Electronic Journal of Statistics
Vol. 7 (2013) 1517–1552
ISSN: 1935-7524
DOI: 10.1214/13-EJS809

Penalized regression, mixed effects models and appropriate modelling

Nancy Heckman[*]

Department of Statistics, University of British Columbia
Vancouver, British Columbia, V6T 1Z4, Canada
e-mail: nancy@stat.ubc.ca

Richard Lockhart[*]

Department of Statistics & Actuarial Science, Simon Fraser University
Burnaby, British Columbia, V5A 1S6, Canada
e-mail: lockhart@sfu.ca

and

Jason D. Nielsen[*]

School of Mathematics and Statistics, Carleton University
Ottawa, Ontario, K1S 5B6, Canada
e-mail: jdn@math.carleton.ca

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 smooth function of time. A computational connection between linear mixed effects modelling and spline smoothing has resulted in a cross-fertilization of these two fields. The connection has popularized the use of spline functions in longitudinal data analysis and the use of mixed effects software in smoothing analyses. 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 not reflect the true variability in the data. We discuss these shortcomings and suggest some easy-to-compute methods to eliminate them.

AMS 2000 subject classifications: Primary 62G08; secondary 62J99.
Keywords and phrases: Linear mixed effects models, penalized smoothing, P-splines, sandwich estimator.

Received May 2012.

Contents

[*] The authors thank the National Science and Engineering Research Council for its support.

1. Introduction

Two approaches have emerged for the analysis of longitudinal data: linear mixed effects modelling and functional data analysis. Linear mixed effects modelling has its roots in parametric statistics, with, for instance, the response variable assumed to be linear in time (see Demidenko [3]). In contrast, functional data analysis has its roots in smoothing (Ramsay [22]). Recently, these two approaches have become intertwined, with researchers in one approach borrowing methods from the other.

Linear mixed effects modellers have borrowed from smoothing research by replacing linear response models with more flexible piecewise polynomial response models, using basis functions such as B-splines. Within-individual correlation can be incorporated via random regression coefficients. See, for instance, Verbyla *et al* [33], Fitzmaurice [8], Ruppert *et al* ([27, 28]) and the extensive work in animal breeding, including work of Meyer ([19, 20]). However, correct modelling of the within-individual correlation is rarely straightforward.

Smoothers have made good use of the fact that, for a specific and known covariance structure, there is a computational equivalence between a particular linear mixed effects model and a standard smoothing approach. This equivalence parallels the well-known correspondence between Bayes estimation and penalized regression. See, for instance, the work on cubic smoothing splines by Kimeldorf and Wahba [17]. Curve estimates using this covariance structure can be calculated quickly and often, easily with existing software (Ngo [21] using the software PROC MIXED in SAS, and White *et al* [35] using the software AS-Reml by Gilmour *et al* [11]). 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 smoothing is not only elegant, but has also proven useful in many applications such as the comparison of human growth curves (Durban *et al*, [5]). However, care must be

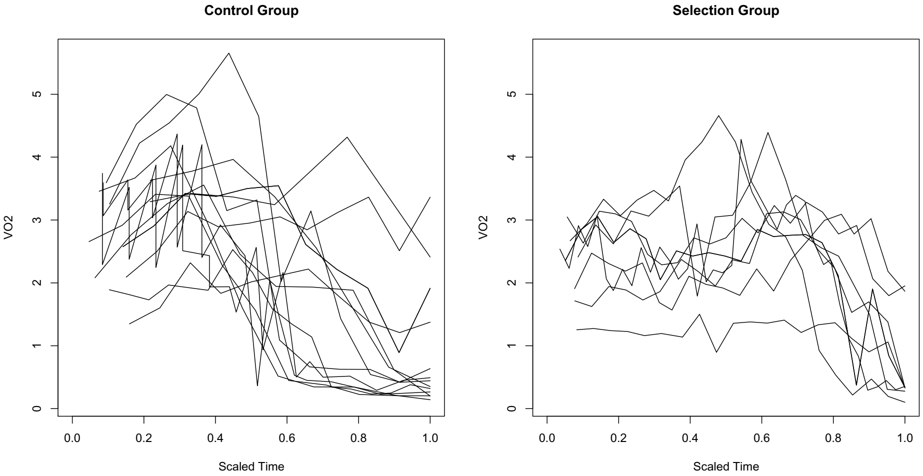

FIG 1. *Velocity of oxygen consumption (V02) of fruit flies, with lifetimes scaled to the unit interval. The selection group has been bred to withstand dry conditions. Data were collected by Donna G. Folk in the laboratory of Timothy J. Bradley at University of California Irvine.*

