Penalized Regression, Mixed Effects Models and Appropriate Modelling

N. HECKMAN
Department of Statistics, University of British Columbia
R. LOCKHART
Department of Statistics & Actuarial Science, Simon Fraser University
J. D. NIELSEN
School of Mathematics and Statistics, Carleton University

May 2009

Abstract

Linear mixed effects methods for the analysis of longitudinal data provide a convenient framework for modelling within-individual correlation across time. Using spline functions allows for flexible modelling of the response as a function of time. A computational connection between linear mixed effects modelling and spline smoothing makes the use of spline functions in longitudinal data analysis even more appealing. However, care must be taken in exploiting this connection, as resulting estimates of the underlying population mean might not track the data well and associated standard errors might be unreasonably large. We discuss these shortcomings and suggest some easy-to-compute methods to eliminate them.

Keywords Linear Mixed Effects Models; Penalized Smoothing; P-splines; Sandwich Estimator.

1 Introduction

Linear mixed effects models have proven useful in fitting longitudinal data. Part of their popularity arises because one can use flexible basis functions such as B-splines to fit smooth response curves, and can easily incorporate correlation within individuals through the use of random regression coefficients. See, for instance, Verbyla et al. (1999), Fitzmaurice et al. (2004), Ruppert et al. (2003), Ruppert, Wand and Carroll (manuscript, 2009) and the extensive work in animal breeding, including work of Meyer (2005, 2007). For a specific and known covariance structure, there is a computational equivalence between a particular linear mixed effects model and a standard smoothing approach. Computation of curve estimates using this covariance structure can be calculated quickly and easily with existing software (Ngo & Wand, 2004, using the software PROC MIXED in SAS, and White et al., 1999, using the software

ASReml, Gilmour et al., 2006). This connection also provides an automatic choice of the amount of smoothing via the estimation of the ratio of variances in the mixed effects model. The connection between linear mixed effects models and the smoothing method is not only elegant, but has also proven useful in many applications such as the comparison of human growth curves (Durbán et al., 2005).

However, care must be taken in exploiting this computational connection. Inappropriate modelling of the mean structure or too much reliance on the assumed specific covariance structure for the random effects can lead to undesirable effects in the population effect estimation, in variance parameter estimation, and in specification of standard errors. In particular, the smoothing-based covariance structure should not be completely trusted since typically its form does not come from subject area modelling.

We first give as an example a data analysis that, although simple and perhaps not the best approach, clearly illustrates some of the problems that might arise with blind application of the standard smoothing/linear mixed model analysis. Figure 1 shows average daily temperatures recorded at 35 Canadian weather stations, where time $t = 1$ corresponds to January 1. This data set is available in the *fda* library in the statistical software package R. Our goal is inference for the "typical" or expected daily temperature in a Canadian weather station. Another goal, estimation of a particular station's "typical" weather curve, is only briefly addressed here.

Panel a) of Figure 2 contains three estimates of the population mean: one estimate is simply the daily mean of the temperatures. The other two estimates are from linear mixed effects models. They are calculated with the R library *lme*, which uses restricted maximum likelihood (REML) estimates of variance components. For a discussion of restricted maximum likelihood estimators, see, for instance, Demidenko (2004). The specifics of our calculations are given in Sections 2 and 3. The two mixed effects estimates are calculated using the same function spaces for the population mean and the individual station effects, but the two estimates use different covariance structures on the random effects, covariance structures commonly used in smoothing. One mixed effects estimate uses a covariance structure corresponding to time "running forward", the other uses a covariance structure corresponding to time "running backward". These two mixed effects estimates do not track the pointwise average well. The two estimates are different, but one is the time-reversed version of the other.

The remaining three panels of Figure 2 show the three estimates described above along with pointwise standard errors. Throughout this paper, we plot error bands as plus or minus one standard error. Panel b) shows the pointwise average, with standard errors given by the pointwise standard deviation divided by $\sqrt{35}$. The bottom two panels show standard errors calculated using estimates of the assumed covariance structure of the linear mixed effects models. Panel c) corresponds to the "running backward" covariance structure, panel d) to the "running forward". Note the widening of the standard error bars to values that are much higher than the standard errors of the pointwise averages, standard error bars so wide that they are clearly nonsensical. Thus from Figure 2, we see that the covariance structure assumed for the random effects can seriously impact both estimates and standard errors. The standard errors in panel b) can only be calculated for a balanced design, that is, when temperatures are recorded on the same 365 days for all stations. In many applications, the design is not balanced and these straightforward standard errors cannot be readily calculated.

Figure 3 contains standard error bars calculated via our recommended methods, as described in Section 3.2. These standard error bars do not show the widening as seen in Figure 2.

The observation that the assumed covariance structure of the random effects in a linear mixed effects model can impact estimates and standard errors is not completely new. Misspecification of the covariance structure might have an effect on the estimated means, and typically can have a big effect on standard errors and on inference. Fortunately, standard errors can be corrected via appropriate sandwich estimators. See Liang & Zeger (1986) for a discussion of these issues in the context of generalized estimating equations. Unfortunately, currently published work on smoothing in mixed effects models has not sufficiently addressed this potential problem. One notable exception is the recent work of Brumback et al. (2009), who note the serious implications of reliance on the smoothing-induced covariance and suggest a computer-intensive bootstrap procedure to rectify the problems.

We propose using function estimators calculated assuming the specific covariance structure in the mixed effects-smoothing formulation, and we recommend some restriction on the function spaces used in modelling. These smoothing based estimators are fast to calculate. To compute standard errors and to make inference, we recommend using a more general covariance structure or using a sandwich estimator. Both recommendations are fast to compute, as they use the output of the smoothing-based analysis. Section 2 contains notation and the general formulation of the linear mixed effects model. Sections 3 and 4 contain detailed calculations and discussion of estimators, predictors, and standard errors. Section 3 covers the conceptually straightforward model in which the population curve is nonrandom. In Section 4, the population curve is random.

2 General Formulation

Data are collected on N independent subjects, with data on subject i, (t_{ij}, Y_{ij}), $j = 1, \ldots, n_i$, modelled as

$$Y_{ij} = f_i(t_{ij}) + \epsilon_{ij} \equiv \mu(t_{ij}) + g_i(t_{ij}) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2), \text{ independent.} \quad (1)$$

To model the population curve μ and individual i's deviation g_i, we use four sets of generic basis functions $\{\psi_{Pj}, 1 \le j \le J_P\}$, $\{\phi_{Pk}, 1 \le k \le K_P\}$, $\{\psi_{Ij}, 1 \le j \le J_I\}$ and $\{\phi_{Ik}, 1 \le k \le K_I\}$:

$$\mu(t) = \sum_j \beta[j]\, \psi_{Pj}(t) \; + \; \sum_k \delta[k]\, \phi_{Pk}(t), \quad (2)$$

$$g_i(t) = \sum_j \beta_i[j]\, \psi_{Ij}(t) \; + \; \sum_k \delta_i[k]\, \phi_{Ik}(t). \quad (3)$$

The β_i's and δ_i's are random effects, β is a fixed effect, and δ can be either fixed, as in Section 3, or random, as in Section 4. Throughout, a subscript of P denotes "population" and a subscript of I denotes "individual".

The model has two components: the choice of the basis functions and the assumed covariance structure of the random coefficients. These two elements determine the covariance between $\mu(t) + g_i(t)$ and $\mu(s) + g_i(s)$, and it is the structure of covariance that is crucial in analysis. Clearly, to preserve this covariance structure, any change of basis must be accompanied by the appropriate change in the covariance of the random coefficients.

In all of our examples, the first summations in (2) and (3) are polynomials and the second summations are splines. Specifically, $\{\psi_{Pj}, \phi_{Pk}, 1 \le j \le J_P, 1 \le k \le K_P\}$ and $\{\psi_{Ij}, \phi_{Ik}, 1 \le j \le J_I, 1 \le k \le K_I, \}$ are bases for spline function spaces. A spline function of degree p with knots $\mathcal{K}_1 < \mathcal{K}_2 < \cdots < \mathcal{K}_K$ is piecewise polynomial of degree p, with the "pieces" defined on the subintervals determined by the knots. The pieces of the polynomial are joined together at the knots so that the function has $p-1$ continuous derivatives. Splines are widely used for flexible fitting of functions (see, for instance, Ramsay & Silverman, 2005). A conceptually simple basis, which we use here, is the power basis: $\psi_j(t) = t^{j-1}, j = 1, \ldots, p+1$, and $\phi_k(t) = \{(t - \mathcal{K}_k)_+\}^p$, $k = 1, \ldots, K$. We do not restrict the knots for modelling μ to be the same as the knots for modelling the g_i's, but we do use the same value of p.

