively it seems desirable to take into account the relative accuracies of the estimated points; see Rice & Silverman (1991), Besse & Cardot (1996), Buckheit et al. (1997) and Besse, Cardot & Ferraty (1997) for interesting applications, variations and extensions of the direct method.

1·3. *A mixed effects approach*

Mixed-effects models have been widely used in the analysis of curve data; see for instance Brumback & Rice (1998). Shi, Weiss & Taylor (1996) and Rice & Wu (2000) suggest using a mixed-effects approach to solve the functional principal components problem. Their model uses a set of smooth basis functions, $b_\ell(t),\ \ell = 1,\ldots,q$, such as B-splines, to represent the curves. Let $Y_i(t)$ be the value for the ith curve at time t and let $b(t) = [b_1(t), b_2(t), \ldots, b_q(t)]^{\mathrm{T}}$ be the vector of basis functions evaluated at time t. Denote by β an unknown but fixed vector of spline coefficients, let γ_i be a random vector of spline coefficients for each curve with population covariance matrix Γ, and let $\varepsilon_i(t)$ be random noise with mean zero and variance σ^2. The resulting mixed effects model has the form

$$Y_i(t) = b(t)\beta + b(t)\gamma_i + \varepsilon_i(t) \quad i = 1,\ldots,N. \tag{1}$$

In practice $Y_i(t)$ is only observed 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 by 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 then becomes

$$Y_i = B_i\beta + B_i\gamma_i + \varepsilon_i \quad i = 1,\ldots,N. \tag{2}$$

The fixed-effects term $B_i\beta$ models the mean curve for the population and the random-effects term $B_i\gamma_i$ allows for individual variation. The principal patterns of variation about the mean curve are referred to as functional principal component curves. Rice & Wu (2000) suggest modelling the patterns of variation of the basis coefficients, γ_i, and then transforming back to the original space. Since Γ is the covariance matrix of the γ_i's, this is achieved by multiplying the eigenvectors of Γ by $b(t)$.

A general approach to fitting mixed effects models of this form uses the EM algorithm to estimate β and Γ (Laird & Ware, 1982). Given these estimates, predictions are obtained for the γ_i's using best linear unbiased prediction (Henderson, 1950). For (2) above, the best linear unbiased prediction for γ_i is

$$\hat{\gamma}_i = (\hat{\Gamma}^{-1}/\sigma^2 + B_i^{\mathrm{T}}B_i)^{-1}B_i^{\mathrm{T}}(Y_i - B_i\hat{\beta}). \tag{3}$$

Using the fitted values of β and Γ one can estimate the mean and principal component curves and

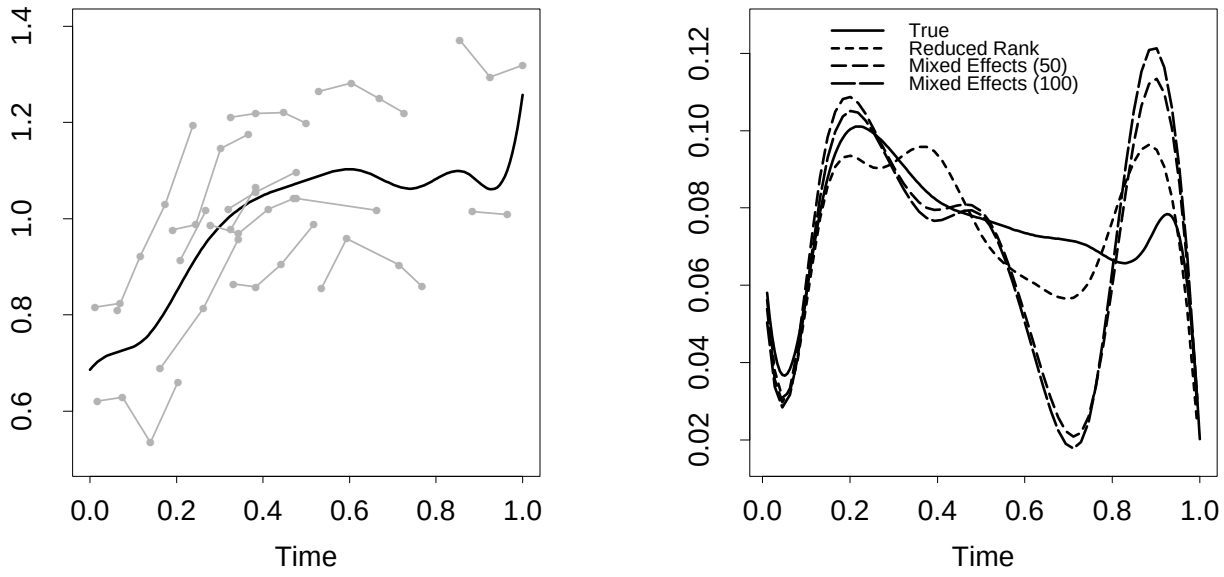

Figure 2: (a) A dataset simulated from a mean function plus one principal component curve plus random normal noise. Both the mean function and principal component are cubic splines with seven equally spaced knots. (b) Estimates for the first principal component curve for the dataset in (a). The solid line is the true principal component curve we are trying to estimate.

by combining these estimates with the prediction for γ_i one can also predict the individual curve $Y_i(t)$.

The mixed effects method has many advantages over the direct method. First, it estimates the curve $Y_i(t)$ using all the observed data points rather than just those from the ith individual. This means that the mixed effects method can be applied when there are insufficient data from each individual curve to use the direct method. Secondly, it uses maximum likelihood to estimate β and Γ. Thus it automatically assigns the correct weight to each observation and the resulting estimators have all the usual asymptotic optimality properties.

1·4. *Some problems with the mixed effects method*

If the dimension of the spline basis is q then in fitting Γ we must estimate $q(q+1)/2$ different parameters. With a sparse data set these estimates can be highly variable. This not only makes the estimates suspect but also means that the likelihood tends to have many local maxima. As a result, the EM algorithm will often fail to converge to the global maximum. Figure 2(a) illustrates a simulated dataset of sixteen curve fragments. Each curve was generated by adding a random multiple of a single principal component curve to a mean function. Random normal noise, with standard deviation of 0.02, was added to produce the final data. Both the mean function and principal component curve are cubic splines with seven equally spaced knots. The goal is to estimate the principal component curve, a difficult problem since there were only fifty-one data points in total.

The direct method can't even be applied to this dataset because there are too few measurements

per curve for estimating a separate spline for each. The mixed effects model can be applied but, after including a constant term, the dimension of the spline basis is eleven. As a result we are attempting to estimate sixty-six parameters using fifty-one data points and so there is no unique representation of Γ. This is not necessarily a problem since it is the first eigenvector of Γ that is of primary interest. In standard principal component analysis it is often possible to estimate well the first few eigenvectors, and hence the first few principal components, even if the fitted covariance matrix is unstable. However, in functional principal component analysis this is generally not the case, as illustrated in Fig. 2(b). The solid line gives the true principal component curve for the dataset. There are also three estimates, each using cubic splines with seven equally spaced knots. The two dashed lines are estimates using the mixed effects method with fifty and one-hundred EM iterations; the algorithm had converged after one-hundred iterations. The mixed effects method's estimates are poor in the second half of the plot. Furthermore, the fit appears to be deteriorating as the number of iterations increases. This suggests that the procedure is over-fitting the data. The forth line is the estimate produced by the reduced rank method introduced in this paper. This method attempts to estimate the principal component curve directly rather than estimating an entire covariance matrix and computing the first eigenvector. This involves estimating fewer parameters and as a result the fitted curve is less variable and generally more accurate. The reduced rank and mixed effects methods are compared on the growth curve data in § 2 and on more extensive simulated data in § 5. In § 3 we present the reduced rank model and compare it to the mixed effects model. § 4 motivates and outlines the reduced rank fitting procedure. The simulations in § 5 suggest that the reduced rank method gives superior fits and is less sensitive to sparse data. Methods for selecting the dimension of the spline basis, choosing the number of principal component curves, and producing confidence intervals are given in § 6. § 7 relates the reduced rank method to standard principal components analysis.