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 features 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 illustrate the problems and our solution using two data sets. One data set is taken from an evolutionary biology experiment involving fruit flies' metabolic characteristics measured at irregular time points. Figure 1 shows measurements of the metabolic variable velocity (i.e. rate) of oxygen consumption (VO2) of fruit flies placed in individual dessication (drying) chambers. The data were collected by Donna G. Folk in the laboratory of Timothy J. Bradley at University of California Irvine in order to study the evolution of physiological traits in fruit flies that were selectively bred to withstand dry conditions (Folk and Bradley [9]). The researchers provided already processed data giving the oxygen consumption rates. The right panel of Figure 1 shows measurements on eight fruit flies from the selectively bred group while the left panel shows measurements from the eighteen fruit flies in the control group, who were randomly bred. Repeated measurements were made on each individual until death, usually within one to two days. The researchers had compared lifetimes of the two groups, but were also interested in comparing the temporal patterns of the V02 measurements, that is, the shape of the curves. We therefore scaled lifetimes to the unit interval, to allow us to focus on shape and ignore lifetime. This re-scaling is a very simple type of curve registration (see Ramsay and Silverman [22]). Due

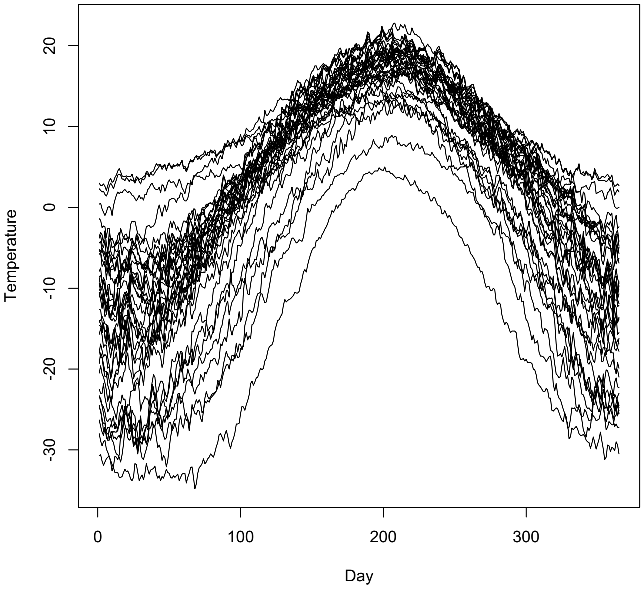

Fig 2. *Average daily temperatures (degrees Celsius) at 35 Canadian weather stations. Day 1 is January 1.*

to the design of the laboratory machinery, fruit flies were measured at different times. This inherent imbalance in the design with respect to the time variable is increased by the individual rescaling of time to the unit interval. Thus, the researchers cannot use pointwise analysis nor standard multivariate analysis to compare the two groups of insects. Our analysis of these data is presented in Section 5.

The other data we analyze were collected in a balanced design. In such a design, we can calculate unbiased estimates of variances and, through direct comparison of variance estimates, we can illustrate some of the problems that might arise with blind application of the standard smoothing/linear mixed model analysis. Also, the widespread availability of this data set, the *CanadianWeather* in the *fda* library in the statistical software package R, allows the reader to check our calculations. Figure 2 shows average daily temperatures recorded at 35 Canadian weather stations, where time $t = 1$ corresponds to January 1. In our analysis, we view the 35 stations as a random sample of all possible Canadian locations, with our goal being inference for the mean daily temperature in Canada. While this is a somewhat unusual goal for these data, it allows us to illustrate the main points of this paper. A more appropriate goal for these data, estimation of a particular station's "typical" weather curve, is not addressed here, although it is briefly mentioned.

We propose fitting data sets such as these in two steps: first, for speed of computation, we fit a standard smoothing-based linear mixed effects model. This yields our proposed estimate of the population curve. Then we use the output of this fit to calculate pointwise sandwich standard errors of our population curve estimate. We make two recommendations for the linear mixed effects model. We recommend that the model used for the individual-level random effects be a submodel of the population curve model, in the sense specified in Theorem 3.1. In addition, we recommend that the covariance structure for individual-level random effects be based on subject area knowledge. If such a rationale for the covariance is not available, we recommend using a "time-neutral" covariance structure, defined in Section 2. These recommendations will, in some cases, reduce or eliminate bias in the estimate of the population curve. While we do not propose anything further for bias correction of the estimate of the population curve, we do show via simulation studies that, if our recommendations are followed, resulting pointwise confidence intervals have good coverage properties (Section 6). This is further supported by our theoretical calculations in Section 4.2.

Figure 3 contains four estimates of the population mean temperature: one estimate is the daily mean of the temperatures. The other three estimates are from linear mixed effects models. They are calculated with the R library *nlme*, which uses restricted maximum likelihood (REML) estimates of variance components. For a discussion of restricted maximum likelihood estimators, see, for instance, Demidenko [3]. The specifics of our calculations are given in Section 3.4. The three mixed effects estimates are calculated using the same function space to model the population mean and the same function space to model the individual station effects, but the three estimates use different covariance structures for 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". The third mixed effects estimate uses a "time neutral" covariance structure. The first two mixed effects estimates do not track the pointwise average well, with the "forward" estimate deviating slightly from the pointwise average near day 365 and the "backward" estimate deviating near day 1. Clearly, both are poor estimates. The "time neutral" estimate tracks the mean fairly well.