Using (1), (2) and (3), we can write the model for subject i's response vector as

$$\begin{aligned} Y_i &= (Y_{i1}, \ldots, Y_{in_i})' \\ &\equiv X_{Pi}\,\beta \;+\; Z_{Pi}\,\delta \;+\; X_{Ii}\,\beta_i \;+\; Z_{Ii}\,\delta_i + \epsilon_i \\ &= [X_{Pi}\;\; Z_{Pi}] \begin{pmatrix} \beta \\ \delta \end{pmatrix} \;+\; [X_{Ii}\;\; Z_{Ii}] \begin{pmatrix} \beta_i \\ \delta_i \end{pmatrix} \;+\; \epsilon_i \\ &\equiv C_i\theta \;+\; C_{Ii}\theta_i + \epsilon_i \\ &\equiv C_i\theta \;+\; \epsilon_i^*. \end{aligned} \tag{4}$$

We assume that $\theta_1, \ldots, \theta_N, \epsilon_1, \ldots, \epsilon_N$ are all independent, mean zero, and normally distributed. We denote the covariance matrix of the θ_i's as Σ_I and we sometimes make the restriction that $\Sigma_I = \Sigma_I^S$ (S for smooth):

$$\Sigma_I^S = \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \sigma_I^2 \mathrm{I} \end{bmatrix}. \tag{5}$$

Here Σ_β is the unrestricted covariance matrix of β_i and $\sigma_I^2\mathrm{I}$ is the restricted covariance matrix of δ_i, that is, the components of δ_i are assumed to be independent and identically distributed with common variance σ_I^2. This restricted model is easy to fit provided the dimension of Σ_β is small. In contrast, the model with unrestricted Σ_I usually causes computational problems unless the dimension of Σ_I is small. We call the model assuming (5) the "smooth g_i" model, for reasons given below. Sometimes we assume that δ is random with covariance matrix equal to a constant times the identity matrix. Again, we refer to this as the "smooth μ" model; again this model is usually easier to fit than one assuming an unrestricted covariance matrix for δ.

We consider four models:

A. μ is fixed, g_i is random: assume that β and δ are non-random and Σ_I is unrestricted.

B. μ is fixed, g_i is random and smooth: assume that β and δ are non-random and that Σ_I equals Σ_I^S as in (5).

C. μ is random and smooth, g_i is random: assume that β is fixed, that δ is $N(0, \sigma^2_{P,C}\mathrm{I})$, is independent of the θ_i's and ϵ_i's, and that Σ_I is unrestricted.

D. μ is random and smooth, g_i is random and smooth: assume that β is fixed, that δ is $N(0, \sigma^2_P\mathrm{I})$, is independent of the θ_i's and ϵ_i's, and that Σ_I equals Σ_I^S as in (5).

We use the notation $\sigma^2_{P,C}$ for the model C parameter to avoid confusion between estimating the variance of a component of δ under the unrestricted model C versus the restricted model D.

To see why the restriction on the covariance matrix of the δ_i's leads to a "smooth" g_i, first consider model B. By Henderson's justification (Robinson, 1991), for fixed $\sigma^2_\epsilon, \sigma^2_I$ and Σ_β, the best linear unbiased predictors of the f_i's in (1) are obtained by minimizing

$$\frac{1}{\sigma^2_\epsilon}\sum_{i,j}\{Y_{ij} - f_i(t_{ij})\}^2 + \sum_i \beta_i'\Sigma_\beta^{-1}\beta_i + \frac{1}{\sigma^2_I}\sum_i \delta_i'\delta_i$$

over β, δ, the β_i's and the δ_i's. That is, we minimize

$$\sum_i \left[\sum_j \{Y_{ij} - f_i(t_{ij})\}^2 + \sigma^2_\epsilon\ \beta_i'\Sigma_\beta^{-1}\beta_i + \frac{\sigma^2_\epsilon}{\sigma^2_I}\ \delta_i'\delta_i\right].$$

The ith summand is simply a penalized least squares regression with penalties on β_i and δ_i. When we model g_i as a spline using the power basis, the penalty on δ_i with "smoothing parameter" $\sigma^2_\epsilon/\sigma^2_I$ is one of those proposed by Eilers & Marx (1996) for P-spline smoothing regression, and is recommended by Ruppert et al. (2003). For instance, if $p = 1$, the penalty on δ_i shrinks g_i to a line with the amount of shrinkage depending on $\sigma^2_\epsilon/\sigma^2_I$. When $p = 1$, the effect of the penalty is similar to penalizing for large second divided differences of g_i (Eilers and Marx, 2004, manuscript). Such a penalty is similar to the second derivative penalty that yields a smoothing spline estimate (Green & Silverman, 1994).

Similarly in model D we can use Henderson's justification to show that the predictors of the f_i's are based on minimizers of

$$\sum_i \left[\sum_j \{Y_{ij} - f_i(t_{ij})\}^2 + \sigma^2_\epsilon\ \beta_i'\Sigma_\beta^{-1}\beta_i + \frac{\sigma^2_\epsilon}{\sigma^2_I}\ \delta_i'\delta_i\right] + \frac{\sigma^2_\epsilon}{\sigma^2_P}\delta'\delta.$$

Thus, when using the power basis, the restrictions in our linear mixed effects model lead us to P-spline smoothing type estimators of μ and the g_i's.

We can reparameterize functions in terms of a Bspline basis, which is computationally more stable than the power basis. Consider, for example, reparameterizing the population mean curve μ and suppose that the knots are equispaced with interknot

distance equal to Δ. Then, if $\gamma \equiv (\gamma_1, \ldots, \gamma_{K+2})'$ is the resulting vector of Bspline coefficients, one can easily show that $\delta[k] = (\gamma_{k+2} - 2\gamma_{k+1} + \gamma_k)/\Delta$, $k = 1, \ldots, K$. Therefore, we can write $\delta'\delta$ as $\gamma' D_2' D_2 \gamma / \Delta^2$ where D_2 is a matrix for taking discrete second divided differences. See Eilers and Marx (2004, manuscript) and Welham et al. (2007) for further discussion of the connections between the truncated power basis and a Bspline basis in penalized smoothing.

Durbán et al. (2005) analyzed growth data using the restricted model D. They modelled μ using the power basis with $p = 1$. They considered a range of models for g_i: g_i equal to a random intercept, a random line, or piecewise linear with the same knots as μ. Their main goal was to carry out various hypothesis tests. While their figures do contain prediction bands for their function estimates, they provide no explanation of their calculation.

Special cases of models A and B have been considered elsewhere. In the animal breeding literature see, for instance, Huisman et al. (2002) and Meyer (2005). Rice & Wu (2001) consider model A in a medical context.

2.1 Temperature Data Example

In all of the temperature data analyses, we model functions as splines of degree $p = 1$, using the power basis representation. Knots are equi-spaced with equal distances from the "edges" of 1 and 365: a K knot sequence is constructed with $\mathcal{K}_j = 1 + 364\ j/(K+1)$, $j = 1, \ldots, K$. The estimates in Figure 2 are based on 41 population knots, $\mathcal{K}_{P1}, \ldots, \mathcal{K}_{P41}$, and 7 individual knots, $\mathcal{K}_{I1}, \ldots, \mathcal{K}_{I7}$. The estimates are based on model B.

Thus $\psi_{P1}(t) = \psi_{I1}(t) = 1$ and $\psi_{P2}(t) = \psi_{I2}(t) = t$. When time is "running forward" $\phi_{Pk}(t) = (t - \mathcal{K}_{Pk})_+$ and $\phi_{Ik}(t) = (t - \mathcal{K}_{Ik})_+$, and when time is "running backward" $\phi_{Pk}(t) = (\mathcal{K}_{Pk} - t)_+$ and $\phi_{Ik}(t) = (\mathcal{K}_{Ik} - t)_+$. In models A and C, there is no difference between the time running forward model and the time running backward model. However, there is a difference between time running forward and time running backward when using models B or D.

We may understand the reason for this difference as follows. For time running forward and with covariance structure as in model B,

$$\text{var}\left(\sum_k \delta_i[k]\phi_{Ik}(t)\right) = \sigma_I^2 \ \times \ \sum_k \left\{(t - \mathcal{K}_{Ik})_+\right\}^2,$$

an increasing function of t. Similarly, for time running backward and with covariance structure as in model B, the variance of $\sum_k \delta_i[k]\phi_{Ik}(t)$ is decreasing in t. This is the cause of the widening of our standard error bands clearly visible in Figure 2, as the model implies that there is more variability in our data at one end of the time scale than the other. The widening becomes even worse if we increase the number of knots used to model the individual random effects. This widening didn't occur when we fit model B or D with the individual random effect g_i equal to a line. This agrees with the analysis in Smith & Wand (2008).

