Fitting Subject-specific Curves to Grouped Longitudinal Data

Djeundje, Viani
Heriot-Watt University, Department of Actuarial Mathematics & Statistics
Edinburgh, EH14 4AS, UK
E-mail: vad5@hw.ac.uk

Currie, Iain
Heriot-Watt University, Department of Actuarial Mathematics & Statistics
Edinburgh, EH14 4AS, UK
E-mail: I.D.Currie@hw.ac.uk

1 A motivating example

In longitudinal studies, a response variable measured over time, may vary across subgroups. A typical example is illustrated by the first three panels of Figure 1; these display simulated growth data (based on the model in Durban *et al.* 2005) of 197 children who suffer from acute lymphoblastic leukaemia, and have received three different treatments. In some cases, a parametric mixed model is sufficient to summarize this type of data. However in Figure 1, a parametric approach does not seem appropriate, so smoothing is incorporated into the modelling process in order to extract the correct patterns from the data. In this setting, the mixed model representation of truncated polynomials is widely used, with a well known covariance structure for the random effects. In this paper, we outline this approach, demonstrate its limitations, and describe a more appropriate approach via penalty arguments.

2 The standard model

We consider n subjects classified into m groups and we denote by $g(i)$ the group to which subject i belongs, by t_{ij} the time of the jth observation on subject i, and by y_{ij} the value of the response variable on subject i at time t_{ij}. We suppose that the data are entered in group order and denote by $\boldsymbol{y}_i$ and $\boldsymbol{t}_i$ the response and time vectors for subject i. Our aim is to extract both the group and subject effects from these data in a smooth fashion. A convenient model for this purpose can be written as

$$y_{ij} = S_{g(i)}(t_{ij}) + \breve{S}_i(t_{ij}) + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $S_{g(i)}(\cdot)$ quantifies the group effect to which subject i belongs, and $\breve{S}_i(\cdot)$ measures the departure of the ith subject from its group effect. A common approach consists of expressing these functions in terms of truncated lines, ie,

$$S_{g(i)}(t) = a_{g(i),0} + a_{g(i),1}t + \sum_{k=1}^{q} u_{g(i),k}(t-\tau_k)_+, \quad \breve{S}_i(t) = \breve{a}_{i0} + \breve{a}_{i1}t + \sum_{l=1}^{\breve{q}} \breve{u}_{il}(t-\breve{\tau}_l)_+, \tag{2}$$

where $x_+ = \max\{x, 0\}$, and $\boldsymbol{\tau} = \{\tau_1, ..., \tau_q\}$ and $\breve{\boldsymbol{\tau}} = \{\breve{\tau}_1, ..., \breve{\tau}_{\breve{q}}\}$ are sets of internal knots at the group and subject levels respectively. If $\boldsymbol{a}_{g(i)} = (a_{g(i),0}, a_{g(i),1})'$, $\boldsymbol{u}_{g(i)} = (u_{g(i),1}, ..., u_{g(i),q})'$, $\breve{\boldsymbol{a}}_i = (\breve{a}_{i0}, \breve{a}_{i1})'$ and $\breve{\boldsymbol{u}}_i = (\breve{u}_{i1}, ..., \breve{u}_{i\breve{q}})'$, then model (1) can be expressed in matrix form as

$$\boldsymbol{y}_i = S_{g(i)}(\boldsymbol{t}_i) + \breve{S}_i(\boldsymbol{t}_i) + \boldsymbol{\varepsilon}_i, \; i = 1, \ldots, n, \tag{3}$$

where

$$S_{g(i)}(\boldsymbol{t}_i) = \boldsymbol{X}_i\boldsymbol{a}_{g(i)} + \boldsymbol{T}_i\boldsymbol{u}_{g(i)}, \quad \breve{S}_i(\boldsymbol{t}_i) = \breve{\boldsymbol{X}}_i\breve{\boldsymbol{a}}_i + \breve{\boldsymbol{T}}_i\breve{\boldsymbol{u}}_i, \tag{4}$$