In Figure 4, we compare several methods of computing standard errors. Throughout this paper, we plot error bands as plus or minus one standard error. Panels a) through c) show three of the four estimates from Figure 3 along with pointwise standard errors. Panel a) shows the pointwise average, with standard errors given by the pointwise standard deviation divided by $\sqrt{35}$. Panels b) and c) show model-based standard errors, that is, standard errors calculated using estimates of the assumed covariance structure of the linear mixed effects models. Panel b) corresponds to the "running backward" covariance structure and panel c) to the "running forward" model. Note the widening of the standard error bars in panels b) and c) to values that are much higher than the standard errors of the pointwise averages, standard error bars so wide that they are nonsensical. The "time neutral" covariance structure yields the model-based standard errors shown in panel d), where we also show the standard errors from panel a). Note that these two types of standard errors are similar in magnitude, but the "time neutral" covariance structure yields standard errors that are almost constant, which contradicts the fact that, in the winter, the variance between city temperatures is much larger than in the summer. Note that the standard errors in panel a) of Figure 4 will only be useful when, on each day, temperatures are recorded for a fairly large number of stations. In many applications, this is not the case and these straightforward standard errors cannot be readily calculated.

The observation that the assumed covariance structure of the random effects in a linear mixed effects model can impact mean 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. Unfortunately, currently published work on smoothing in mixed effects models has largely ignored this potential problem. Two notable exceptions are Brumback *et al* [1] and Djeundje and Currie [4], who both note the serious implications of reliance on the smoothing-induced covariance. The former suggest a computer-intensive bootstrap procedure to rectify the problems. The latter specifically note the problem of fanning as illustrated in Figure 4 and suggest a penalized approach to reduce fanning and also to reduce bias in the estimate of the population curve. This penalized approach should work well when the data are generated from a a specific covariance structure that matches the penalty. Section 4.2 contains further discussion of Djeundje and Currie's method, along with discussion of general issues of assessing variability in an estimated regression function. Our simulation studies in Section 6 indicate that, when the covariance structure does not match the penalty, Djeundje and Currie's variability bands severely underestimate the sampling variability in the estimates of the population mean.

In summary, from Figures 3 and 4, we see that the covariance structure assumed for the random effects can seriously impact both estimates and standard errors. Fortunately, under certain conditions, mean estimates are not effected by the assumed covariance structure (see Theorem 3.1). Appropriate standard errors can be easily constructed via sandwich estimators (see Sections 3.2 and 4.3). Figure 5 contains standard error bars calculated via our recommended methods. These standard error bars do not show the widening as seen in Figure 4 and can be calculated even when we do not have multiple observations on each day. See Liang and Zeger [18] for a discussion of variance misspecification in the context of generalized estimating equations. See also Szpiro *et al* [32], who propose a sandwich estimator-based solution in a slightly different context.

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. This assumption is unorthodox in classical linear mixed effects modelling but is common in smoothing (see, for instance, Ruppert *et al* [28]) and in Gaussian Process Regression, a popular technique in machine learning (Rasmussen and Williams, [23]). Section 5 contains analysis of the fruit fly data set shown in Figure 1. Section 6 contains the results of a simulation study.

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

$$Y_{ij} = f_i(t_{ij}) + \epsilon_{ij} \equiv \mu(t_{ij}) + g_i(t_{ij}) + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2), \text{ independent.} \tag{2.1}$$

We model the population curve μ via a set of $J_P + K_P$ basis functions $\{\psi_{Pj}, 1 \leq j \leq J_P, \phi_{Pk}, 1 \leq k \leq K_P\}$:

$$\mu(t) = \sum_{j=1}^{J_P} \boldsymbol{\beta}[j] \ \psi_{Pj}(t) \ + \ \sum_{k=1}^{K_P} \boldsymbol{\delta}[k] \ \phi_{Pk}(t). \tag{2.2}$$

We model individual i's deviation g_i via basis functions $\gamma_{Ij}, j = 1, \ldots, L$:

$$g_i(t) = \sum_{j=1}^{L} \boldsymbol{\theta}_i[j] \ \gamma_{Ij}(t). \tag{2.3}$$

In all of our analyses, $\boldsymbol{\beta} = (\boldsymbol{\beta}[1], \ldots, \boldsymbol{\beta}[J_P])'$ is a fixed effect and the $\boldsymbol{\theta}_i$'s, $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_i[1], \ldots, \boldsymbol{\theta}_i[L])'$, are random effects. We consider $\boldsymbol{\delta} = (\boldsymbol{\delta}[1], \ldots, \boldsymbol{\delta}[K_P])'$ as 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".