Details of calculations of estimators and standard errors are given in the following sections. In summary, when μ is a nonrandom function, we estimate μ by the maxi-

mum likelihood estimator under the restricted model B. We propose easy-to-compute standard errors of this estimate, valid under the unrestricted model A. When μ is a random function, we use the restricted model D to calculate the best linear unbiased predictor of μ and then use the unrestricted model C for easy-to-compute prediction bands. Throughout, we ignore any model-based bias, that is, we assume that (1) - (3) are exact. We do not consider estimation or prediction of μ under models A or C because fitting linear mixed effects models with so many variance parameters is computationally challenging.

The techniques we use in our calculations are not new. Many calculations in the linear mixed effects model appear in Demidenko (2004) and Ruppert et al. (2003). However we present these calculations in a way that clearly shows when we are relying on the smoothing model covariance structure of models B and D and when we are simply using the more general models A and C. We also discuss in Section 4 interpretations of various techniques for error bars of a predictor of μ when μ is random. When μ is random, we should assess variability of the predictor about μ, not about E(μ).

3 Non-random μ

3.1 Estimation of μ under model B

Consider data generated according to (1) through (4) under either model A or B. Since δ is a fixed effect, μ is non-random; thus when we talk about an estimate of $\mu(t)$ and a standard error of the estimator, our meaning is clear. The estimator of θ under the assumptions of model B is simply the generalized least squares estimate, minimizing

$$\sum (Y_i - C_i\theta)'(\Sigma_i^{*S})^{-1}(Y_i - C_i\theta),$$

with Σ_i^{*S} denoting the variance of ϵ_i^* under the "smooth g_i" model B,

$$\Sigma_i^{*S} = X_{Ii}\Sigma_\beta X_{Ii}' \; + \; \sigma_I^2 \; Z_{Ii}Z_{Ii}' + \sigma_\epsilon^2 \mathrm{I}. \tag{6}$$

Therefore the estimator of θ for known variance parameters is

$$\begin{aligned} \tilde{\theta} = \begin{pmatrix} \tilde{\beta} \\ \tilde{\delta} \end{pmatrix} \quad &= \left(\sum C_i'(\Sigma_i^{*S})^{-1}C_i\right)^{-1} \sum C_i'(\Sigma_i^{*S})^{-1}Y_i \\ &\equiv \sum H_i(\Sigma_1^{*S}, \ldots, \Sigma_N^{*S})\; Y_i \equiv \sum H_i Y_i. \end{aligned} \tag{7}$$

The estimator of $\mu(t)$ for given Σ_i^{*S}'s is then $\tilde{\mu}(t) = \sum_j \tilde{\beta}[j]\psi_{Pj}(t) + \sum_k \tilde{\delta}[k]\phi_{Pk}(t)$.

A linear mixed effects model fit of model B yields restricted maximum likelihood variance estimators $\hat{\Sigma}_\beta$, $\hat{\sigma}_I^2$, $\hat{\sigma}_\epsilon^2$, and thus yields $\hat{H}_i = H_i(\hat{\Sigma}_1^{*S}, \ldots, \hat{\Sigma}_N^{*S})$, an estimator of H_i. The maximum likelihood estimator $\hat{\theta}$ is then equal to $\sum \hat{H}_i Y_i$ and the maximum likelihood estimator of $\mu(t)$, $\hat{\mu}(t)$, is gotten in the obvious way from $\hat{\theta}$. The method also provides estimators, $\hat{\theta}_i$ $i = 1, \ldots, N$, of the best linear unbiased predictors of the θ_i's. These estimators, commonly called the estimated best linear unbiased predictors, are gotten by substituting covariance estimates into the expressions for the best linear unbiased predictors.

3.2 Calculation of standard errors

The estimator $\tilde{\mu}$ is derived under the assumption that model B holds. In this section, we calculate the standard deviation of $\tilde{\mu}(t)$ valid under the unrestricted model A. We then use this standard deviation to compute a standard error of $\hat{\mu}(t)$ by plugging in variance parameter estimates that are appropriate under model A. We ignore variability caused by estimation of the variance parameters that appear in $\hat{\mu}$.

The variance of $\tilde{\mu}(t)$ is easily calculated from $\mathrm{var}(\tilde{\theta})$ using variance/covariance rules as

$$\mathrm{var}(\tilde{\theta}) = \sum H_i \, \mathrm{var}(\epsilon_i^*) \, H_i'. \tag{8}$$

Keep in mind that H_i contains model B variance parameters while var(ϵ_i^*) contains model A variance parameters.

If the restricted model B holds, then the covariance matrix of ϵ_i^* is equal to Σ_i^{*S} and $\mathrm{var}(\tilde{\theta})$ simplifies to $(\sum \mathcal{C}_j'(\Sigma_j^{*S})^{-1}\mathcal{C}_j)^{-1}$, which we can estimate by $(\sum \mathcal{C}_j'(\hat{\Sigma}_j^{*S})^{-1}\mathcal{C}_j)^{-1}$ where $\hat{\Sigma}_j^{*S}$ is obtained by fitting model B. This expression was used to estimate the variance of $\hat{\theta}$ needed to construct the standard error bars in Figure 2, panels c) and d). Clearly, we do not want to use this model-based covariance, as it gives unrealistic standard errors for our estimate of μ.

To construct standard errors, we require an estimator of $\mathrm{var}(\epsilon_i^*)$ in (8), an estimator that is valid under model A. The variance of ϵ_i^* is

$$\mathrm{var}(\epsilon_i^*) = C_{Ii}\Sigma_I C_{Ii}' + \sigma_\epsilon^2 \mathrm{I} \tag{9}$$

and thus we require an estimator of σ_ϵ^2 and an unrestricted estimator of $\Sigma_I = \mathrm{var}(\theta_i)$. We estimate Σ_I by $S_{\hat{\theta}}$, the sample covariance matrix of the $\hat{\theta}_i$'s, our estimators of the best linear unbiased predictors gotten from fitting model B:

$$S_{\hat{\theta}} = \frac{1}{N-1}\sum_i \left(\hat{\theta}_i - \sum \hat{\theta}_j/N\right)\left(\hat{\theta}_i - \sum \hat{\theta}_j/N\right)'. \tag{10}$$

We estimate σ_ϵ^2 by

$$\hat{\sigma}_\epsilon^2 = \frac{1}{\mathrm{df}} \ \mathrm{trace} \sum_i (Y_i - C_i\hat{\theta} - C_{Ii}\hat{\theta}_i)(Y_i - C_i\hat{\theta} - C_{Ii}\hat{\theta}_i)'. \tag{11}$$

where

$$\mathrm{df} = \sum_1^N n_i - \mathrm{length}(\theta) + \mathrm{df}_{\mathrm{adj}} - \sum_1^N \mathrm{length}(\theta_i)$$

and $\mathrm{df}_{\mathrm{adj}}$ corrects for parameter over-counting, adjusting for the fact that some of the population level basis functions are equal to the individual level basis functions. In our case, with a slope and intercept at both the population and individual level, $\mathrm{df}_{\mathrm{adj}} = 2 +$ the number of common population and individual knots. Another sensible estimate of σ_ϵ^2 can be gotten by ordinary least squares, with no shrinkage in estimation of any basis function coefficient.

Our estimator of $\mathrm{var}(\epsilon_i^*)$ is $\widehat{\mathrm{var}}(\epsilon_i^*) = C_{Ii} S_{\hat{\theta}} C_{Ii}' + \hat{\sigma}_\epsilon^2 \mathrm{I}$, and we use this in (8) to estimate the variance of $\hat{\theta}$:

$$\widehat{\mathrm{var}}(\hat{\theta}) = \sum \hat{H}_i \, (C_{Ii} S_{\hat{\theta}} C_{Ii}' + \hat{\sigma}_\epsilon^2 \mathrm{I}) \, \hat{H}_i'. \tag{12}$$

In the special case that that the population knots and individual knots are the same and the t_{ij}'s do not depend on i, our formula for the degrees of freedom simplifies: with $n = n_i$ and $K =$ the number of knots, df $= Nn - N(K+2)$. The resulting estimates of σ_ϵ^2 and $\mathrm{var}(\hat{\theta})$ agree with Demidenko's (2004, pp 61 ff).