2. THE GROWTH CURVE DATA

Here we fit the reduced rank and mixed effects procedures to the growth curve data illustrated in Fig. 1. Estimates for the mean function and first principal component using natural cubic splines with four, nine and fourteen equally spaced knots are shown in Fig. 3(a) - (f). The two methods produce fairly similar estimates of the principal component curves but some differences are apparent. Not surprisingly, both procedures display more variability as the number of knots increases. However, a sharp peak near the age of 13 followed by a leveling off is apparent in all three of the reduced rank fits. This is consistent with the residual plot in Fig. 1(b). The mixed effects procedure only displays a strong peak for the nine-knot fit. The peak in the four-and fourteen-knot fits is much less well defined. There is also an anomalous dip in the nine-knot mixed effects method fit around the age of 22. Naturally, given the sparseness of the data, one must be careful not to over-interpret the results. Figures 3(g) and (h) give new estimates for the mean and first principal component of the growth curve data using natural cubic splines with knots at ages $12, 14, 16$ and 18, a spacing suggested by previous experience with this data set. This gives added flexibility during the puberty period when variability among individuals is likely to be highest. As with the previous knot selection, there appears to be a peak near age 13 that is much more marked in the reduced rank fit.

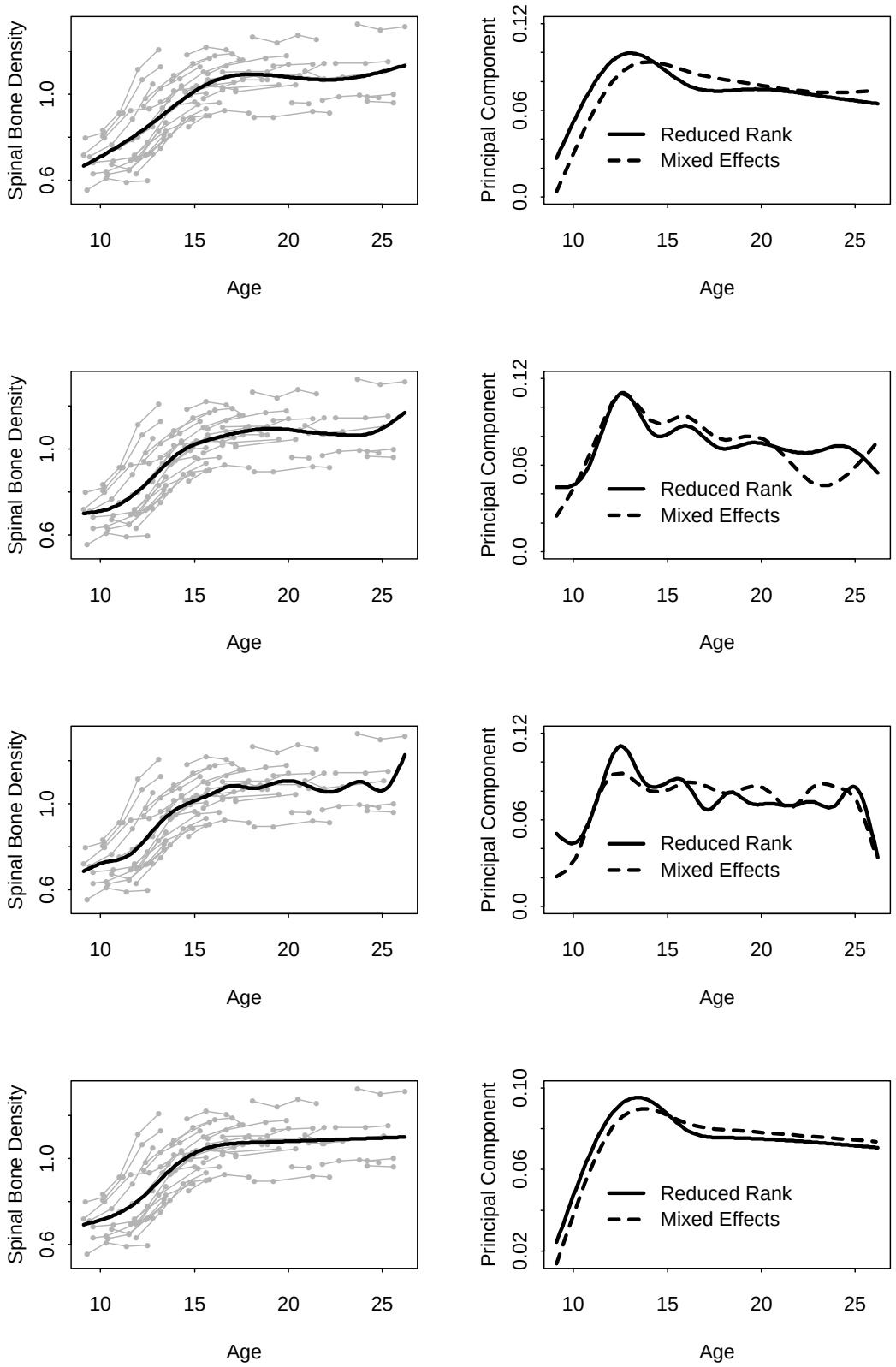

Figure 3: Three different estimates for (a), (c) and (e) the mean function and (b), (d) and (f) the first principal component using natural cubic splines with four, nine and fourteen equally spaced knots. (g) and (h) show mean function and first principal component, respectively, using a natural cubic spline with knots at ages 12, 14, 16 and 18 years.

3. THE REDUCED RANK MODEL

Here we develop our reduced rank model and show that one can interpret the mixed effects model in terms of this framework. In the process, the reasons for some of the mixed effects method's defects become apparent. Let $Y_i(t)$ be the measurement at time t for the ith individual or curve. Let $\mu(t)$ be the overall mean function, let f_j be the jth principal component function and let $f = (f_1, f_2, \ldots, f_k)^{\mathrm{T}}$. To estimate k principal component curves we first define a general additive model

$$\begin{aligned} Y_i(t) &= \mu(t) + \sum_{j=1}^{k} f_j(t)\alpha_{ij} + \varepsilon_i(t) \quad i = 1, \ldots, N \\ &= \mu(t) + f(t)^{\mathrm{T}}\alpha_i + \varepsilon_i(t) \quad i = 1, \ldots, N, \end{aligned} \tag{4}$$

subject to the orthogonality constraint $\int f_j f_l = \delta_{jl}$, the Kronecker δ. The random vector α_i gives the relative weights on the principal component functions for the ith individual and $\varepsilon_i(t)$ is random measurement error. The α_i's and ε_i's are all assumed to have mean zero. The α_i's are taken to have a common covariance matrix, Σ, and the measurement errors are assumed uncorrelated with a constant variance of σ^2. If Σ is diagonal one can interpret (4) as a principal factor decomposition of the covariance kernel of $Y_i(t)$. A more general structure, R, could be assumed for the error term covariance matrix. This would increase the flexibility of the model but would involve estimating extra parameters. For this paper we have opted for the simpler covariance structure.