Using (2.1), (2.2) and (2.3), we can write the general model for subject i's response vector as

$$\begin{aligned} \mathbf{Y}_i &= (Y_{i1}, \ldots, Y_{in_i})' \\ &\equiv \mathbf{X}_{Pi} \, \boldsymbol{\beta} \ + \ \mathbf{Z}_{Pi} \, \boldsymbol{\delta} \ + \ \mathbf{C}_{Ii} \, \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i \\ &\equiv \mathbf{C}_i \boldsymbol{\theta} \ + \ \mathbf{C}_{Ii} \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i \\ &\equiv \mathbf{C}_i \boldsymbol{\theta} \ + \ \boldsymbol{\epsilon}_i^*. \end{aligned} \tag{2.4}$$

We assume that $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N$ (and $\boldsymbol{\delta}$, when $\boldsymbol{\delta}$ is assumed random) are independent, mean zero and normally distributed. When $\boldsymbol{\delta}$ is random, we will assume that $\text{var}(\boldsymbol{\delta})$ is an unknown positive constant times $\boldsymbol{\Sigma}_\delta$, a known covariance matrix. We denote $\text{var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_I$ and we consider two general models for $\boldsymbol{\Sigma}_I$, one with $\boldsymbol{\Sigma}_I$ unrestricted and the other with $\boldsymbol{\Sigma}_I = \boldsymbol{\Sigma}_I^R$ of some particular known form with just a few unknown parameters. The primary purpose of restricting the form of $\boldsymbol{\Sigma}_I$ is to facilitate computation in fitting the linear mixed effects model. In addition, some restricted forms have a connection to smoothing and so have become popular in mixed model smoothing.

We propose fitting the parameters of the model defined by (2.1), (2.2) and (2.3) using a restricted covariance for $\text{var}(\boldsymbol{\theta}_i)$. However, to avoid the problems in standard errors sometimes caused by mis-specifying the covariance structure, we propose using sandwich-type standard errors based on the unrestricted Σ_I.

While our results hold in general, in all of our examples and in our simulation studies, we assume that μ and the g_i's are linear splines with equally spaced knots and we suppose that $t \in (t_{\min}, t_{\max})$ where $t_{\min} = \min_{i,j}\{t_{ij}\}$ and $t_{\max} = \max_{i,j}\{t_{ij}\}$. We do not restrict the knots for modelling μ to be the same as the knots for modelling the g_i's. A linear spline on the interval $[t_{\min}, t_{\max}]$ with K interior knots, $\mathcal{K}_1 < \mathcal{K}_2 < \cdots < \mathcal{K}_K$, with $\mathcal{K}_1 > t_{\min}$ and $\mathcal{K}_K < t_{\max}$, is continuous and piecewise linear with the "pieces" defined on the subintervals determined by the knots. With K interior knots, we require $K+2$ basis functions and we consider the following bases. These bases were also considered by Djeundje & Currie [4].

1. The power basis with time "running forward": $\psi_1(t) \equiv 1$, $\psi_2(t) = t$ and $\phi_k(t) = (t - \mathcal{K}_k)_+$, $k = 1, \ldots, K$, where $u_+ = u$ if $u > 0$, $= 0$ otherwise.
2. The power basis with time "running backward": $\psi_1(t) \equiv 1$, $\psi_2(t) = 1 - t$ and $\phi_k(t) = (\mathcal{K}_k - t)_+$, $k = 1, \ldots, K$.
3. The "time neutral" Bspline basis, composed of triangular functions: let $\mathcal{K}_0 = t_{\min}$ and $\mathcal{K}_{K+1} = t_{\max}$. For $k = 2, \ldots, K+1$, define γ_{Ik}, the kth basis function, to be continuous and piecewise linear with support $[\mathcal{K}_{k-2}, \mathcal{K}_k]$, with $\gamma_{Ik}(\mathcal{K}_{k-2}) = \gamma_{Ik}(\mathcal{K}_k) = 0$ and $\gamma_{Ik}(\mathcal{K}_{k-1}) = 1$. Define γ_{I1} to be linear with support on $[\mathcal{K}_0, \mathcal{K}_1)$ with $\gamma_{I1}(\mathcal{K}_0) = 1$ and $\gamma_{I1}(\mathcal{K}_1) = 0$. Define $\gamma_{I,K+2}$ to be linear with support on $[\mathcal{K}_K, \mathcal{K}_{K+1}]$ with $\gamma_{I,K+2}(\mathcal{K}_K) = 0$ and $\gamma_{I,K+2}(\mathcal{K}_{K+1}) = 1$.

Splines are widely used for flexible fitting of functions. See, for instance, Ramsay and Silverman [22]. See Eilers and Marx [7] and Welham *et al* [34] for extensive discussion of the connections between the truncated power basis and a Bspline basis in penalized smoothing.

We consider four models for $f_i = \mu + g_i$ in (2.1) - (2.3).