The variance estimator in (12) relies on the assumed form of the variance of ϵ_i^* given in model A. If this form is suspect, if, for instance, the covariance matrix of ϵ_i is not a constant times the identity, then the following general sandwich estimator of the variance of $\hat{\theta}$ might be preferred:

$$\widehat{\mathrm{var}}_s(\hat{\theta}) = \sum \hat{H}_i \, (Y_i - C_i\hat{\theta})(Y_i - C_i\hat{\theta})' \, \hat{H}_i'. \tag{13}$$

Robert-Granié et al. (2002) consider such a sandwich estimator when fitting a simple random regression model assuming a specific variance structure that depends on covariates.

3.3 Balanced Design

We call a design for (4) balanced if $C_i \equiv C$ and $C_{Ii} \equiv C_I$. For such a design, the variance of ϵ_i^* does not depend on i. In this case, the model B estimator of θ in (7) simplifies, after some algebra, to $\tilde{\theta} = \{C'[\Sigma_i^{*S}]^{-1}C\}^{-1}C'[\Sigma_i^{*S}]^{-1}\bar{Y}$ which only depends on the data via $\bar{Y} = \sum Y_i/N$. Since $H_i \equiv H$ does not depend on i, the sandwich variance estimator in (13) is equal to $\widehat{\mathrm{var}}_s(\hat{\theta}) = \hat{H} \, \sum (Y_i - \bar{Y})(Y_i - \bar{Y})' \, \hat{H}'$. Thus, we see that estimating the variance of the ϵ_i^*'s in (8) via model B based residuals is equivalent to estimating the variance using the sample variance of the Y_i's.

Suppose that the design is balanced and that model (4) holds with Σ_θ denoting the possibly restricted covariance matrix of θ_i. Then the maximum likelihood estimator of θ when variance parameters are known is simply the generalized least squares estimate

$$\tilde{\theta}_G = \{C'[\mathrm{var}(\epsilon_i^*)]^{-1}C\}^{-1}C'[\mathrm{var}(\epsilon_i^*)]^{-1}\bar{Y} \tag{14}$$

with $\mathrm{var}(\epsilon_i^*) = C_I \Sigma_\theta C_I' + \sigma_\epsilon^2 \mathrm{I}$.

Under an additional condition on C and C_I, given in the following theorem, the estimator $\tilde{\theta}_G$ is equal to the ordinary least squares estimator and thus does not depend on the assumed covariance matrix. Under the same condition explicit formulas for the maximum likelihood and restricted maximum likelihood estimators for Model A covariance parameters can be given; see the end of this section.

Suppose that model (4) holds with $C_i \equiv C$ and $C_{Ii} \equiv C_I$. If the column space of C_I is contained in the column space of C, then the $\tilde{\theta}_G$ in (14) and $\hat{\theta}$, the corresponding maximum likelihood estimator when variance parameters are unknown, are equal to the ordinary least squares estimate $\hat{\theta}_O \equiv (C'C)^{-1}C'\bar{Y}$.

The proof of Theorem 3.3 uses Theorem 3.3 below, which is a modified, more general version of Theorem 2 of Section 2.3 of Demidenko (2004). The proof of Theorem 3.3 is given after that of Theorem 3.3.

Suppose that G is a matrix of full column rank, that M is a symmetric matrix with $M + \mathrm{I}$ invertible and that the column space of M is contained in the column space of G. Then

$$\{G' (M + \mathrm{I})^{-1} G\}^{-1} G' (M + \mathrm{I})^{-1} = (G'G)^{-1} G'.$$

We take transposes and show that

$$(M + \mathrm{I})^{-1} G\{G' (M + \mathrm{I})^{-1} G\})^{-1} - G (G'G)^{-1} = 0. \tag{15}$$

Define temporarily $Q = G' (M + \mathrm{I})^{-1} G$ and $P = \mathrm{I} - G (G'G)^{-1} G'$. The left hand side of (15) is then

$$\begin{aligned} \left\{(M + \mathrm{I})^{-1} G - G (G'G)^{-1} Q\right\} Q^{-1} &= \left\{(M + \mathrm{I})^{-1} G - G (G'G)^{-1} G' (M + \mathrm{I})^{-1} G\right\} Q^{-1} \\ &= P (M + \mathrm{I})^{-1} G Q^{-1}. \end{aligned}$$

The matrix P projects onto the orthogonal complement of the column space of G and so $PG = 0$. Also, since the column space of M is in the column space of G, we see $PM = 0$. Since $\left\{\mathrm{I} - M (M + \mathrm{I})^{-1}\right\} (M + \mathrm{I}) = \mathrm{I}$ we get

$$P (M + \mathrm{I})^{-1} G Q^{-1} = P \left\{\mathrm{I} - M (M + \mathrm{I})^{-1}\right\} G Q^{-1} = PGQ^{-1} = 0.$$

[of Theorem 3.3]

Write $\mathrm{var}\,(\epsilon_i^*) \equiv \sigma_\epsilon^2 (M+\mathrm{I})$ where $M = C_I \Sigma_\theta C_I' / \sigma_\epsilon^2$. Then $\tilde{\theta}_G = \{C' (M + \mathrm{I})^{-1} C\}^{-1} C' (M + \mathrm{I})^{-1} \bar{Y}$ and the Ordinary Least Squares estimator is $\hat{\theta}_O = (C'C)^{-1} C' \bar{Y}$. Since the column space of C_I lies in the column space of C, the column space of $C_I B$ also lies in the column space of C for any matrix B. Thus the column space of M lies in the column space of C. The result follows directly from Theorem 3.3.

Demidenko (2004, pp 61ff) establishes the conclusion of Theorem 3.3 in a balanced design under the stronger condition $C = C_I$; he then gives, under this same condition, explicit formulas for the maximum likelihood and restricted maximum likelihood estimates of Σ_I and σ_ϵ^2 for the unrestricted model A. Careful reading of his proof shows that these formulas remain valid whenever generalized least squares reduces to ordinary least squares. Thus under the conditions of Theorem 3.3 we find that the restricted and unrestricted maximum likelihood estimators of σ_ϵ^2 under model A are equal and given by

$$\hat{\sigma}_\epsilon^2 = \sum_{i=1}^{N} (Y_i - C\hat{\theta}_0)' \left\{\mathrm{I} - C_I (C_I' C_I)^{-1} C_I'\right\} (Y_i - C\hat{\theta}_0) / \{N(n - m)\}$$

where C_I is m by m. The maximum likelihood estimator of Σ_I is

$$\hat{\Sigma}_{I,ml} = (C_I' C_I)^{-1} C_I' S C_I (C_I' C_I)^{-1} - \hat{\sigma}_\epsilon^2 (C_I' C_I)^{-1}$$

where

$$S = \sum (Y_i - C\hat{\theta}_O)(Y_i - C\hat{\theta}_O)' / N.$$

To get the restricted maximum likelihood estimator of Σ_I replace the N in the denominator of S by $N - 1$. See Demidenko (2004, p 63).

3.4 Temperature data

Figures 2 and 3 contain estimates of the mean temperature curve along with pointwise standard errors. All plots assume that the population curve μ is non-random. Figure 2 panels c) and d) were created using 41 population knots and 7 individual knots, with model-based standard errors, that is standard errors based on (8) with var(ϵ_i^*) calculated under the restricted model B. For comparison, we also computed pointwise average temperatures and pointwise standard deviations divided by the square root of 35. These are shown in panel b). The standard errors in panels c) and d) show unrealistic widening. Panel a) contains the three estimates from panels b), c) and d). Note the poor tracking of the pointwise average in panel a).

In Figure 3, estimation of μ involved 41 population knots and 6 individual knots. In panel a) of Figure 3, the standard errors were calculated using the sandwich variance estimator in (13). Panel b) compares the pointwise standard errors of average daily temperature with this sandwich estimator and with the variance estimator in (12). We have not plotted the model-based standard errors but they exhibit the same undesirable fanning behaviour shown in Figure 2. Indeed, model-based standard errors exhibit fanning for a wide range of choices of number of knots.

For the analysis of the weather data using spline functions of degree p, if the knots for the g_i's are a subset of the knots for the population curve μ, then the column space of C_I is contained in the column space of C. Therefore, by Theorem 1, $\tilde{\mu}(t_i)$ does not depend on the specific basis functions or on the assumed covariance structure of the station-specific random effects. Consequently our "forward time" and "backward time" estimates of $\mu(t_i)$ are the same. In Figure 2, the knot choices do not satisfy the conditions of Theorem 1, and we see that the two estimates of μ are different, as expected. In Figure 3, the knot choices do satisfy the conditions of Theorem 1.

It is important to remember that both model-based and sandwich standard errors for $\hat{\mu}$ are affected by the assumed covariance structure. Even if the "forward" estimator of μ is the same as the "backward" estimator of μ, the model B based standard errors of the "forward" estimator will, in general, be different from those of the "backward" estimator. In plots not shown here, the standard errors using (12) or (13) also exhibit fanning, albeit mild, unless the conditions of Theorem 1 hold.