In order to fit this model when the data are measured at only a finite number of time points it is necessary to place some restrictions on the form of the mean and principal component curves. We choose to represent μ and f using a basis of spline functions (Silverman, 1985; Green & Silverman, 1994). Let $b(t)$ be a spline basis with dimension q. Let Θ and θ_μ be, respectively, a q by k matrix and a q-dimensional vector of spline coefficients. Then

$$\begin{aligned} \mu(t) &= b(t)^{\mathrm{T}}\theta_\mu, \\ f(t)^{\mathrm{T}} &= b(t)^{\mathrm{T}}\Theta. \end{aligned}$$

The resulting restricted model has the form

$$Y_i(t) = b(t)^{\mathrm{T}}\theta_\mu + b(t)^{\mathrm{T}}\Theta\alpha_i + \varepsilon_i(t), \quad i = 1, \ldots, N, \tag{5}$$

$$\varepsilon_i(t) \sim (0, \sigma^2), \quad \alpha_i \sim (0, D)$$

subject to

$$\Theta^{\mathrm{T}}\Theta = I, \quad \int b(t)^{\mathrm{T}} b(t)\,dt = 1, \quad \int\int b(t)^{\mathrm{T}} b(s)\,dt\,ds = 0. \tag{6}$$

The equations in (6) impose orthogonality constraints on the principal component curves. Note that, if one does not assume a special structure for the covariance matrix of the α_i's, Θ and Σ will be confounded. Thus we restrict the covariance matrix to be diagonal and denote it by D.

For each individual i, let $t_{i1}, t_{i2}, \ldots, t_{in_i}$ be the possibly different time points at which measure-

ments are available. Then

$$\begin{aligned} Y_i &= (Y_i(t_{i1}), \ldots, Y_i(t_{in_i}))^{\mathrm{T}}, \\ B_i &= (b(t_{i1}), \ldots, b(t_{in_i}))^{\mathrm{T}}. \end{aligned}$$

Note that B_i is the spline basis matrix for the ith individual. To approximate the orthogonality condition in (6) we choose $b(\cdot)$ so that $B^{\mathrm{T}}B = I$, where B is the basis matrix evaluated on a fine grid of time points. For instance, in the growth curve example the time interval was divided into 172 periods of $1/10$th of a year each.

The reduced rank model can then be written as

$$Y_i = B_i\theta_\mu + B_i\Theta\alpha_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{7}$$

$$\Theta^{\mathrm{T}}\Theta = I, \quad \varepsilon_i \sim (0, \sigma^2 I), \quad \alpha_i \sim (0, D).$$

Fitting this model involves estimating θ_μ, Θ, D and σ^2. A fitting procedure is presented in § 4. In practice q, the dimension of the spline, and k, the number of principal components, must also be chosen. Methods for making these choices are suggested in § 6. Note that the reduced rank model can also be interpreted as a mixed effects model with a rank constraint on the covariance matrix. This latter approach dates back to Anderson (1951).

Recall that, in the mixed effects model (2), γ_i is a random vector with unrestricted covariance matrix. Hence we can reparameterise γ_i as

$$[\Theta, \Theta^*]\begin{pmatrix} \alpha_i \\ \alpha_i^* \end{pmatrix}$$

where Θ and α_i are defined as in (7), Θ^* is a q by $q-k$ dimensional matrix which is orthogonal to Θ, and α_i^* is a random vector of length $q-k$ with a diagonal covariance matrix. As a result the mixed effects model can be written as

$$Y_i = B_i\theta_\mu + B_i\Theta\alpha_i + B_i\Theta^*\alpha_i^* + \varepsilon_i, \quad i = 1, \ldots, N. \tag{8}$$

Thus the reduced rank model is a submodel of the mixed effects model. In the reduced rank model the α_i^*'s are set to zero and no attempt is made to estimate the additional parameters, Θ^*. To fit k principal component curves using the mixed effects method Rice & Wu (2000) calculate the first k eigenvectors of the estimate for Γ; recall that Γ is the covariance matrix of the γ_i's. In other words, even though Θ^* is estimated in the mixed effects procedure it is never used. By employing the mixed effects method and then setting the α_i^*'s to zero one is simply fitting the reduced rank model using a different algorithm.

We call the likelihood obtained from the mixed effects fit, after setting the α_i^*'s to zero, the constrained mixed effects likelihood. Since the mixed effects and reduced rank methods can be considered as two different approaches to fitting the reduced rank model, the constrained mixed effects and reduced rank likelihoods can be meaningfully compared. For example, Table 1 provides the loglikelihoods up to a constant term for the three different fits to the growth curve data. The reduced rank likelihood must be at least as large as that of the constrained likelihood. However, note that the reduced rank likelihood is in fact strictly higher.

Number of knots	Loglikelihood	
	Constrained	Reduced rank
4	380.63	389.22
9	394.75	409.81
14	399.00	411.36

Table 1: Loglikelihoods for the fits in Fig. 3(a) through (f).

4. FITTING THE REDUCED RANK MODEL

4·1. *Preamble*

In a functional principal component analysis setting the primary goal is to estimate μ and f. A secondary goal is the prediction of the α_i's, which, when combined with the estimates of μ and f, give predictions for the individual curves. Since we are assuming a spline fit to the functions this is equivalent to estimating θ_μ and Θ and predicting the α_i's. Note that θ_μ, Θ, σ^2 and D are all unknown parameters. The elements of D give a measure of the variability explained by each principal component curve and σ^2 provides a measure of the variability left unexplained. To derive a fitting procedure we appeal to maximum likelihood and penalised least squares ideas which in this instance lead to the same algorithm.

4·2. *Maximum likelihood*

Assume that the α_i's and ε_i's are normally distributed. Then

$$Y_i \sim N(B_i\theta_\mu, \sigma^2 I + B_i\Theta D\Theta^{\mathrm{T}}B_i^{\mathrm{T}}) \qquad i = 1,\ldots,N, \tag{9}$$

and the observed likelihood for the joint distribution of the Y_i's is

$$\prod_{i=1}^{N} \frac{1}{(2\pi)^{n_i/2}|\sigma^2 I + B_i\Theta D\Theta^{\mathrm{T}}B_i^{\mathrm{T}}|^{1/2}} \exp\left\{-\frac{1}{2}(Y_i - B_i\theta_\mu)^{\mathrm{T}}(\sigma^2 I + B_i\Theta D\Theta^{\mathrm{T}}B_i^{\mathrm{T}})^{-1}(Y_i - B_i\theta_\mu)\right\}. \tag{10}$$

Unfortunately to maximise this likelihood over $\theta_\mu, \Theta, \sigma^2$ and D is a difficult non-convex optimisation problem. If the α_i's were observed the joint likelihood would simplify to

$$\prod_{i=1}^{N} \frac{1}{(2\pi)^{(n_i+k)/2}\sigma^{n_i}|D|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - B_i\theta_\mu - B_i\Theta\alpha_i)^{\mathrm{T}}(Y_i - B_i\theta_\mu - B_i\Theta\alpha_i) - \frac{1}{2}\alpha_i^{\mathrm{T}}D^{-1}\alpha_i\right\}. \tag{11}$$

This is a much easier expression to maximise which suggests treating the α_i's as missing data and employing the EM algorithm (Dempster, Laird & Rubin, 1977). Details of our optimization routine can be obtained from the web site www-rcf.usc.edu/∼gareth.