A. μ is fixed, g_i is random: assume that $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are non-random and $\mathrm{var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_I$ is unrestricted.
B. μ is fixed, g_i is random but with modelled covariance: assume that $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are non-random and that $\mathrm{var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_I^R$, some restricted form.
C. μ is random, g_i is random: assume that $\boldsymbol{\beta}$ is fixed, that $\boldsymbol{\delta}$ is $N(0, \sigma_{P,C}^2 \boldsymbol{\Sigma}_\delta)$ for some known $\boldsymbol{\Sigma}_\delta$ and unknown $\sigma_{P,C}^2$, that $\boldsymbol{\delta}$ is independent of the $\boldsymbol{\theta}_i$'s and $\boldsymbol{\epsilon}_i$'s, and that $\mathrm{var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_I$ is unrestricted.
D. μ is random, g_i is random but with modelled covariance: assume that $\boldsymbol{\beta}$ is fixed, that $\boldsymbol{\delta}$ is $N(0, \sigma_P^2 \boldsymbol{\Sigma}_\delta)$ for some known $\boldsymbol{\Sigma}_\delta$ and unknown σ_P^2, that $\boldsymbol{\delta}$ is independent of the $\boldsymbol{\theta}_i$'s and $\boldsymbol{\epsilon}_i$'s, and that $\mathrm{var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_I^R$, some restricted form.

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

It is important to keep in mind that the model defined in (2.1)-(2.3) 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 this covariance that is crucial in analysis. Clearly, a change of basis without the accompanying change in assumptions about the covariance of the random coefficients may change the covariance of $\mu + g_i$. Notice that, in models A and C, the form for the covariance of the g_i's depends only on the space spanned by the basis functions and not on the choice of basis. In contrast, in models B and D, the basis for the g_i's and the form of $\mathbf{\Sigma}_I^R$ are both important in defining the model for the covariance of $\mu + g_i$.

When modelling μ in (2.2) in our data analyses and simulation study, we take the ψ_{Pj}'s and ϕ_{Pj}'s to be either the linear "time running forward" or "time running backward" basis functions. When μ is random, as in models C and D, we take $\mathbf{\Sigma}_\delta$ equal to the identity.

When modelling the g_i's in (2.3) in our data analyses and simulation studies, we use piecewise linear functions and we consider two restrictions on $\text{var}(\boldsymbol{\theta}_i)$. When using a power basis for g_i, with either time running forward or time running backward, we take

$$\text{var}(\boldsymbol{\theta}_i) \equiv \mathbf{\Sigma}_I^p = \begin{bmatrix} \mathbf{\Sigma}_\beta & 0 \\ 0 & \sigma_I^2 \mathbf{I} \end{bmatrix}. \tag{2.5}$$

Here $\mathbf{\Sigma}_\beta$ is the two by two unrestricted covariance matrix of $(\boldsymbol{\theta}_i[1], \boldsymbol{\theta}_i[2])$, the coefficients of $\gamma_{I1}(t) \equiv 1$ and $\gamma_{I2}(t) = t$, and $\sigma_I^2\mathrm{I}$ is the restricted covariance matrix of $(\boldsymbol{\theta}_i[3], \ldots, \boldsymbol{\theta}_i[K + 2])$, the coefficients of the power functions. This covariance structure was used by Durban *et al* [5]. When using the Bspline basis for g_i, we take

$$\text{var}(\boldsymbol{\theta}_i) = \sigma_I^2 \ \mathbf{I}. \tag{2.6}$$

We now discuss the covariance structures induced by these assumptions. For "time running forward" and with covariance structure as in model B or D with $\text{var}(\boldsymbol{\theta}_i) = \mathbf{\Sigma}_I^p$ as in (2.5),

$$\text{var}\left(\sum_{k=3}^{K_I+2} \boldsymbol{\theta}_i[k]\gamma_{Ik}(t) \right) = \sigma_I^2 \ \times \ \sum_{k=1}^{K_I} \left\{ (t - \mathcal{K}_k)_+ \right\}^2 ,$$

an increasing function of t. Similarly, for "time running backward" and with the same $\text{var}(\boldsymbol{\theta}_i)$, the variance of $\sum_{k=3}^{K_I+2} \boldsymbol{\theta}_i[k]\gamma_k(t)$ is decreasing in t. Assuming more variability in individual random effects at one end of the time scale than the other leads to the "drifting" of the estimate of μ in Figure 3 and to the unacceptable widening of the standard error bands in panels b) and c) of Figure 4.

Consider the covariance induced by the linear B-spline basis with equi-spaced knots, with the difference between knots equal to Δ. Suppose that either model B or model D holds, with $\text{var}(\boldsymbol{\theta}_i) = \sigma_I^2\mathbf{I}$. Letting $\mathcal{I}_k = [\mathcal{K}_{k-1}, \mathcal{K}_k \ )$, $k = 1, \ldots, K_I$, and $\mathcal{I}_{K_I+1} = [\mathcal{K}_{K_I}, \mathcal{K}_{K_I+1} \ ]$,

$$\text{var}(g_i(t)) = \frac{\sigma_I^2}{\Delta^2} \ \times \ \sum_{k=1}^{K_I+1} \{t \in \mathcal{I}_k\} \left[(t - \mathcal{K}_{k-1})^2 \ + (t - \mathcal{K}_k)^2 \right].$$

In particular, at the equi-spaced knots the variance is constant: $\text{var}(g_i(\mathcal{K}_k)) = \sigma_I^2$. On the interval between knots, the variance is quadratic with minimum value at the midpoint.