4 Random μ

4.1 Prediction of μ under model D

Suppose data are generated according to (1) through (4) under either model C or D, so that the population effect δ is random. Under either model we write

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_N \end{bmatrix} \begin{pmatrix} \beta \\ \delta \end{pmatrix} + \begin{pmatrix} \epsilon_1^* \\ \epsilon_2^* \\ \vdots \\ \epsilon_N^* \end{pmatrix} \equiv \mathcal{C}\,\theta + \epsilon^*. \qquad (16)$$

To calculate the estimators of β and the best linear unbiased predictor of δ, we assume that the restricted model D holds. The variance of ϵ_i^* under model D is Σ_i^{*S} as defined in (6). Let v_0 be a row vector of zeroes the same length as β and v_1 a row vector of ones the same length as δ. Let

$$\mathcal{S} = \text{diag}(\Sigma_1^{*S}, \Sigma_2^{*S}, \ldots, \Sigma_N^{*S}) \text{ and } \mathrm{I}_0 = \text{ diag } (v_0, v_1).$$

Then, for known variance parameters, under model D, $\tilde{\beta}$, the maximum likelihood estimator of β, and $\tilde{\delta}$, the best linear unbiased predictor of δ, can be found via Henderson's justification (Robinson, 1991), as the minimizers of

$$(Y - \mathcal{C}\theta)'\mathcal{S}^{-1}(Y - \mathcal{C}\theta) \ + \ \frac{1}{\sigma_P^2}\,\theta'\mathrm{I}_0\theta = (Y - \mathcal{C}\theta)'\mathcal{S}^{-1}(Y - \mathcal{C}\theta) \ + \ \frac{1}{\sigma_P^2}\delta'\delta$$

and so

$$\begin{aligned}
\tilde{\theta} &= \left[\mathcal{C}'\mathcal{S}^{-1}\mathcal{C} + \frac{1}{\sigma_P^2}\mathrm{I}_0\right]^{-1}\mathcal{C}'\mathcal{S}^{-1}Y \\
&= \left(\sum C_i'(\Sigma_i^{*S})^{-1}C_i \ + \ \frac{1}{\sigma_P^2}\mathrm{I}_0\right)^{-1} \ \sum C_i'(\Sigma_i^{*S})^{-1}Y_i \\
&\equiv \sum \mathcal{H}_i(\Sigma_1^{*S}, \ldots, \Sigma_N^{*S}, \sigma_P^2)Y_i \equiv \sum \mathcal{H}_i Y_i.
\end{aligned}$$

Therefore, our predictor of $\mu(t)$ in model D for known variance parameters is

$$\tilde{\mu}(t) = \tilde{\theta}'(\psi_{P1}(t), \ldots, \psi_{PJ_P}(t), \phi_{P1}(t), \ldots, \phi_{PK_P}(t)) \ \equiv \tilde{\theta}' f(t).$$

A linear mixed effects model fit of model D yields restricted maximum likelihood estimators of σ_P^2, Σ_β, σ_I^2 and σ_ϵ^2, and thus estimators of $\mathcal{H}_i$, denoted $\hat{\mathcal{H}}_i$. The fit also produces estimators of the best linear unbiased predictors of θ and the θ_i's. The estimator of the best linear unbiased predictor of θ is simply $\hat{\theta} = \sum \hat{\mathcal{H}}_i Y_i$. The predictor of $\mu(t)$, denoted $\hat{\mu}(t)$, is gotten in the obvious way from $\hat{\theta}$.

4.2 Assessing variability of the predictor of $\mu(t)$

In a random effects model, we have several ways to construct intervals for $\mu(t)$ based on the best linear unbiased predictor $\tilde{\mu}(t)$. Since we are interested in $\mu(t)$ and not the population fixed effect $\mathrm{E}\{\mu(t)\}$, we construct intervals based on a measure of the magnitude of $\tilde{\mu}(t) - \mu(t)$. So, for instance, we do not construct intervals of the form $\tilde{\mu}(t)$ $\pm[\text{var}\{\tilde{\mu}(t)\}]^{1/2}$ since $\text{var}\{\tilde{\mu}(t)\} = \text{var}\{\tilde{\mu}(t) - \sum \beta_j \psi_{Pj}(t)\}$ measures variability of $\tilde{\mu}(t)$ about the population fixed effect, not about $\mu(t)$.

We study two measures of magnitude: $e_\delta^2(t) = \mathrm{E}[\ \{\tilde{\mu}(t) - \mu(t)\}^2 \mid \delta]$ and $e^2(t) = \mathrm{E}\{\tilde{\mu}(t) - \mu(t)\}^2$, and discuss how we might use these measures to construct intervals that are likely to contain $\mu(t)$. The measure $e_\delta^2(t)$ provides inference that holds for each realization of μ, and thus seems the most sensible, as our data set has been generated by only one realization of μ. We can also justify the use of $e_\delta^2(t)$ by thinking of the randomness of μ as merely a mechanism for smoothing. In either case, there is really

just one μ of interest, leading us to think of μ as fixed in our inference. The measure $e^2(t)$ provides inference that holds on average over all realizations of μ. It may perform poorly for some realizations of μ and perform well for others.

Below we calculate $e^2_\delta(t)$ and $e^2(t)$ assuming that the unrestricted model C holds. In Section 4.3, we present estimators of these two measures, estimators that are appropriate under model C.

To calculate $e^2_\delta(t)$ and $e^2(t)$, let

$$A = \mathcal{C}'\mathcal{S}^{-1}\mathcal{C} + \frac{1}{\sigma_P^2}\mathrm{I}_0$$

and write, using (16) and some algebra,

$$\begin{aligned}
\tilde{\theta} - \theta &= A^{-1}\mathcal{C}'\mathcal{S}^{-1}Y - \theta \\
&= -\frac{1}{\sigma_P^2}A^{-1}\mathrm{I}_0\theta + A^{-1}\mathcal{C}'\mathcal{S}^{-1}\epsilon^* \\
&= -\frac{1}{\sigma_P^2}\, A^{-1}\begin{pmatrix} 0 \\ \delta \end{pmatrix} + A^{-1}\mathcal{C}'\mathcal{S}^{-1}\epsilon^*. \qquad (17)
\end{aligned}$$

Consider the first measure:

$$\begin{aligned}
e^2_\delta(t) &= \mathrm{E}[\{\tilde{\mu}(t) - \mu(t)\}^2|\delta] = \mathrm{E}[\,\{f(t)'\,(\tilde{\theta} - \theta)\}^2 \mid \delta] \\
&= f(t)'\ \mathrm{E}\{\,(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)'\,|\delta\}\ f(t) \\
&= f(t)'\ (\,B_{\theta|\delta}B'_{\theta|\delta}\ +\ V_{\theta|\delta}\,)\ f(t)
\end{aligned}$$

where, by (17),

$$B_{\theta|\delta}\ =\ \mathrm{E}(\,\tilde{\theta} - \theta\,|\delta)\ =\ -\,\frac{1}{\sigma_P^2}\,A^{-1}\begin{pmatrix} 0 \\ \delta \end{pmatrix}$$

and

$$\begin{aligned}
V_{\theta|\delta} &= \mathrm{var}(\tilde{\theta}|\delta) = \mathrm{var}(\tilde{\theta} - \theta|\delta) \\
&= A^{-1}\ \mathcal{C}'\mathcal{S}^{-1}\,\mathrm{var}(\epsilon^*)\,\mathcal{S}^{-1}\mathcal{C}\ A^{-1} \\
&= A^{-1}\ \sum C_i'(\Sigma_i^{*S})^{-1}\,\mathrm{var}(\epsilon_i^*)\,(\Sigma_i^{*S})^{-1}C_i\ A^{-1} \\
&\equiv \sum \mathcal{H}_i\,\mathrm{var}(\epsilon_i^*)\,\mathcal{H}_i' \equiv V_\theta, \qquad (18)
\end{aligned}$$

since $V_{\theta|\delta}$ doesn't depend on δ and therefore is not random. So

$$e^2_\delta(t) = f(t)'\left(\frac{1}{\sigma_P^4}\,A^{-1}\begin{bmatrix} 0 & 0 \\ 0 & \delta\delta' \end{bmatrix}A^{-1} + V_\theta\right)f(t). \qquad (19)$$

In the Appendix, we show that $\mathrm{pr}(|\tilde{\mu}(t) - \mu(t)| \le z_{\alpha/2}\,e_\delta(t) \mid \delta) \ge 1 - \alpha$ where the probability is calculated under the unrestricted model C and $z_{\alpha/2}$ is the 1-$\alpha/2$ quantile of the standard normal distribution. So $\tilde{\mu}(t) \pm z_{\alpha/2}\ e_\delta(t)$ is a sensible conservative interval for $\mu(t)$, one that performs well for each realization of μ. In the Appendix, we see that the interval may be unnecessarily conservative if $\delta'\delta$ is large.