4·3. *Penalised least squares*

The same algorithm can be motivated using penalised least squares. With this approach one chooses θ_μ, Θ and the α_i's to minimise the sum of squared residuals between the data points and predicted values, subject to a penalty term on the α_i's, namely minimise

$$\sum_{i=1}^{N} \left\{ (Y_i - B_i\theta_\mu - B_i\Theta\alpha_i)^{\mathrm{T}}(Y_i - B_i\theta_\mu - B_i\Theta\alpha_i) + \sigma^2 \sum_{j=1}^{k} \frac{\alpha_{ij}^2}{D_{jj}} \right\}. \tag{12}$$

The algorithm is as follows. Fix the values of σ^2, D and the α_i's, and minimise (12) with respect to θ_μ and Θ, giving values identical to those from the M-step of the EM algorithm. Next minimise (12) with respect to the α_i's while holding all other parameters fixed. The values of the α_i's will be identical to those from the E-step of the EM algorithm. Finally, refit σ^2 and D using the standard sample variance estimates. If the same initial values are used, iterating these three steps until convergence will yield the same final estimates as the maximum likelihood procedure of § 4·2. Note that the coefficient in the penalty term is σ^2/D_{jj}. Since D_{jj} is the variance of the α_{ij}'s the terms with lower variance are penalised more heavily.

5. THE REDUCED RANK AND MIXED EFFECTS METHODS COMPARED

In § 3 we noted that the primary difference between the reduced rank and the constrained mixed effects methods lies in the fitting procedures. Thus it is legitimate to compare the two methods directly using likelihoods. To do this we ran two simulation studies. In order to make the simulated data more realistic and interpretable we based them on the growth curve data. In the first study the data were generated from the mean function and principal component curve corresponding to the reduced rank fit shown in Fig. 3(g) and (h). Forty-eight curve fragments were generated using the same time points as the growth curve data. The mixed effects and reduced rank procedures were fitted to ten such datasets using natural cubic splines with the correct knot selection. The generating curves for the second study were obtained just as in the first study except that splines with seven equally spaced knots were used. Sixteen curve fragments were generated using the time points from a randomly selected subset of the original forty-eight partial growth curves. The mixed effects and reduced rank procedures were fitted to ten such datasets again using cubic splines with the correct knot selection. The datasets in this simulation were more difficult to fit because of the smaller sample size and the higher dimensionality of the splines.

Figure 4(a) shows the estimates for the principal component from the mixed effects and reduced rank fits on a dataset from the first simulation study. The accuracy of the fit is typical of the mixed effects procedure. The reduced rank fit was superior to the mixed effects fit for all ten datasets. The ratio of the true variance to the estimated variance gives a measure of goodness of fit. Figure 4(b) shows a plot of the variance ratio versus the loglikelihood. The constrained mixed effects and reduced rank fits are represented, respectively, by squares and triangles, fits corresponding to the same dataset being joined up. This plot illustrates two key points. First, the variance ratios for the mixed effects fits are almost all greater than one, suggesting that the method tends to overfit the data; the reduced rank method performs much better in this respect. Secondly, the reduced rank procedure gives a higher likelihood on all ten datasets.

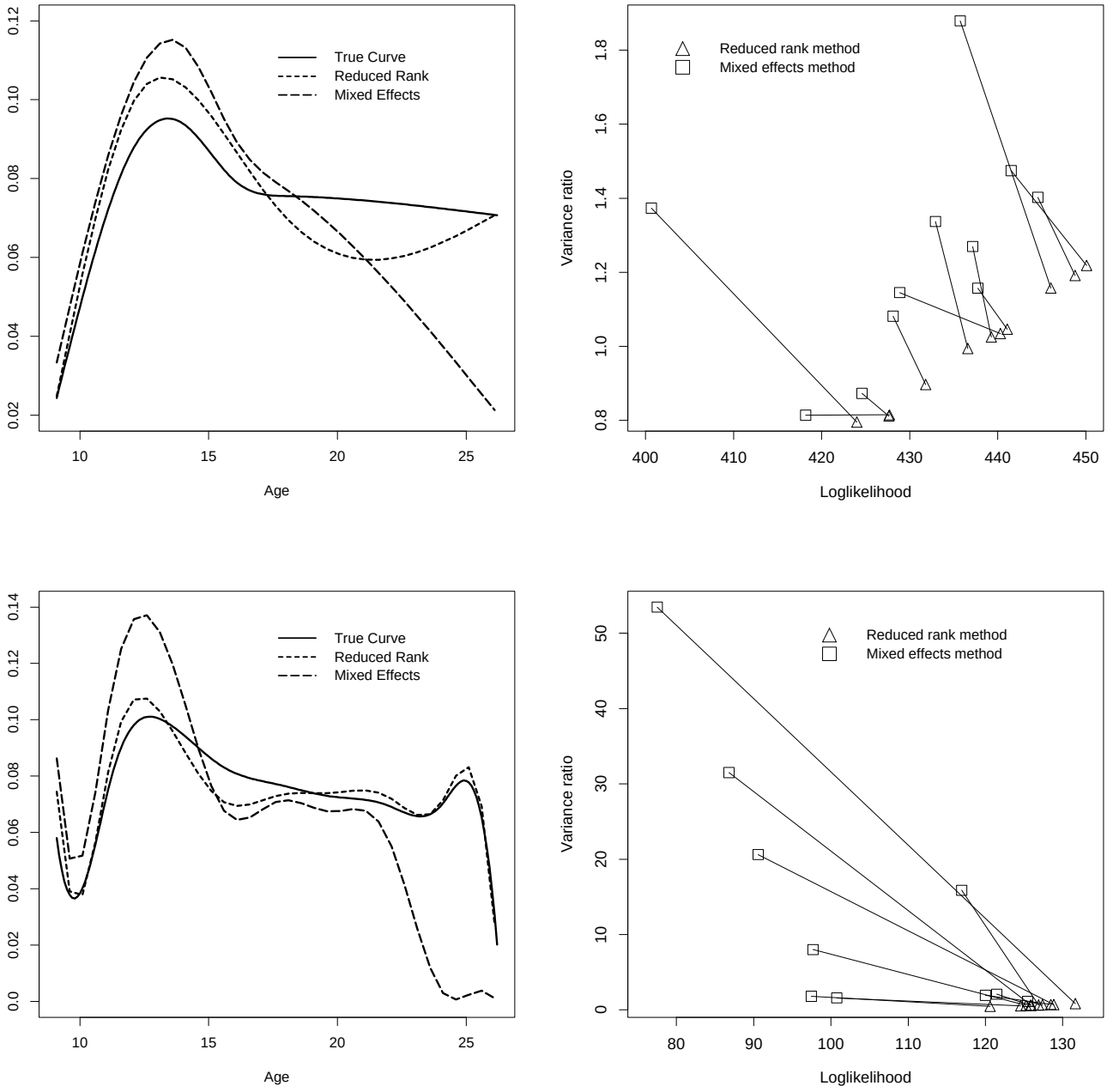

Figure 4: Results from two simulation studies. Figures (a) and (c) give the first principal component curve for two datasets from, respectively, the first and second simulation studies. Figures (b) and (d) give the corresponding plots of variance ratio versus loglikelihood for the datasets in each simulation study. Fits to the same dataset are joined.