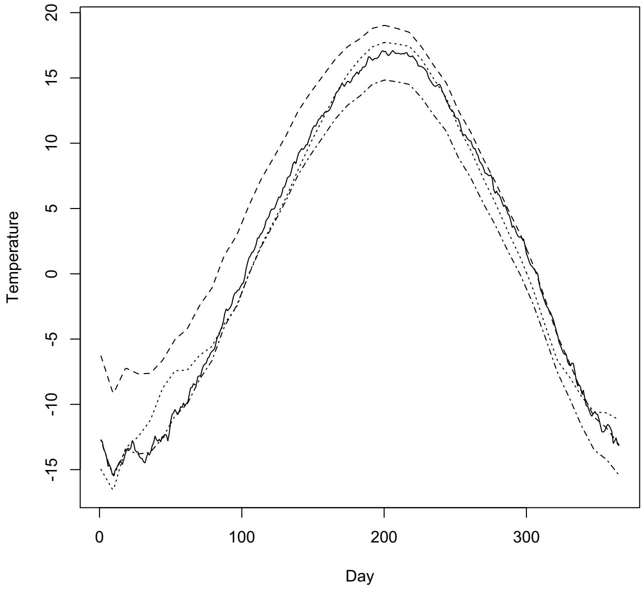

FIG 3. *The plot contains four estimates of μ, the typical weather curve, as described in Sections 2 and 3. The solid line is the pointwise average, the dotted line is the "time neutral" estimate, the dot-dashed line is the "time running forward" estimate, and the long dashed line is the "time running backward" estimate. The latter three estimates are based on piecewise linear functions with 41 equispaced interior population knots and 7 equispaced interior individual knots.*

We call the model for g_i assuming either (2.5) for the linear power basis or (2.6) for the linear Bspline basis a "smooth g_i" model. Likewise, we call our piecewise linear model for random μ a "smooth μ model" when $\boldsymbol{\Sigma}_\delta$ is the identity matrix. To see why, consider model B with power basis functions and $\text{var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_I^p$ as in (2.5). Write $\boldsymbol{\beta}_i = (\boldsymbol{\theta}_i[1], \boldsymbol{\theta}_i[2])'$ and $\boldsymbol{\delta}_i = (\boldsymbol{\theta}_i[3], \ldots, \boldsymbol{\theta}_i[K_I + 2])'$. By Henderson's justification [26], for fixed $\sigma_\epsilon^2, \sigma_I^2$ and $\boldsymbol{\Sigma}_\beta$, the best linear unbiased predictors (the BLUPs) of the f_i's in (2.1), (2.2) and (2.3) are obtained by minimizing

$$\frac{1}{\sigma_\epsilon^2} \sum_{i,j} \{Y_{ij} - f_i(t_{ij})\}^2 + \sum_i \boldsymbol{\beta}_i' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta}_i + \frac{1}{\sigma_I^2} \sum_i \boldsymbol{\delta}_i' \boldsymbol{\delta}_i$$

over $\boldsymbol{\beta}, \boldsymbol{\delta}$, the $\boldsymbol{\beta}_i$'s and the $\boldsymbol{\delta}_i$'s. That is, we minimize

$$\sum_i \left[\sum_j \{Y_{ij} - f_i(t_{ij})\}^2 + \sigma_\epsilon^2 \; \boldsymbol{\beta}_i' \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\beta}_i + \frac{\sigma_\epsilon^2}{\sigma_I^2} \; \boldsymbol{\delta}_i' \boldsymbol{\delta}_i \right].$$

The ith summand is a penalized least squares regression with penalties on $\boldsymbol{\beta}_i$ and $\boldsymbol{\delta}_i$. When we model g_i as a spline using the power basis, the penalty on $\boldsymbol{\delta}_i$ with "smoothing parameter" $\sigma_\epsilon^2/\sigma_I^2$ is one of those proposed by Eilers and Marx [6] for P-spline smoothing, and is recommended by Ruppert *et al* [28]. To