Now consider the second measure, $e^2(t)$. To calculate $e^2(t)$, we simply take the expectation of (19) under model C:

$$e^2(t) = f(t)' \left(\frac{1}{\sigma_P^4} A^{-1} \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{P,C}^2 \mathrm{I} \end{bmatrix} A^{-1} + V_\theta \right) f(t). \tag{20}$$

We argue here that, on average over realizations of μ (with probability $1-\alpha$), $\mu(t)$ will lie in the interval $\tilde{\mu}(t) \pm z_{\alpha/2}\, e(t)$. From (17), $\mathrm{E}(\tilde{\theta}-\theta) = 0$ and so $\mathrm{E}\{\tilde{\mu}(t)-\mu(t)\} = 0$. Thus $e^2(t) = \mathrm{var}\{\tilde{\mu}(t) - \mu(t)\}$ and so $\mathrm{pr}(\mu(t) \in \tilde{\mu}(t) \pm z_{\alpha/2}\, e(t)) = 1 - \alpha$.

If the variance model is correctly specified, that is, if model D holds, then $e^2(t)$ has an additional interpretation that further supports its use in constructing inference intervals. When the variance is correctly specified, $\tilde{\mu}(t) = E\{\mu(t)|Y_1, \ldots, Y_N\}$ and $\mathrm{var}\{\mu(t)|Y_1, \ldots, Y_N\} = \mathrm{E}[\{\mu(t) - \tilde{\mu}(t)\}^2|Y_1, \ldots, Y_N]$, which, in the normal model, does not depend on $Y_1, \ldots, Y_N$. So $\mathrm{var}\{\mu(t)|Y_1, \ldots, Y_N\} = \mathrm{E}[\mathrm{var}\{\mu(t)|Y_1, \ldots, Y_N\}]$ $= e^2(t)$. That is, $e^2(t)$ is equal to the posterior variance of $\mu(t)$ given the data. This posterior variance is commonly used for assessing variability of the posterior mean.

Ruppert et al. (2003, Section 6.4) discuss the analogues of V_θ, $e_\delta^2(t)$ and $e^2(t)$ in the case that $N = 1$, that is, in the case of P-spline smoothing regression. To calculate confidence intervals for $\mu(t)$, they compare $e^2(t)$ and $f(t)'V_\theta f(t)$, and state they prefer the former. They don't consider confidence intervals based on $e_\delta^2(t)$. Their calculations assume that the smoothing model D holds, that is, that Σ_I is as in (5), while our calculations hold under the more general model C. The forms of our (19) and (20) allow for easier interpretation and comparison.

4.3 Estimating $e_\delta^2(t)$ and $e^2(t)$

We have defined $\tilde{\mu}(t)$ and $\hat{\mu}(t)$, predictors of $\mu(t)$, in the restricted model D. In Section 4.2, we defined two measures of the variability of $\tilde{\mu}(t)$, $e_\delta^2(t)$ in (19) and $e^2(t)$ in (20). Our calculations for $e_\delta^2(t)$ and $e^2(t)$ are valid under the unrestricted model C. In this section, we define estimators of $e_\delta^2(t)$ and $e^2(t)$ that are also valid under the unrestricted model C. We then use these estimators to define prediction intervals for $\mu(t)$ centered at $\hat{\mu}(t)$. Keep in mind that $A, \mathcal{S}$ and σ_P^2 are estimated by fitting model D. Variance parameters appearing in other parts of $e_\delta^2(t)$ and $e^2(t)$ must be estimated using the unrestricted model C.

Both $e_\delta^2(t)$ and $e^2(t)$ contain the unknown V_θ. We can estimate V_θ in two ways: using model C to estimate $\mathrm{var}(\epsilon_i^*)$ in (18) or using a more general sandwich estimator. To use model C, write $\mathrm{var}(\epsilon_i^*) = C_{Ii}\Sigma_I C_{Ii}' + \sigma_\epsilon^2 \mathrm{I}$. We estimate Σ_I and σ_ϵ^2 as in Section 3.2 equations (10) and (11), but using the $\hat{\theta}_i$'s and $\hat{\theta}$ gotten from fitting model D. To estimate V_θ by a sandwich estimator, we replace $\mathrm{var}(\epsilon_i^*)$ in (18) with $(Y_i - C_i\hat{\theta})(Y_i - C_i\hat{\theta})'$.

To estimate $e_\delta^2(t)$ we must predict $\delta\delta'$, while to estimate $e^2(t)$, we must estimate $\sigma_{P,C}^2$. Let $\hat{\delta}$ be the estimator of the best linear unbiased predictor of δ, calculated by fitting model D. We predict $\delta\delta'$ by $\hat{\delta}\hat{\delta}'$. We estimate $\sigma_{P,C}^2$ by the sample variance of the

components of $\hat{\delta}$:

$$\hat{\sigma}^2_{P,C} = \frac{1}{K_P - 1} \sum_k \left(\hat{\delta}[k] - \sum \hat{\delta}[j]/K_P \right)^2 .$$

4.4 Temperature data

We constructed a figure (not shown) analogous to Figure 2, except based on model D with μ random using $e^2(t)$. The figure showed the same features as in Figure 2: estimates of μ did not track the pointwise average, pointwise prediction bands were unreasonably wide at 365 days if time was "running forward" or wide at day 1 if time was "running backward".

Figure 4 shows estimates of μ using model D with 41 population knots and 6 individual knots. The top panel shows the estimate and pointwise prediction bars calculated with the sandwich estimate of $e^2(t)$. The bottom panel compares three prediction bars: the pointwise standard deviation divided by $\sqrt{35}$, and those gotten from estimating $e^2(t)$ via model C and via the sandwich method. We see that the prediction error based on sandwich estimation smoothly tracks the pointwise standard errors while the D-model-based prediction error bars are even smoother, but clearly show the effect of knot placement. Plots of prediction bars based on estimating $e^2_\delta(t)$ were almost identical to those based on $e^2(t)$ at the beginning of the year, and only slightly different at the end of the year.

5 Discussion

The use of linear mixed effects modelling as a smoothing tool in the analysis of longitudinal data analysis has increased, with many researchers taking advantage of readily available mixed effects model software. Incorporating spline functions into the analysis allows more flexible estimators than those from traditional parametric methods. However, use of these spline models raises some concerns.

In the literature, typically the population curve is modelled as a spline function while individual curves are modelled simply, as just a random intercept, or perhaps as a random line. However, a richer individual model might be necessary for the data. For instance, it's clear from Figure 1 that the station effect cannot be modelled adequately by a random line. As another example, consider studying growth via height/age data. Using only a random intercept for the individual effect on height yields predicted growth rates that do not vary across individuals. Thus a richer model would be needed. See Smith & Wand (2008).

Unfortunately, for a rich individual model the unrestricted models A and C are computationally difficult to fit. But estimators and predictors in the restricted models B and D are fast to compute and are good, provided knots are chosen appropriately. For estimating the population curve, we recommend that the knots used in modelling the individual curves be a subset of the knots used in modelling the population curve. As stated in Theorem 1, when μ is non-random and the design is balanced in a certain sense, doing so will yield a population curve estimate that doesn't depend on the assumed covariance structure.

The choice of the covariance structure for the population curve and the individual deviations can have a large effect on the standard errors. If possible, one should use a covariance structure that arises from the application. For instance, an appropriate covariance structure for the temperature data should reflect the fact that January 1 is just one day after December 31, and so temperatures on these two days are highly correlated. If an appropriate covariance structure cannot be determined a priori, standard errors should be based on an unrestricted covariance structure because model-based standard errors can be quite unrealistic. We recommend using standard errors based on (12), (13), (19) or (20).

Other authors have proposed alternatives to model-based standard errors. Crainiceanu et al. (2007) assume that the δ_i's are uncorrelated, but with possibly different variances, replacing $\sigma_I^2 \mathrm{I}$ in (5) with a diagonal matrix. Sun et al. (2007) study a slightly more complex model for (4). In order to reduce computational cost they propose two stage estimation: regression to estimate the mean parameters and random effects, followed by method of moments to estimate variance parameters.

Historically, linear mixed effects models have been used to estimate fixed effects. Using them to predict random effects, as in the prediction of μ in Section 4, raises conceptual problems in the interpretation of μ and in how one should construct prediction intervals. We propose basing prediction errors on either of two measures of variability: the conditional mean squared error and the unconditional mean squared error. When the model used to calculate the estimator of μ is correct, then the predictor of μ is the posterior mean and the unconditional mean squared error is simply the posterior variance. The posterior variance is a usual measure of variability of the posterior mean.