and $\boldsymbol{\varepsilon}_i$ is the error vector for subject i; here $\boldsymbol{X}_i = \breve{\boldsymbol{X}}_i = [\mathbf{1} : \boldsymbol{t}_i]$, and $\boldsymbol{T}_i$ and $\breve{\boldsymbol{T}}_i$ are made up of truncated lines for subject i at the group and subject level respectively. Two issues need to be addressed: *smoothness* and *identifiability*. A standard approach uses a rich set of knots and then deals simultaneously with smoothness and identifiability by applying the following normal constraints:

$$\boldsymbol{u}_{g(i)} \sim \mathcal{N}\left(\mathbf{0}, \sigma_P^2 \boldsymbol{I}_q\right), \quad \breve{\boldsymbol{a}}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \quad \breve{\boldsymbol{u}}_i \sim \mathcal{N}\left(\mathbf{0}, \sigma_S^2 \boldsymbol{I}_{\breve{q}}\right), \tag{5}$$

where $\boldsymbol{\Sigma}$ is a 2×2 symmetric positive definite matrix, and $\boldsymbol{I}_q$ and $\boldsymbol{I}_{\breve{q}}$ indicate identity matrices of sizes q and $\breve{q}$ respectively; see Coull *et al.* (2001), Ruppert *et al.* (2003) and Durban *et al.* (2005). We refer to (5) as the *standard covariance* assumption and to the model defined by (4) and (5) as the *standard model*. One advantage of this model is that it can be fitted to data with the `lme` function in the `R` package `nlme`, as described by Durban *et al.* (2005). However, the investigation of this approach in terms of its ability to separate appropriately the two inter-connected effects (ie, the group and subject effects) has received little attention. In a recent paper, Djeundje & Currie (2010) illustrated some of its problems. In this section we describe these problems and explain why the approach will fail, especially as far as the identifiability of the group and subject effects is concerned.

First, we consider the growth data introduced earlier; this data set contains 197 children and the number of observations per child varies from 1 to 21; for this reason, we follow Durban *et al.* (2005) and use $q = 40$ and $\breve{q} = 10$ knots at the treatment and subject levels. To illustrate our point, we consider two scenarios:

Scenario 1: We use the representation (4); we refer to this representation as the *forward basis*.

Scenario 2: We consider the representation (4), but replace the basis functions $(t - \tau_k)_+$ and $(t - \breve{\tau}_l)_+$ by $(\tau_k - t)_+$ and $(\breve{\tau}_l - t)_+$ respectively; we refer to this representation as the *backward basis*.

The fitted children (subject) means from both these scenarios (obtained by adding the fitted children effects to their group effects) are nearly identical, and sit appropriately on the data. It is easy to suppose that this goodness of fit at the global level implies that the fitted group and subject effects are appropriately extracted. However, the lower right panel in Figure 1 illustrates a fitted group effect under these two scenarios; we see that this fitted group effect together with the corresponding confidence bands depends on whether the forward or backward basis is used.

Second, we perform a simulation exercise to clarify this point further. For this purpose we consider $n = 60$ subjects divided into $m = 3$ groups of sizes $n_1 = 20$, $n_2 = 15$ and $n_3 = 25$. We consider the special case of *balanced data* with $M = 25$ observations made on each subject at equally-spaced time points on $[0, 1]$; we denote this common time vector by $\boldsymbol{t}$. We generate the response data for a single simulation as follows:

- We define the true group effects by the following quadratic functions: $S_k(t) = (t - \frac{k}{4})^2,\ k = 1, 2, 3.$
- We define the true subject curves by $\breve{S}_i(t) = A_i \times \sin(\phi_i t + \varphi_i)$, where $A_i \sim \mathcal{N}(0, \sigma_A^2)$ and $\phi_i, \varphi_i \sim \mathcal{U}[0, \theta]$, $i = 1, ..., n$. The results presented in this paper use $(\sigma_A^2, \theta) = (0.25^2, 2\pi)$, but we note that large values of θ increase the flexibility of these subject curves.
- Finally, we simulate response data according to model (1), with σ^2 set to 0.1^2. An illustration of such simulated data for group 1 is shown in the upper left panel of Figure 2.