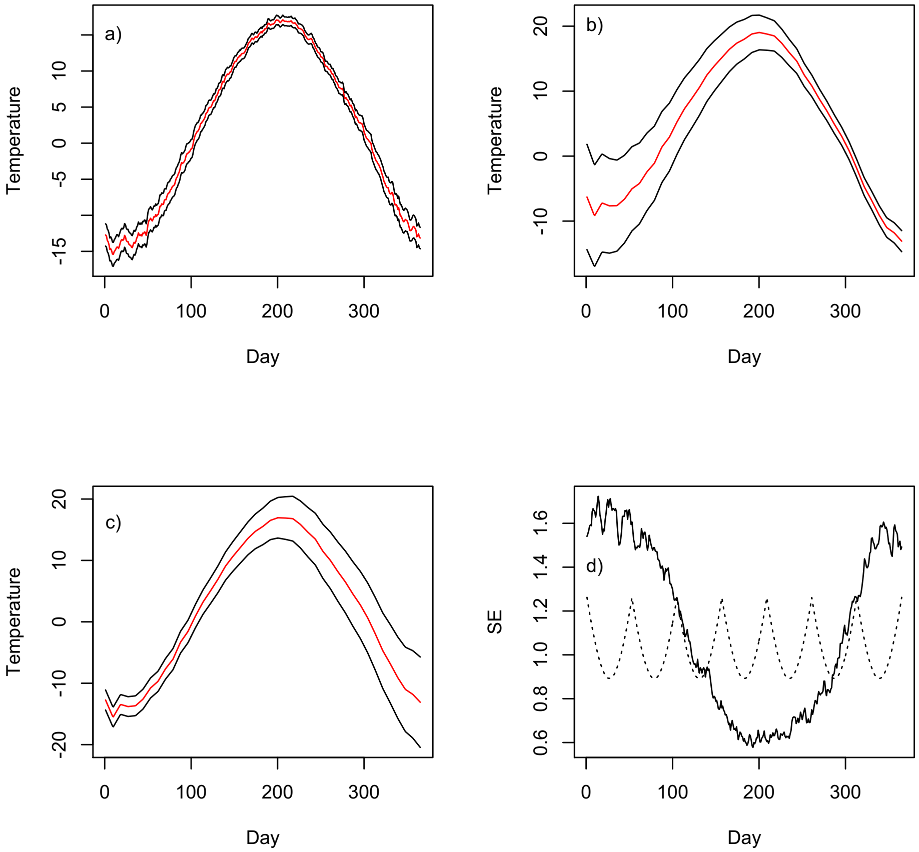

FIG 4. *Plotted are three of the estimates of μ shown in Figure 3, along with corresponding pointwise standard errors. In panel a) the estimate of μ is the pointwise average and the standard errors are the pointwise standard deviations of the 35 temperatures, divided by $\sqrt{35}$. Panels b) and c) show analogous information but with estimates using model B and standard errors calculated using equation (3.4) based on the restricted model B. Panel b) contains the "time running backward" estimate and panel c) contains the "time running forward". Panel d) compares the pointwise standard errors of panel a) to the model B-based standard errors using (3.4) for the "time neutral" estimate of μ (dashed line).*

consider the effect of the penalty, suppose that the penalty on $\boldsymbol{\delta}_i$ is large, that is, that $\sigma_\epsilon^2/\sigma_I^2$ is large. Then the BLUPs of the $\boldsymbol{\delta}_i$'s will be close to 0 and so the BLUP of g_i will be close to a line. We can thus say that the penalty on $\boldsymbol{\delta}_i$ shrinks the BLUP of g_i to a line, with the amount of shrinkage depending on $\sigma_\epsilon^2/\sigma_I^2$. Therefore, the effect of the penalty is similar to penalizing for large second divided differences of g_i and so the penalty is similar to the second derivative penalty that yields a smoothing spline estimate [12]. In fact, when the knots are equi-spaced with $\mathcal{K}_k = (k-1)\Delta$, then one can show that $\boldsymbol{\delta}_i[k]$ is exactly equal to the second divided difference $[g_i((k+2)\Delta) - 2g_i((k+1)\Delta) + g_i(k\Delta)]/\Delta$. See Eilers and Marx [7].

Similarly in model D with $\boldsymbol{\Sigma}_\delta$ equal to the identity matrix, by Henderson's justification, the predictors of the f_i's are based on minimizers of

$$\sum_i \left[\sum_j \{Y_{ij} - f_i(t_{ij})\}^2 + \sigma_\epsilon^2 \ \boldsymbol{\beta}_i' \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\beta}_i + \frac{\sigma_\epsilon^2}{\sigma_I^2} \ \boldsymbol{\delta}_i' \boldsymbol{\delta}_i \right] + \frac{\sigma_\epsilon^2}{\sigma_P^2} \boldsymbol{\delta}' \boldsymbol{\delta}.$$

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.