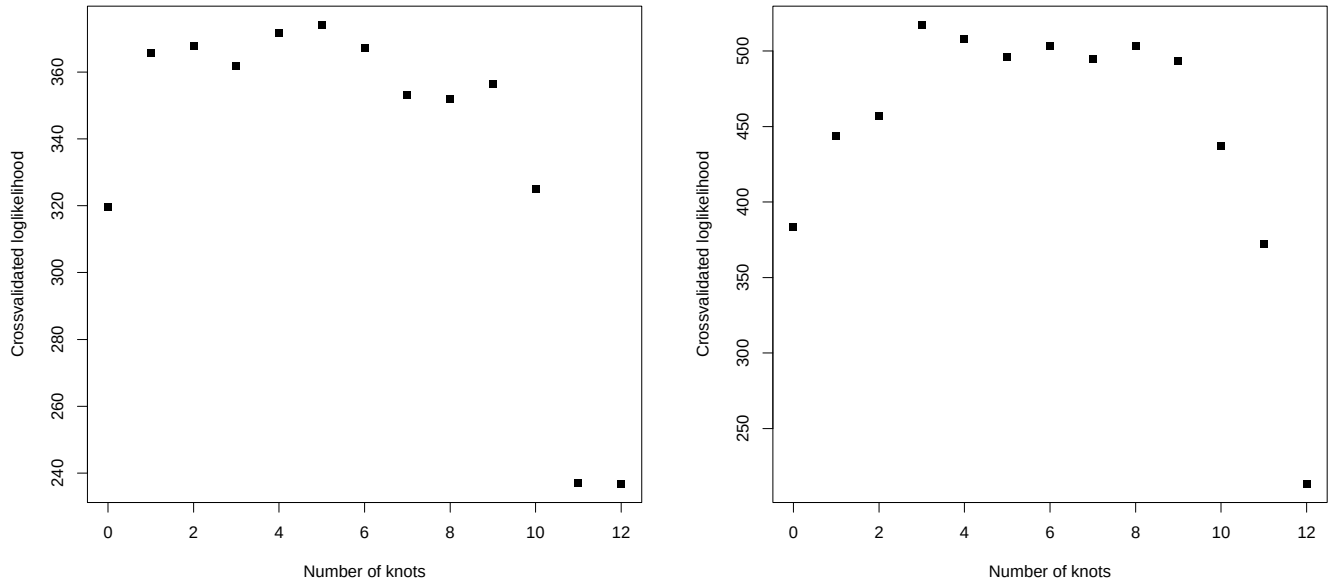

Figure 5: Crossvalidated loglikelihoods for (a) the growth curve data and (b) a dataset simulated from a spline with four knots.

Figures 4(c) and (d) give the corresponding plots for the second simulation study. Once again the reduced rank procedure has done a substantially better job at estimating the principal component. It is clear that the large number of parameters involved in estimating the full covariance matrix has had a deleterious effect on the mixed effects method fit. Correspondingly, the overfitting problem has drastically increased. The mixed effects method underestimates the variance by up to a factor of fifty. In addition, the more the variance is underestimated the worse the corresponding likelihood estimate becomes. Again, the reduced rank procedure consistently produces better variance estimates and higher likelihoods.

6. Model selection and inference

6·1. *Selection of the number of knots in the spline basis*

A natural approach is to calculate the crossvalidated loglikelihood for different numbers of knots and to select the number corresponding to the maximum. All examples in this section use ten-fold crossvalidation, which involves removing 10% of the curves as a test set, fitting the model to the remaining curves, calculating the loglikelihood on the test set, and then repeating the process nine more times. For the growth curve data, Fig. 5(a) shows crossvalidated loglikelihood estimates for models involving between zero and twelve evenly spaced knots. It appears that the optimal number of knots is between four and six, and we opted for the more parsimonious model with four knots.

To test the validity of this procedure we generated data from the model of the first simulation study in § 5 except that the random noise had a smaller standard deviation. Figure 5(b) shows that

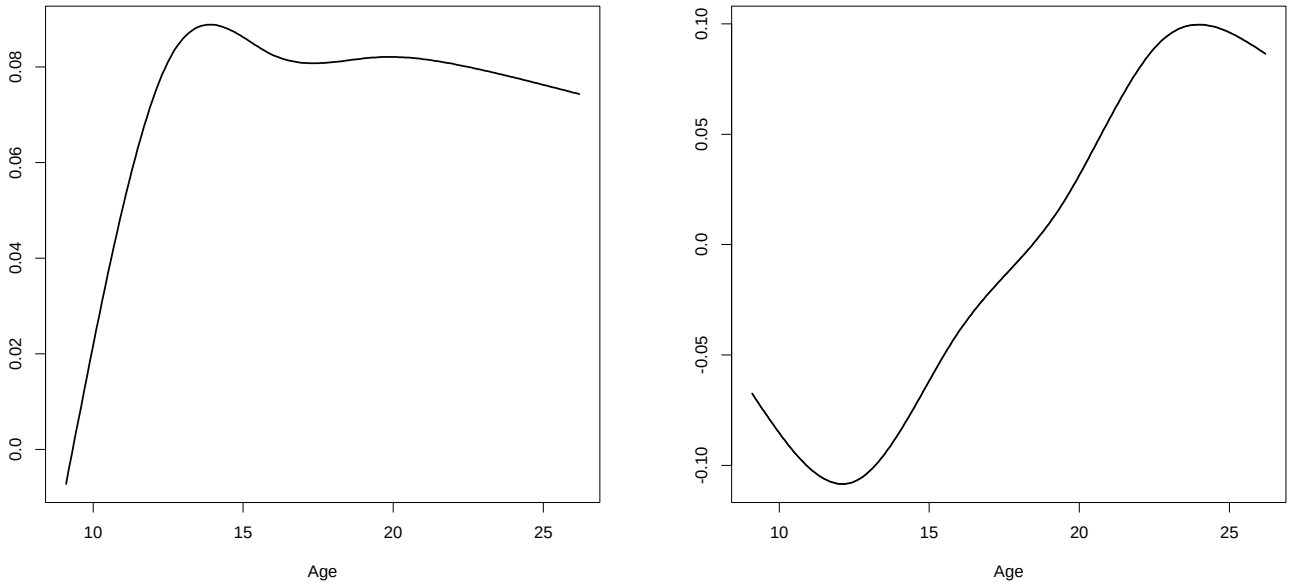

Figure 6: The principal component curves that result when the reduced rank method is fitted to the growth curve data with the optimal rank $k = 2$. Notice that the second principal component models differences in the slopes of individual curves.

the crossvalidated likelihood is maximised for three knots with the second largest likelihood corresponding to the correct value of four knots. The procedure seems to be selecting approximately the correct number of knots but this simulation illustrates that the plot should be treated as a guide rather than an absolute rule.

Crossvalidation is a computationally expensive procedure. Rice & Wu (2000) suggest using AIC and BIC which require fewer computations. In the datasets they examined AIC, BIC and crossvalidation all produced qualitatively comparable results.

6·2. *Selection of the rank, k*

With functional principal component analysis it is particularly important to identify the number of important principal components, k, because the fits of the different components are not independent. As examples involving the mixed effects procedure have shown, choosing to fit too many principal components can degrade the fit of them all. In this section we outline two alternative procedures for choosing the number of principal components. Note that this is equivalent to selecting the rank of the covariance matrix.

A natural first approach is to calculate the proportion of variability explained by each principal component. It is difficult to compute this quantity directly in functional principal component analysis. However, if σ^2 is close to zero and the curves are all measured at similar time points it can be shown that the proportion of the total variation in the α_i's associated with each component is a good approximation. Recall that D is the diagonal covariance matrix of the α_i's so the desired

proportion is simply

$$\frac{D_{jj}}{\text{trace}(D)}. \tag{13}$$

For the growth curve data, in view of the results of § 6·1., we fitted the reduced rank method using a cubic spline with four knots, which allows a choice of up to six principal components. The first principal component explains approximately 74% of the variability. The second principal component explains nearly all of the remaining variability, about 24%, and so may also be useful, the remaining components are unnecessary and should not be included in the model because they may cause over-fitting problems. Figures 6(a) and (b) show the principal components obtained when the reduced rank method is fitted with $k = 2$. The first component was discussed in § 2. Note that the second principal component captures differences in the slopes of individual curves; a positive weighting for this component would indicate a curve with a greater than average slope, and a negative weighting a curve with less than average slope.