We want to evaluate the standard model in terms of its ability to recover the true underlying group curves as well as the behaviour of the related confidence bands. The upper right panel of Figure 2 displays the fitted group effect corresponding to the data in the left panel. Specifically, this graphic shows for the effect of group 1: (i) the true effect (dashed black line), (ii) the fitted effect

using the forward basis (red), (iii) the fitted effect using the backward basis (green), and (iv) the fitted effect using the *penalty* approach (blue, and described in the next section). The bias in the fitted effect and the associated confidence bands arising from the standard model (with either the forward or the backward basis) is worrying.

We will now evaluate this approach asymptotically over a set of simulations as follows:

- We perform and store $N = 100$ simulations.
- For each simulation r, $1 \leq r \leq N$, we
  - fit the standard model and estimate the group effect $S_k(\boldsymbol{t})$ as $\hat{S}_k^{(r)}(\boldsymbol{t})$, $k = 1, 2, 3$,
  - compute the standard deviation $SD_k^{(r)}(\boldsymbol{t})$ about $\hat{S}_k^{(r)}(\boldsymbol{t})$, $k = 1, 2, 3$,
  - compute the average mean square errors as $MSE^{(r)} = \frac{\sum_{k=1}^{m} \|\hat{S}_k^{(r)}(\boldsymbol{t}) - S_k(\boldsymbol{t})\|^2}{mM}$, $m = 3$, $M = 25$.
- Finally, we compute the mean standard deviation as $SD_k = \frac{\sum_{r=1}^{N} SD_k^{(r)}(\boldsymbol{t})}{N}$, $k = 1, 2, 3$.

In the lower left panel of Figure 2, the red and green dots show the $MSE^{(r)}$ obtained under the forward and backward bases respectively, while the blue dots refer to the *penalty* approach (presented in the next section). The lower right panel shows the mean standard deviation SD_k with the same colour specification (each group is denoted by a different line style). We note that there is a close connection between the upper and lower right panels: the upper panel shows the widening fan effect of the confidence bands (for group one) arising from fitting the standard model (with the forward or backward basis) to a single simulated sample. This widening suggests that the standard deviation will increase or decrease (depending on the basis orientation). The lower panel does indeed confirm this widening (increasing or decreasing) under the standard model.

We have conducted an extensive investigation into the effect of different values of the parameters $(\sigma_A^2, \theta, \sigma^2)$. One important finding was that the discrepancy between the fitted effects and the true underlying effects with the standard model increases as the underlying subject curves become more flexible, ie, as θ increases; with small θ the subject curves are close to linear and the bias is relatively small.

The reason for the bias with the standard model is the covariance specification (5). In (5), the parameters σ_P^2 and σ_S^2 determine the *smoothness* of the group and subject effects respectively; the intention is that their *identification* be controlled by the covariance component $\boldsymbol{\Sigma}$. However, $\boldsymbol{\Sigma}$ can only separate the linear components of these two effects. There are two cases. First, if the subject effects are not too flexible, (ie, close to linear) then the variance parameter σ_S^2 will tend to be small and the standard model can successfully recover the two effects. Second, if the subject effects are sufficiently non-linear (as in the growth data example or in the simulated data in the upper left panel of Figure 2), then the non-linear components $\boldsymbol{T}_i\boldsymbol{u}_{g(i)}$ and $\breve{\boldsymbol{T}}_i\breve{\boldsymbol{u}}_i$ in (4) become substantial; there is nothing in the standard covariance (5) that enables the appropriate separation of the two effects. In the next section, we derive a more appropriate covariance structure via a penalty argument; this provides a solution to the problem.