Another conceptual problem lies in the role played by the individual random effects. On the one hand, these random effects serve as a model for individual departures from the mean population response. On the other hand, they serve simply as a tool for smoothing individual level responses. These two roles can be aligned if the function space for the population curve contains the function space for the individual effects.

Our two stage methodology, a smoothing-based linear mixed effects fit followed by method of moment estimation of variance parameters, provides fast and flexible analysis of longitudinal data, analysis that is robust to variance model misspecification.

Acknowledgement

This research was supported by grants from the National Science and Engineering Research Council of Canada.

Appendix

Confidence intervals based on mean squared error are commonly used. However, to our knowledge, the rationale has not been published, so we give it here. To apply these results to $e_\delta^2(t)$, simply replace probabilities, expectations and variances with conditional probabilities, expectations and variances.

Consider a parameter θ and an estimator $\hat{\theta}$, assumed to be normally distributed. Let $b = \mathrm{E}(\hat{\theta}) - \theta$, $\sigma^2 = \mathrm{var}(\hat{\theta})$ and $m^2 = \mathrm{E}(\hat{\theta} - \theta)^2 = b^2 + \sigma^2$. We show that $\mathrm{pr}(|\hat{\theta} - \theta| \geq z_{\alpha/2}\, m) < \alpha$. Write

$$\mathrm{pr}(\hat{\theta} - \theta \geq z_{\alpha/2}\, m) = \mathrm{pr}\bigg(\frac{\hat{\theta} - \mathrm{E}(\hat{\theta})}{\sigma} \geq \frac{z_{\alpha/2}\, m - b}{\sigma}\bigg) = \mathrm{pr}\bigg(Z \geq \frac{z_{\alpha/2} m - b}{\sigma}\bigg)$$

where Z follows a standard normal distribution. Similarly,

$$\mathrm{pr}(\hat{\theta} - \theta \leq -z_{\alpha/2}\, m) = \mathrm{pr}\bigg(Z \leq \frac{-z_{\alpha/2}\, m\, - b}{\sigma}\bigg).$$

Consider the function

$$\begin{aligned} H(b) &= \mathrm{pr}(|\hat{\theta} - \theta| > z_{\alpha/2}\, m) \\ &= \mathrm{pr}\bigg(Z \geq \frac{z_{\alpha/2}\, m - b}{\sigma}\bigg) + \mathrm{pr}\bigg(Z \leq \frac{-z_{\alpha/2}\, m\, - b}{\sigma}\bigg) \\ &= \mathrm{pr}\bigg(Z \geq m^* - \frac{b}{\sigma}\bigg) + \mathrm{pr}\bigg(Z \leq -m^* - \frac{b}{\sigma}\bigg). \end{aligned}$$

Clearly $H(b)$ is no larger than $H(0) = \mathrm{pr}(|Z| \geq z_{\alpha/2} m/\sigma)$. So, since $m \geq \sigma$, $H(b) \leq \mathrm{pr}(|Z| \geq z_{\alpha/2}) = \alpha$. The discrepancy between $H(b)$ and α will be large if b^2 is large.

References

BRUMBACK, B. A., BRUMBACK, L. C. & LINDSTROM, M. J. (2009). *Longitudinal Data Analysis*. Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G., eds. Handbooks of Modern Statistical Methods. Boca Raton, Florida: Chapman & Hall/CRC Press, pp. 291–318.

CRAINICEANU, C. M., RUPPERT, D., CARROLL, R. J., JOSHI, A. & GOODNER, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics* **16**, 265–88.

DEMIDENKO, E. (2004). *Mixed Models: Theory and Applications*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Interscience.

DURBÁN, M., HAREZLAK, J., WAND, M. P. & CARROLL, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine* **24**, 1153–67.

EILERS, P. H. C. & MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

FITZMAURICE, G. M., LAIRD, N. M. & WARE, J. H. (2004). *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Interscience.

GILMOUR, A., GOGEL, B., CULLIS, B. R. & THOMPSON, R. (2006). *ASReml User Guide Release 2.0*. Hemel Hempstead, U.K.: VSN International Ltd.

GREEN, P. J. & SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

HUISMAN, A. E., VEERKAMP, R. F. & VAN ARENDONK, J. A. M. (2002). Genetic parameters for various random regression models to describe the weight data of pigs. *Journal of Animal Science* **80**, 575–82.

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

MEYER, K. (2005). Random regression analyses using B-splines to model growth of Australian Angus cattle. *Genetics Selection Evolution* **37**, 473–500.

MEYER, K. (2007). WOMBAT - a tool for mixed model analyses in quantitative genetics by REML. *Journal of Zheijang University Science B* **8**, 815–21.

NGO, L. & WAND, M. P. (2004). Smoothing with mixed model software. *Journal of Statistical Software* **9**, 1–54.

RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. New York: Springer, 2nd ed.

RICE, J. A. & WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–9.

ROBERT-GRANIÉ, C., HEUDE, B. & FOULLEY, J.-L. (2002). Modelling the growth curve of Maine-Anjou beef cattle using heteroskedastic random coefficients models. *Genetics Selection Evolution* **34**, 423–45.

ROBINSON, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15–51.

RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

SMITH, A. D. A. C. & WAND, M. P. (2008). Streamlined variance calculations for semiparametric mixed models. *Statistics in Medicine* **27**, 435–48.

SUN, Y., ZHANG, W. & TONG, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics* **35**, 2795–2814.

VERBYLA, A. P., CULLIS, B. R., KENWARD, M. G. & WELHAM, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal Of The Royal Statistical Society Series C* **48**, 269–311.

WELHAM, S. J., CULLIS, B. R., KENWARD, M. G. & THOMPSON, R. (2007). A comparison of mixed model splines for curve fitting. *Australian & New Zealand Journal of Statistics* **49**, 1–23.

WHITE, I. M. S., THOMPSON, R. & BROTHERSTONE, S. (1999). Genetic and environmental smoothing of lactation curves with cubic splines. *Journal of Dairy Science* **82**, 632–8.

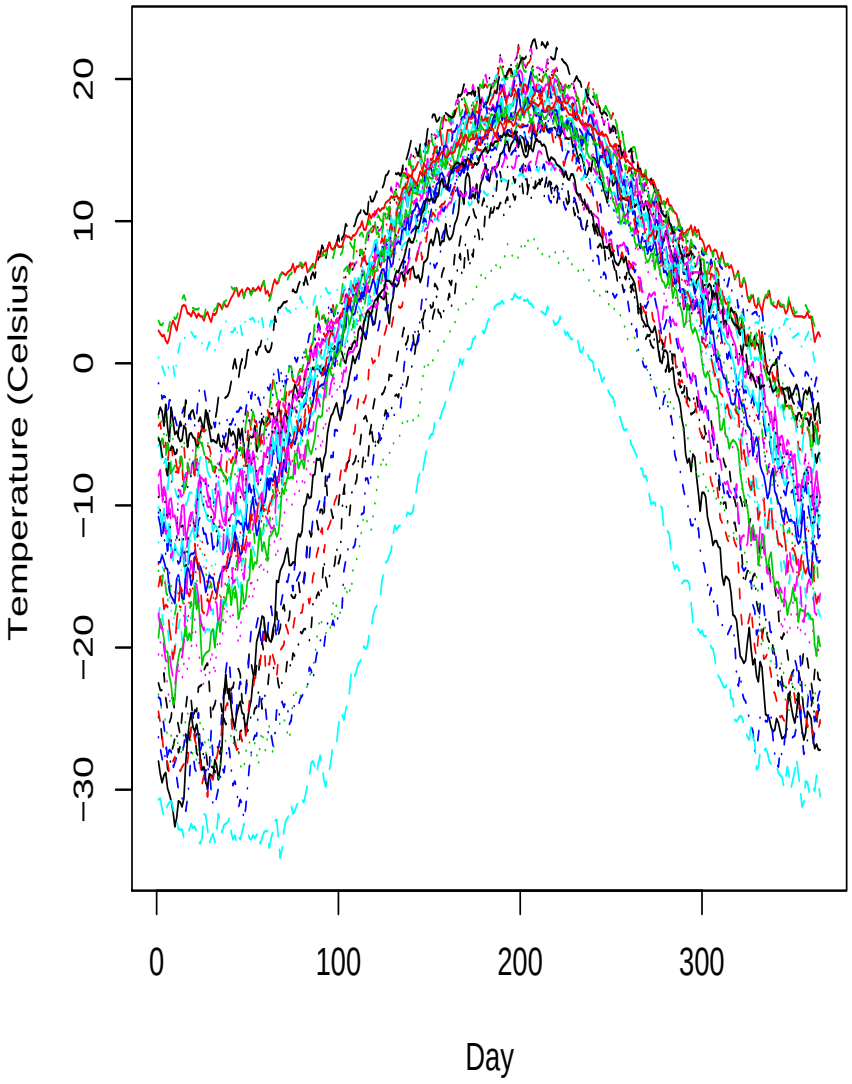

Figure 1: Average daily temperatures (degrees Celsius) at 35 Canadian weather stations. Day 1 is January 1.