A second procedure for estimating the number of principal components involves calculating the loglikelihood for the reduced rank method as k varies between 0 and 6. Provided that the fitting algorithm has converged to the global maximum, the loglikelihood is guaranteed to increase as k increases, but the increase should level off when the optimal rank is reached. A plot of the loglikelihood versus k for the growth curve data, not shown here, reveals a large jump in likelihood between $k = 0$ and $k = 1$ and that the plot has clearly levelled off after $k = 2$. To determine whether the jump between $k = 1$ and $k = 2$ is large enough to warrant using the second principal component, note that twice the difference in loglikelihoods is asymptotically χ^2_5, if truly $k = 1$ since the model with $k = 2$ involves fitting five extra parameters. Twice the observed difference in loglikelihoods is 19.28 yielding a p-value of 0.002. This suggests that the second principal component is significant. However, since this dataset is sparse one should use caution when invoking an asymptotic result.

To check the accuracy of the two procedures we tested them on the simulated dataset from §6.1, generated using a single principal component curve. It turned out that the first procedure, which calculates the proportion of variation explained, worked well; the first principal component explained about 96% of the variability. However, the second procedure, using the loglikelihood, was more ambiguous. The plot suggested that there could be anywhere from one to three principal components and when the χ^2 rule was applied, $k = 3$ was chosen. From our experience the first procedure appears to be more reliable.

6·3. *Confidence intervals*

The bootstrap can be used to produce pointwise confidence intervals for the overall mean function, the principal components and the individual curves. There are two obvious ways to bootstrap curve data. The first involves resampling the individual curves. The second involves resampling the estimated α_i's and residuals and generating new partial curves based on these values. The first method has the advantage of not requiring any parametric assumptions, while the second has the advantage that the bootstrap datasets have observations at the same time points as the original dataset. When the data are sparse, especially in the tails, the first procedure performs poorly, and we therefore present results using the second procedure on the growth curve data.

We generated one-hundred bootstrap datasets and fitted the reduced rank method with $k = 2$ to

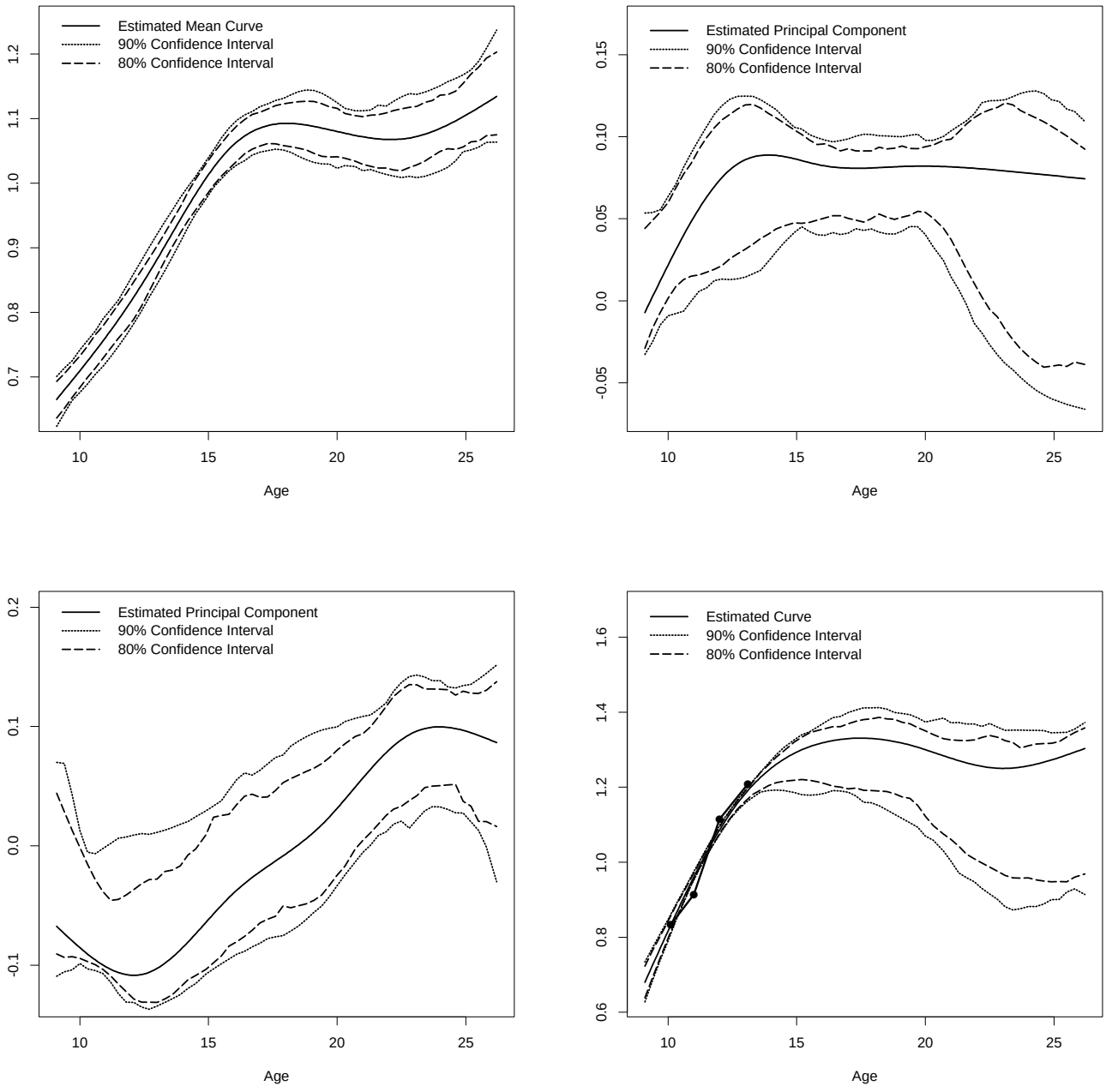

Figure 7: 80% and 90% pointwise confidence intervals for the mean function (a), both principal components (b) and (c) and an individual curve (d), for the growth curve data. The observed data for the individual in (d) is shown by the four circles.

each. Using the bootstrap percentile method (Efron & Tibshirani, 1993) we produced pointwise confidence intervals by taking the $\alpha/2$ and $1-\alpha/2$ quantile at each time point. Figure 7 gives 80% and 90% confidence intervals for the mean function, the principal components and an individual curve for the growth curve data. Despite the sparsity of the data, the intervals for the mean function are relatively narrow with some widening in the right tail where there were few observations. The confidence intervals for the first principal component are much wider, particularly in the right tail. The large dip in the confidence band in this region occurs because approximately 20% of the bootstrap principal component curves exhibited an inverted U shape. There appear to be two distinctly different possible shapes for this component. Interestingly, given the variability of the first component, the intervals for the second component follow the general shape of the estimated curve quite tightly. In Fig. 7(d) the four circles show the observed data values for one of the forty-eight individuals. As one would expect the intervals are very narrow near the observed points and fan out as one extrapolates.