3 Penalty approach

From now on, we express the group and subject effects in terms of B-splines, and we smooth using the P-spline system of Eilers & Marx (1996). A more general approach which includes both B-splines and/or truncated polynomials is described in Djeundje & Currie (2010). In terms of B-spline bases, the components of model (1) can be expressed in matrix form as

$$S_{g(i)}(\boldsymbol{t}_i) = \boldsymbol{B}_i\boldsymbol{\alpha}_{g(i)}, \quad \breve{S}_i(\boldsymbol{t}_i) = \breve{\boldsymbol{B}}_i\breve{\boldsymbol{\alpha}}_i, \tag{6}$$

where $\boldsymbol{B}_i$ and $\breve{\boldsymbol{B}}_i$ are B-spline bases for subject i at the group and subject levels, and $\boldsymbol{\alpha}_{g(i)}$ and $\breve{\boldsymbol{\alpha}}_i$ are regression coefficients. Following Eilers & Marx (1996), we achieve *smoothness* by penalizing the roughness of the B-spline coefficients at both levels, and for *identifiability*, we follow Djeundje & Currie (2010) and shrink the subjects' B-spline coefficients towards 0. This yields the following constraints:

$$\|\boldsymbol{\Delta}\boldsymbol{\alpha}_{g(i)}\|^2 < \rho, \quad \|\breve{\boldsymbol{\Delta}}\breve{\boldsymbol{\alpha}}_i\|^2 < \breve{\rho}_1, \qquad \|\breve{\boldsymbol{\alpha}}_i\|^2 < \breve{\rho}_2, \tag{7}$$

for some well chosen constants $(\rho, \breve{\rho}_1, \breve{\rho}_2)$; here and below, $\boldsymbol{\Delta}$ and $\breve{\boldsymbol{\Delta}}$ represent second order difference operators of appropriate size. Using Lagrange arguments, the penalized residual sum of squares, PRSS, of (1) & (6) under the constraints (7) can be expressed as

$$\texttt{PRSS} = \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\alpha}_{g(i)} - \breve{\boldsymbol{B}}_i\breve{\boldsymbol{\alpha}}_i\|^2 + \sum_{k=1}^{m} \boldsymbol{\alpha}_k'\boldsymbol{P}_k\boldsymbol{\alpha}_k + \sum_{i=1}^{n} \breve{\boldsymbol{\alpha}}_i'\breve{\boldsymbol{P}}_i\breve{\boldsymbol{\alpha}}_i, \tag{8}$$

where

$$\boldsymbol{P}_k = \lambda\boldsymbol{\Delta}'\boldsymbol{\Delta}, \quad \breve{\boldsymbol{P}}_i = \breve{\lambda}\breve{\boldsymbol{\Delta}}'\breve{\boldsymbol{\Delta}} + \breve{\delta}\boldsymbol{I}_{\breve{c}} \tag{9}$$

represent the penalty matrices at the group and subject levels respectively, and $\breve{c}$ is the number of columns in $\breve{\boldsymbol{B}}_i$. In these expressions, λ and $\breve{\lambda}$ are the smoothing parameters at the group and subject levels, while $\breve{\delta}$ is the shrinkage/identifiability parameter; these three parameters play (inversely) the equivalent role as ρ, $\breve{\rho}_1$ and $\breve{\rho}_2$ in (7). We can now fit the model in the least squares sense by minimizing the PRSS with respect to the regression coefficients, and use criteria like AIC, BIC or GCV to estimate the smoothing/identifiability parameters, as described in Djeundje & Currie (2010). Nowadays, smooth models are often expressed as mixed models, since estimates of the variance/smoothing parameters obtained from the mixed model representation and restricted likelihood tend to behave well (Reiss and Ogdon, 2009). Further, our simulation study displays a mixed model structure. Using the singular value decomposition of the penalty component $\boldsymbol{\Delta}'\boldsymbol{\Delta}$, we re-parameterize the PRSS as