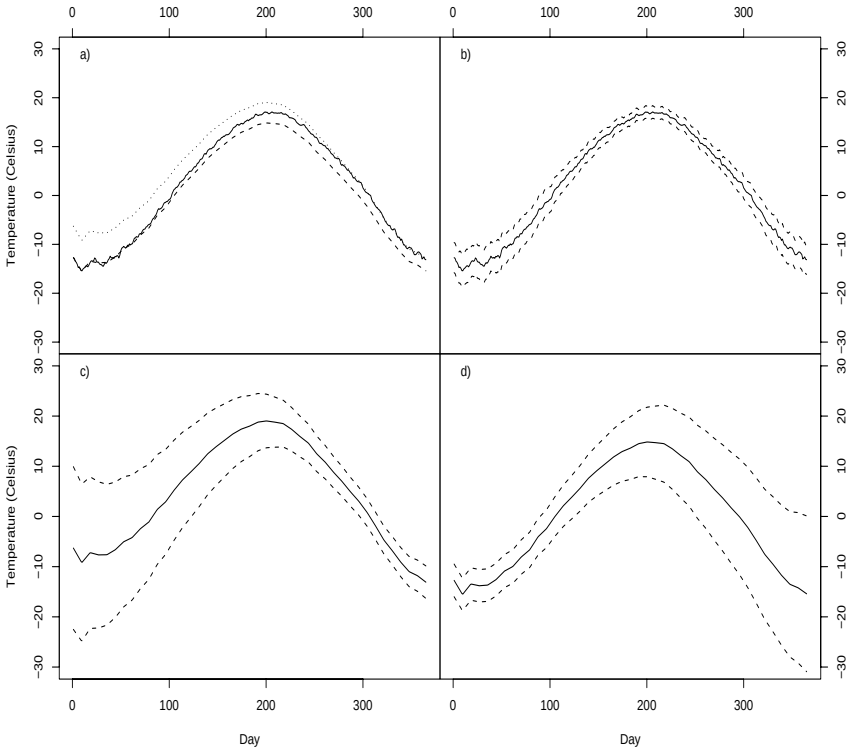

Figure 2: Panel a) contains three estimates of μ, the typical weather curve, as described in Sections 2 and 3, with μ considered non-random. The solid line is the pointwise average, the dashed line is the time "running forward" estimate, and the dotted line is the time "running backward" estimate. These two estimates use 41 population knots and 7 individual knots. In the remaining panels, these three estimates of μ are shown along with bands at plus and minus one standard error. In panel b), the standard errors are simply the pointwise standard deviations of the 35 temperatures, divided by the square root of 35. The bottom two panels contain standard errors calculated using the restricted model B, as described in Section 3. Panel c) contains the "running backward" estimate and panel d) contains the "running forward" estimate.

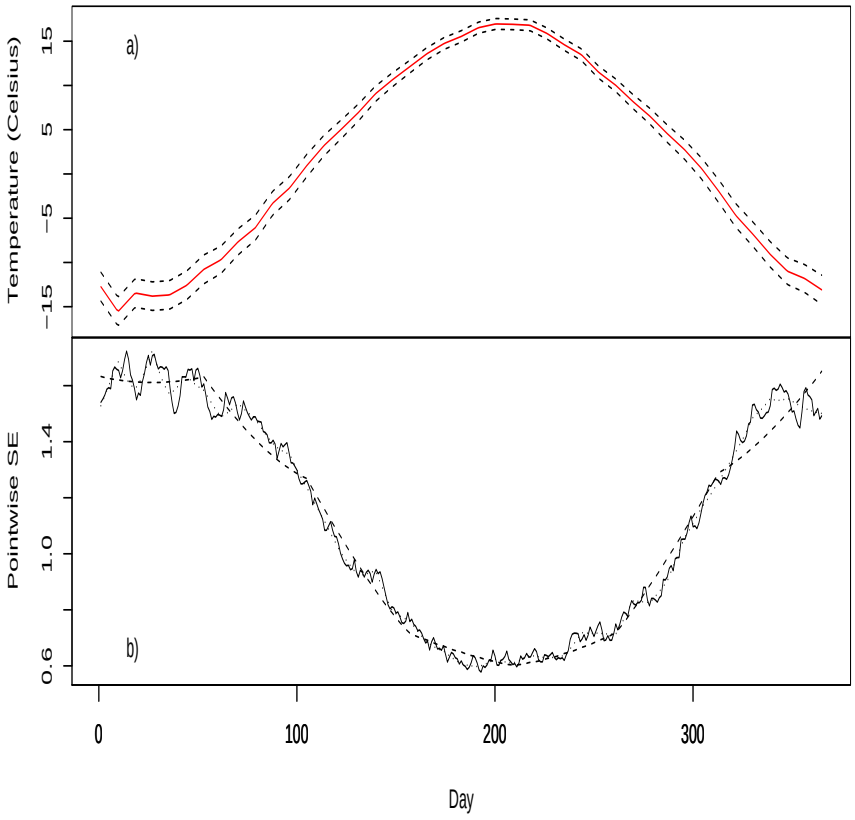

Figure 3: The panels provide information about μ, the typical weather curve, estimated using the techniques of Section 3, where μ is assumed non-random. The model uses 41 population knots and 6 individual knots. Panel a) contains the estimate of μ and pointwise standard errors gotten from the sandwich estimator in (13). Panel b) shows pointwise standard errors calculated using the sandwich estimator (dotted line), the model C estimator of (12) (long dashed line) and the pointwise standard deviation of the temperatures divided by square root of 35 (solid line).

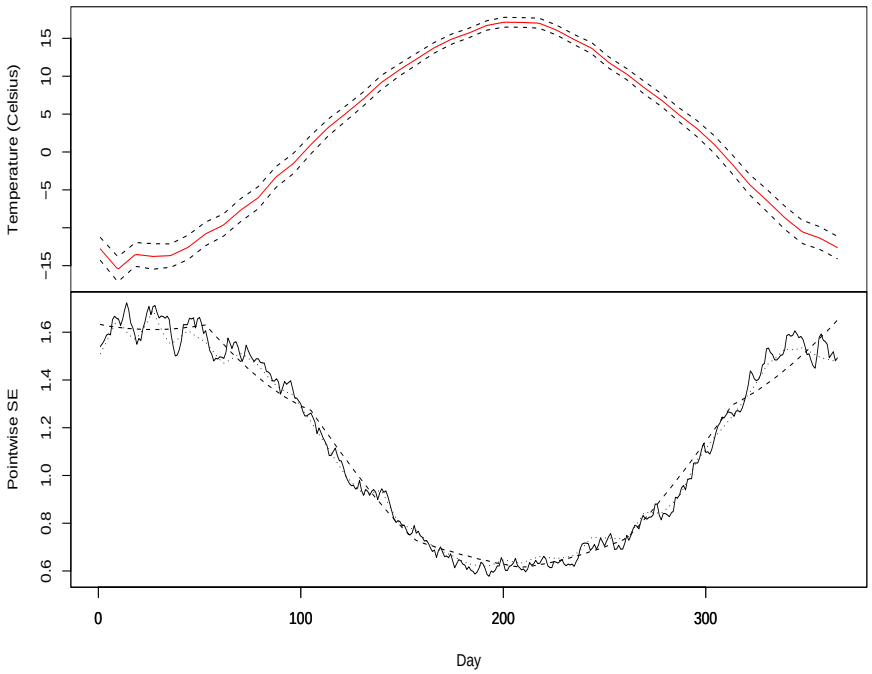

Figure 4: The plots provide information about μ, the typical weather curve, estimated using the techniques of Section 4, where μ is assumed to be random. Prediction bar calculations are given in Section 4.3. The model uses 41 population knots and 6 individual knots. The top panel shows the estimate and pointwise prediction bars calculated with the sandwich estimate of $e^2(t)$. The bottom panel compares three prediction bars: the pointwise standard deviation divided by $\sqrt{35}$, (solid line) and the prediction bars gotten from estimating $e^2(t)$ via model C (dashed line) and the sandwich method (dotted line).