7. Comparison of the reduced rank method and classical principal components

We begin by considering the linear model

$$X_i = \theta_\mu + \Theta\alpha_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{14}$$

$$\varepsilon_i \sim N(0, \Sigma), \quad \alpha_i \sim N(0, D),$$

where the X_i are q-dimensional data vectors and Θ is an orthogonal matrix. The solutions to factor analysis and standard principal components can be derived from this model. If Σ is diagonal, fitting (14) via maximum likelihood yields the factor analysis solution. If Σ is further restricted to have the form $\sigma^2 I$, then the limit as σ^2 approaches zero of the maximum likelihood estimates gives the classical principal components solution. Taking this limit is equivalent to minimising

$$\sum_{i=1}^{N} ||X_i - \theta_\mu - \Theta\alpha_i||^2. \tag{15}$$

In this context, the columns of Θ represent the principal components and the α_i's are weightings for the components. Recall from § 3 that the reduced rank model is

$$Y_i = B_i\theta_\mu + B_i\Theta\alpha_i + \varepsilon_i, \qquad \mathrm{cov}(\varepsilon_i) = \sigma^2 I. \tag{16}$$

If the covariance structure of the ε_i's were relaxed to be an arbitrary diagonal matrix then the reduced rank model would become a generalisation of the factor analysis model. However, we will not pursue this point further. Instead we concentrate on generalizations of principal components. Referring to (12), one sees that if, in analogy with classical principal component analysis, σ^2 is sent to zero in (16) then the procedure for fitting the reduced rank model simply minimises

$$\sum_{i=1}^{N} ||Y_i - B_i\theta_\mu - B_i\Theta\alpha_i||^2. \tag{17}$$

Let $\hat{\gamma}_i = (B_i^{\mathrm{T}} B_i)^{-1} B_i^{\mathrm{T}} Y_i$. Note that $\hat{\gamma}_i$ is the least squares estimates of the spline coefficients for the ith curve. Then one can transform (17) into

$$\begin{aligned} &\sum_{i=1}^{N} ||Y_i - B_i\hat{\gamma}_i||^2 + \sum_{i=1}^{N} ||B_i\hat{\gamma}_i - B_i\theta_\mu - B_i\Theta\alpha_i||^2 \\ &= C(Y) + \sum_{i=1}^{N} ||\hat{\gamma}_i - \theta_\mu - \Theta\alpha_i||^2_{B_i^{\mathrm{T}}B_i} \end{aligned} \tag{18}$$

Therefore, since $C(Y)$ is a constant with respect to the parameters, to minimise (17) it is sufficient to minimise

$$\sum_{i=1}^{N} ||\hat{\gamma}_i - \theta_\mu - \Theta\alpha_i||^2_{B_i^{\mathrm{T}}B_i}. \tag{19}$$

Note that if B_i is not full column rank then the indeterminate parts of $\hat{\gamma}_i$ are given weight zero by the metric $B_i^{\mathrm{T}} B_i$. Suppose that all curves are measured at the same set of time points. Then $B_i = B$ is the common spline basis matrix. Without loss of generality one may assume that $B^{\mathrm{T}} B = I$, and so minimising (19) is equivalent to performing standard principal components on the spline coefficients. Note that this is the approach taken by the direct method of § 1·2.

Standard principal components takes q dimensional data and finds the $k\,(<q)$-dimensional plane that minimises the squared Euclidean distance to each point. As seen above, when all curves are measured at the same time points the reduced rank method also finds the best fitting plane using the Euclidean metric. It is apparent from (19) that when the curves are not sampled at identical time points the reduced rank procedure still identifies the best fitting plane. However, the distance between the plane and each data point is measured relative to the metric $B_i^{\mathrm{T}} B_i$ which may be different for each individual. Taking this view of the reduced rank method as a generalisation of classical principal component analysis provides some useful geometric intuition. One of the difficulties with visualising the functional principal components problem is that the curves are points in an infinite-dimensional space. Equation (19) shows that one can visualise the data as lying in a single q-dimensional space at the expense of assigning each point a unique distance metric. Figure 8 provides a pictorial representation of such a non-Euclidean principal components fit.

ACKNOWLEDGMENTS

The authors would like to thank the Editor and referee for many constructive suggestions. Trevor Hastie was partially supported by grants from the National Science Foundation and the National Institutes of Health.

A APPENDIX

In this section we provide details of the Reduced Rank fitting algorithm. Steps 1 and 2 make up the M-step and Step 3 makes up the E-step.

1. Given current estimates for α_i, θ_μ and Θ we estimate σ^2 and D as

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{\sum n_i}\sum_{i=1}^{N} E[\varepsilon_i^T \varepsilon_i | Y_i] \\ &= \frac{1}{\sum n_i}\sum_{i=1}^{N} E[(Y_i - B_i\hat{\theta}_\mu - B_i\hat{\Theta}\alpha_i)^T (Y_i - B_i\hat{\theta}_\mu - B_i\hat{\Theta}\alpha_i) | Y_i] \\ &= \frac{1}{\sum n_i}\sum_{i=1}^{N} \Big((Y_i - B_i\hat{\theta}_\mu - B_i\hat{\Theta}\hat{\alpha}_i)^T (Y_i - B_i\hat{\theta}_\mu - B_i\hat{\Theta}\hat{\alpha}_i) \\ &\quad + \text{trace}[B_i\hat{\Theta}(\hat{D}^{-1} + \hat{\Theta}^T B_i^T B_i \hat{\Theta}/\hat{\sigma}^2)^{-1}\hat{\Theta}^T B_i^T] \Big)\end{aligned} \tag{20}$$

$$\hat{D}_{jj} = \frac{1}{N}\sum_{i=1}^{N} E[\alpha_{ij}^2 | Y_i] = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{\alpha}_{ij}^2 + (\hat{D}^{-1} + \hat{\Theta}^T B_i^T B_i \hat{\Theta}/\hat{\sigma}^2)_{jj}^{-1} \right) \tag{21}$$

Equations (20) and (21) derive from the facts that

$$E(X^2|Y) = (E(X|Y))^2 + Var(X|Y) \tag{22}$$

and

$$\alpha_i | Y_i \sim N\left((\sigma^2 D^{-1} + \Theta^T B_i^T B_i \Theta)^{-1} \Theta B_i^T (Y_i - B_i\theta_\mu), (D^{-1} + \Theta^T B_i^T B_i \Theta/\sigma^2)^{-1} \right) \tag{23}$$

2. Given current estimates for σ^2, D and α_i we estimate Θ and θ_μ by minimizing

$$\sum_{i=1}^{N} \left[(Y_i - B_i\hat{\theta}_\mu - B_i\hat{\Theta}\hat{\alpha}_i)^T (Y_i - B_i\hat{\theta}_\mu - B_i\hat{\Theta}\hat{\alpha}_i) \right] \tag{24}$$

Minimizing (24) involves a second iterative procedure, in which each column of Θ is estimated separately holding all other columns fixed. First notice that

$$\begin{aligned}&\sum_{i=1}^{N} ||Y_i - B_i\theta_\mu - B_i\Theta\alpha_i||^2 \\ = &\sum_{i=1}^{N} ||(Y_i - B_i\Theta\alpha_i) - B_i\theta_\mu||^2\end{aligned}$$

so the estimate for θ_μ is

$$\hat{\theta}_\mu = \left(\sum_{i=1}^{N} B_i^T B_i \right)^{-1} \sum_{i=1}^{N} B_i^T (Y_i - B_i\hat{\Theta}\hat{\alpha}_i) \tag{25}$$

To estimate the columns of Θ we note that

$$\begin{aligned} & \sum_{i=1}^{N} ||Y_i - B_i\theta_\mu - B_i\Theta\alpha_i||^2 \\ = & \sum_{i=1}^{N} ||(Y_i - B_i\theta_\mu - \alpha_{i2}B_i\theta_2 - \cdots - \alpha_{ik}B_i\theta_k) - \alpha_{i1}B_i\theta_1||^2 \end{aligned}$$

Therefore the estimate for θ_1 is

$$\hat{\theta}_1 = \left(\sum_{i=1}^{N} \widehat{\alpha_{i1}^2} B_i^T B_i\right)^{-1} \sum_{i=1}^{N} B_i^T(\hat{\alpha}_{i1}(Y_i - B_i\hat{\theta}_\mu) - \widehat{\alpha_{i1}\alpha_{i2}} B_i\hat{\theta}_2 - \cdots - \widehat{\alpha_{i1}\alpha_{ik}} B_i\hat{\theta}_k) \tag{26}$$

We repeat this procedure for each column of Θ and iterate until there is no further change.