$$\texttt{PRSS} = \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_{g(i)} - \boldsymbol{Z}_i\boldsymbol{b}_{g(i)} - \breve{\boldsymbol{B}}_i\breve{\boldsymbol{\alpha}}_i\|^2 + \lambda\sum_{k=1}^{m} \boldsymbol{b}_k'\boldsymbol{b}_k + \sum_{i=1}^{n} \breve{\boldsymbol{\alpha}}_i\breve{\boldsymbol{P}}_i\breve{\boldsymbol{\alpha}}_i, \tag{10}$$

with matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ defined as

$$\boldsymbol{X}_i = [\boldsymbol{1} : \boldsymbol{t}_i], \quad \boldsymbol{Z}_i = \boldsymbol{B}_i\boldsymbol{U}_i\boldsymbol{\Lambda}_i^{-1/2}; \tag{11}$$

here $\boldsymbol{\Lambda}_i$ is the diagonal matrix formed by the positive eigenvalues of $\boldsymbol{\Delta}'\boldsymbol{\Delta}$, and $\boldsymbol{U}_i$ is the matrix whose columns are the corresponding eigenvectors.

Setting $\boldsymbol{y} = vec(\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$, $\boldsymbol{\beta} = vec(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m)$ and $\boldsymbol{u} = vec(\boldsymbol{b}_1, ..., \boldsymbol{b}_m, \breve{\boldsymbol{\alpha}}_1, ..., \breve{\boldsymbol{\alpha}}_n)$, we can show that the minimization of the PRSS in (10) with respect to the regression coefficients corresponds to the maximization of the log-likelihood of $(\boldsymbol{y}, \boldsymbol{u})$ arising from the mixed model representation

$$\boldsymbol{y}|\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\ \sigma^2\boldsymbol{I}), \ \ \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0},\ \sigma^2\boldsymbol{P}^{-1}), \tag{12}$$

where

$$\boldsymbol{P} = blockdiag(\underbrace{\lambda\boldsymbol{I}_{c-2}, ..., \lambda\boldsymbol{I}_{c-2}}_{m\ times},\ \breve{\boldsymbol{P}}_1, ..., \breve{\boldsymbol{P}}_n) \tag{13}$$

is the full penalty matrix, $\boldsymbol{\beta}$ is the *fixed effect*, $\boldsymbol{u}$ is the *random effect*, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are appropriate regression matrices; here c is the number of columns in $\boldsymbol{B}_i$. We note that in the balanced case (as in our simulation) we can write $\boldsymbol{X}_i = \boldsymbol{X}_g$, $\boldsymbol{Z}_i = \boldsymbol{Z}_g$ and $\breve{\boldsymbol{B}}_i = \breve{\boldsymbol{B}}$, $i = 1, \ldots, n$, and then in (12) we have $\boldsymbol{X} = blockdiag(\boldsymbol{1}_{n_1} \otimes \boldsymbol{X}_g, \ldots, \boldsymbol{1}_{n_m} \otimes \boldsymbol{X}_g)$ and $\boldsymbol{Z} = [blockdiag(\boldsymbol{1}_{n_1} \otimes \boldsymbol{Z}_g, \ldots, \boldsymbol{1}_{n_m} \otimes \boldsymbol{Z}_g) : \boldsymbol{I}_n \otimes \breve{\boldsymbol{B}}]$.

An illustration of the effectiveness of the penalty approach fitted via the mixed model representation is shown in Figure 2: (a) the top right panel shows the fitted group effect (blue lines) (b) the lower left panel compares the $MSE^{(r)}$ of the penalty approach (blue dots) with the $MSE^{(r)}$ of the standard approach (c) the lower right panel compares the mean standard deviation for the three groups for the two approaches. These graphics demonstrate that the penalty approach is successful in removing both the bias in the estimates and the fanning effect in the confidence intervals seen with the standard model (at least in the cases discussed here).