3. The E-step consists of predicting α_i and $\alpha_i\alpha_i^T$.

$$\hat{\alpha}_i = E(\alpha_i|Y_i, \hat{\theta}_\mu, \hat{\Theta}, \hat{\sigma}^2, \hat{D}) = (\hat{\sigma^2}\hat{D}^{-1} + \hat{\Theta}^T B_i^T B_i\hat{\Theta})^{-1}\hat{\Theta}^T B_i^T(Y_i - B_i\hat{\theta}_\mu) \tag{27}$$

$$\widehat{\alpha_i\alpha_i^T} = E(\alpha_i\alpha_i^T|Y_i, \hat{\theta}_\mu, \hat{\Theta}, \hat{\sigma}^2, \hat{D}) = \hat{\alpha}_i\hat{\alpha}_i^T + (\hat{D}^{-1} + \hat{\Theta}^T B_i^T B_i\hat{\Theta}/\hat{\sigma}^2)^{-1} \tag{28}$$

Both predictions make use of equations (22) and (23).

4. We then return to Step 1 and repeat until we reach convergence.

5. The matrix Θ produced by this procedure will not be orthogonal. We orthogonalize it by producing the reduced rank estimate for Γ,

$$\hat{\Gamma} = \hat{\Theta}\hat{D}\hat{\Theta}^T \tag{29}$$

and setting Θ equal to the first k eigenvectors of $\hat{\Gamma}$.

REFERENCES

ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22**, 327–51.

BACHRACH, L., HASTIE, T., WANG, M., NARASIMHAN, B. & MARCUS, R. (1999). Bone mineral acquisition in healthy asian, hispanic, black and caucasian youth; a longitudinal study. *Clinical Endocrinol. Metabol.* **84**, 4702–12.

BESSE, C. & CARDOT, H. (1996). Spline approximation of the prediction of a functional autoregressive process of order 1 (in French). *Can. Statist.* **24**, 467–87.

BESSE, C., CARDOT, H. & FERRATY, F. (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data. *Comp. Statist. Data Anal.* **24**, 255–70.

BRUMBACK, B. & RICE, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Am. Statist. Assoc.* **93**, 961–76.

BUCKHEIT, J., OLSHEN, R., BLOUCH, K. & MYERS, B. (1997). Modeling of progressive glomerular injury in humans with lupus nephritis. *Am. Physiol.-Renal Physiol.* **42**, F158–69.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *R. Statist. Soc. B* **39**, 1–38.

EFRON, B. & TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. London : Chapman and Hall.

GASSER, T., MULLER, H.-G., KOHLER, W., MOLINARI, L. & PRADER, A. (1984). Nonparametric regression analysis of growth curves. *Ann. Statist.* **12**, 210–29.

GREEN, P. & SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models : A roughness penalty approach*. London : Chapman and Hall.

HENDERSON, C. R. (1950). Estimation of genetic parameters (abstract). *Ann. Math. Statist.* **21**, 309–10.

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

RAMSAY, J. & SILVERMAN, B. (1997). *Functional Data Analysis.* New York : Springer.

RICE, J. & SILVERMAN, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *R. Statist. Soc. B* **53**, 233–43.

RICE, J. & WU, C. (2000). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics (To appear)* .

SHI, M., WEISS, R. & TAYLOR, J. (1996). An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics* **45**, 151–64.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *R. Statist. Soc. B* **47**, 1–52.

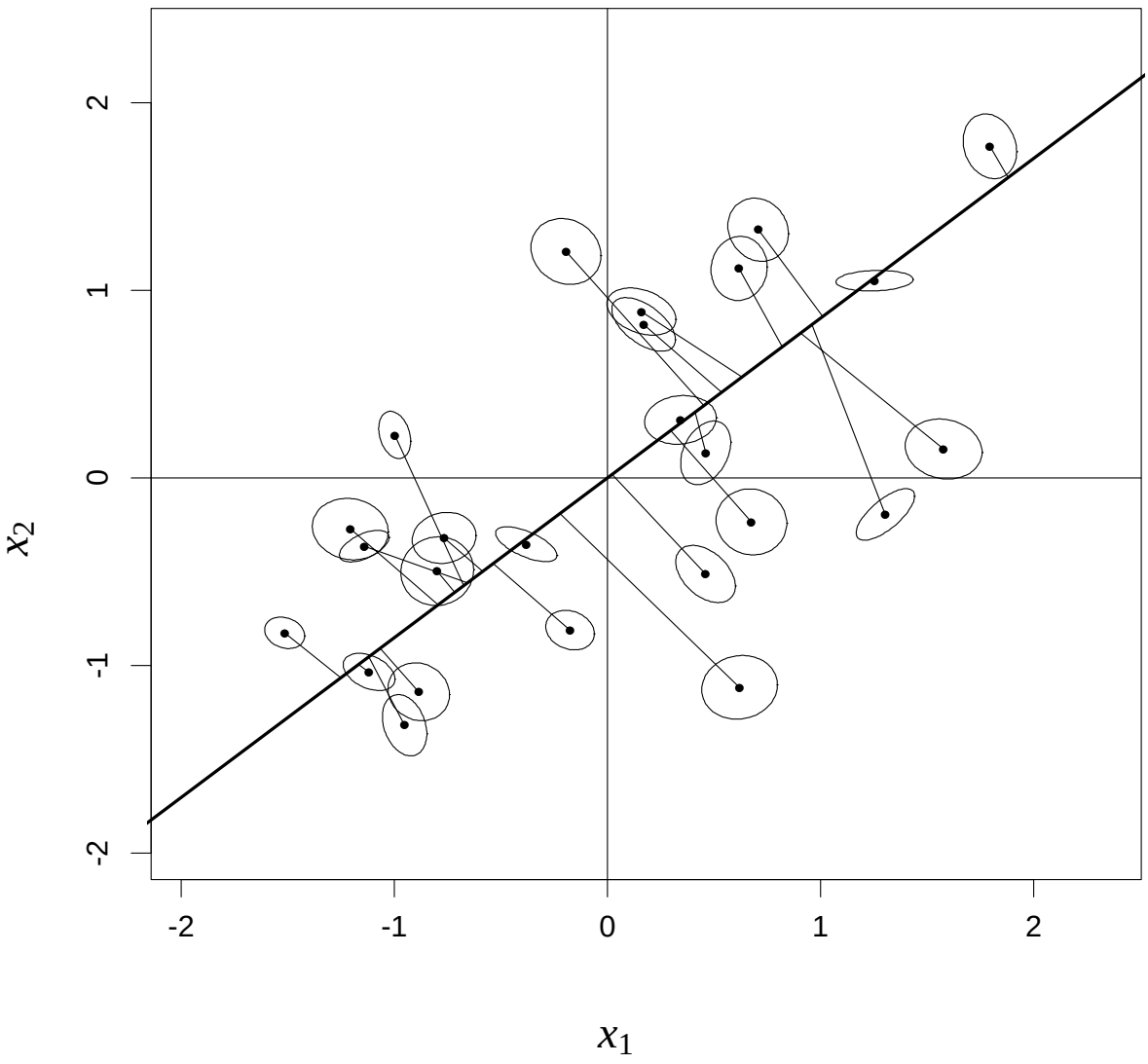

Figure 8: A depiction of non-Euclidean principal components in R^2. Each point has an associated metric Σ_i with which to measure distance. We seek the line that minimises the sum of squared distances to the points.