4 Conclusion

In this short paper we have emphasized the importance of dealing appropriately with both *smoothness* and *identifiability* in the analysis of grouped longitudinal data. We have described a penalty approach which allows us to address both these issues directly. The penalty approach yields a model which can be reformulated as a mixed model and we have demonstrated that the estimates of the group effects from this model are free of bias and of fanning of the associated confidence intervals. A fuller description of this work is in preparation together with software to implement our approach.

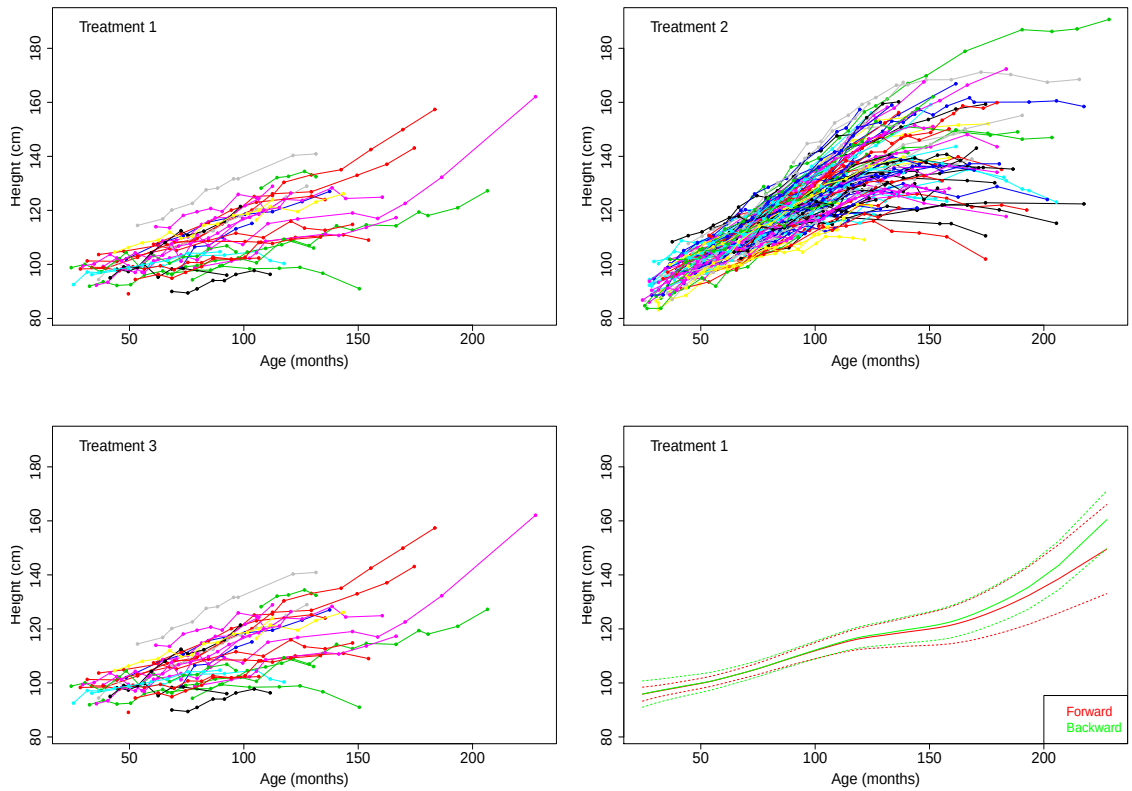

Figure 1: Graphics related to the height measurements of 197 children suffering from acute lymphoblastic leukaemia, and receiving three different treatments. The first three panels show the data. The last panel shows the fitted treatment effect (for group one) from the standard approach with the forward basis (red) and the backward basis (green).

REFERENCES

Coull, B. A., Ruppert, D. and Wand, M. P. (2001). Simple incorporation of interactions into additive models. *Biometrics*, **57**, 539-545.

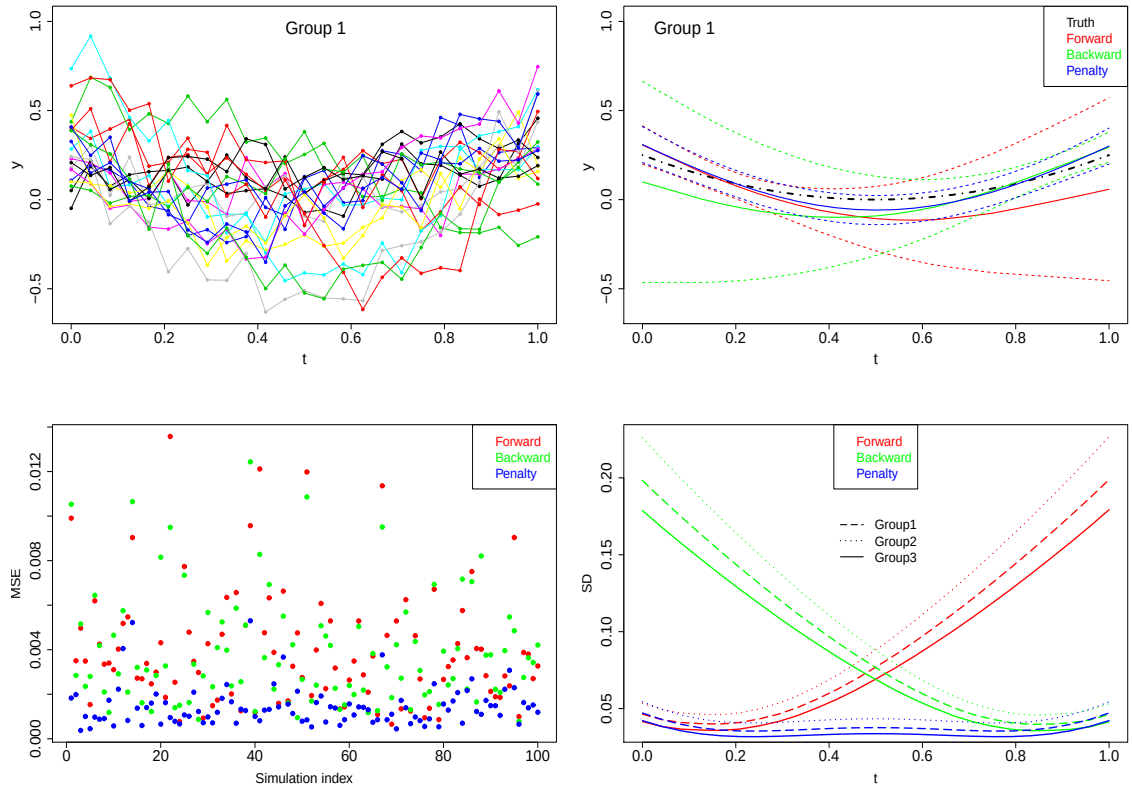

Figure 2: Some graphics related to the simulated data. Top left: a sample of the simulated data. Top right: the fitted group effect for the data in the top left panel. Lower left: comparison of the mean square error. Lower right: Comparison of the standard deviation; each group is denoted by a different line style.

Djeundje, V. A. B. and Currie, I. D. (2010). Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. *Electronic Journal of Statistics*, **4**, 1202-1224.

Durban, M., Harezlak, J., Wand, M. P. and Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153-1167.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89-121.

Reiss, P. T. and Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society, Series B*, **71**, 505-523.

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

ABSTRACT

The fitting of subject-specific curves (also known as factor-by-curve interactions) is important in the analysis of longitudinal data. One approach is to use a mixed model with the curves described in terms of truncated lines and the randomness expressed in terms of normal distributions. This approach gives simple fitting with standard mixed model software. We show that these normal assumptions can lead to biased estimates of and inappropriate confidence intervals for the population effects. We describe an alternative approach which uses a penalty argument to derive an appropriate covariance structure for a mixed model for such data. The effectiveness of our methods is demonstrated with the analysis of some growth curve data and a simulation study.