Conversely, we can rewrite many penalized least squares criteria in terms of the estimation criterion in a linear mixed effects model. Suppose that the f_i's have the basis expansions as in (2.1), (2.2) and (2.3) for some basis functions, not necessarily Bsplines or power functions. Suppose that our estimates of the $\boldsymbol{\beta}[j]$'s, $\boldsymbol{\delta}[k]$'s and $\boldsymbol{\theta}_i$'s are the minimizers of the penalized least squares criterion $\sum_{i,j} \{\mathbf{Y}_{ij} - f_i(t_{ij})\}^2 + \lambda \sum_i \boldsymbol{\theta}_i' \boldsymbol{\Omega} \boldsymbol{\theta}_i$ where $\boldsymbol{\Omega}$ is a known symmetric non-negative definite $L \times L$ matrix and λ is unknown. If $\boldsymbol{\Omega}$ is invertible, then we can use a linear mixed model in which the covariance matrix of the $\boldsymbol{\theta}_i$'s is restricted to be $\boldsymbol{\Sigma}_I^R \equiv (\lambda \boldsymbol{\Omega})^{-1}$. This is exactly model B or D for $\boldsymbol{\theta}_i$. If $\boldsymbol{\Omega}$ is not invertible, then we can reparameterize the problem to a linear mixed model that is similar to the model for $\boldsymbol{\theta}_i$ in model B or D, as follows. Let $\boldsymbol{\Omega} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}'$ be the eigendecomposition of $\boldsymbol{\Omega}$, with $\boldsymbol{\Lambda}$ diagonal with first k diagonal elements equal to 0. Partition $\mathbf{Q}$ as $\mathbf{Q} = [\mathbf{Q}_1 : \mathbf{Q}_2]$ where $\mathbf{Q}_1$ is L by k and $\mathbf{Q}_2$ is L by $L - k$. Define the reparameterization as $\boldsymbol{\theta}_i^{*\prime} = \boldsymbol{\theta}_i' \mathbf{Q} = (\boldsymbol{\theta}_i' \mathbf{Q}_1, \boldsymbol{\theta}_i' \mathbf{Q}_2) \equiv (\boldsymbol{\theta}_{i1}^{*}{}', \boldsymbol{\theta}_{i2}^{*}{}')$. Then, letting $\boldsymbol{\gamma}_I(t) = (\gamma_{I1}(t), \ldots, \gamma_{IL}(t))'$ and $\boldsymbol{\gamma}_I^*(t) = \mathbf{Q}' \boldsymbol{\gamma}_I(t)$,

$$f_i(t) = \mu(t) + \boldsymbol{\theta}_i' \boldsymbol{\gamma}_I(t) = \mu(t) + \boldsymbol{\theta}_i^{*\prime} \boldsymbol{\gamma}_I^*(t)$$

and the penalized sum of squares criterion becomes

$$\sum_{i,j} \{Y_{ij} - f_i(t_{ij})\}^2 + \lambda \sum_i \boldsymbol{\theta}_{i2}^{*}{}' \boldsymbol{\Lambda}^* \ \boldsymbol{\theta}_{i2}^*$$

with $\boldsymbol{\Lambda}^*$ a diagonal matrix containing the non-zero eigenvalues of $\boldsymbol{\Omega}$. Thus, the penalized sum of squares criterion is the same as a linear mixed effects model criterion where the $\boldsymbol{\theta}_{i1}^*$'s are fixed effects and the $\boldsymbol{\theta}_{i2}^*$'s are random effects with covariance matrix $(\lambda \boldsymbol{\Lambda}^*)^{-1}$. Some restrictions on the $\boldsymbol{\theta}_{i2}^*$'s may be needed, to ensure identifiability. A slight modification leads us directly to model B or D and avoids the problem of non-identifiability: suppose that the $\boldsymbol{\theta}_{ij}^*$'s are independent with means equal to the zero vector and with the covariance of $\boldsymbol{\theta}_{i1}^* = \boldsymbol{\Sigma}_I$, unrestricted, and the covariance matrix of the $\boldsymbol{\theta}_{i2}^*$'s equal to the restricted matrix $(\lambda \boldsymbol{\Lambda}^*)^{-1}$.

Special cases of models A and B have been considered elsewhere. In the animal breeding literature see, for instance, Huisman *et al* [15] and Meyer [19]. Rice and Wu [24] consider model A in a medical context. In these models where μ is fixed, the only smoothness of μ comes from the smoothness of the basis functions. We do not include a penalty term, as in the random μ case, where the penalty term forces further smoothness of μ. Our model for fixed μ is a standard model in spline regression. See, for instance, Stone *et al* [30] or Friedman [10].

Durban *et al* [5] analyzed growth data by reformulating a penalized spline approach with μ fixed into an analysis using the restricted model D, modelling μ as random, using the degree 1 power basis. They considered a range of models for g_i: g_i equal to a random intercept, a random line with unspecified covariance matrix for the slope and intercept, and g_i piecewise linear with the same knots as μ, with $\operatorname{cov}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_I^P$ as in (2.5). 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.

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 maximum 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 BLUP of μ and then use the unrestricted model C for easy-to-compute pointwise prediction bands. Throughout, we ignore any model-based bias, that is, we assume that (2.1) - (2.3) are exact.

In this article, we only consider inference for μ – we do not study prediction of the individual effects. But this prediction is straightforward, predicting g_i by substituting our estimates of $\boldsymbol{\theta}_i$ into (2.3).

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. For instance, the model with unrestricted $\boldsymbol{\Sigma}_I$ caused us computational problems when the dimension of $\boldsymbol{\Sigma}_I$ wasn't small. In our analysis of the fruit fly data, we found that R's *lme* would not converge when we used five or more knots in modelling g_i, that is, when the dimension of the unrestricted $\boldsymbol{\Sigma}_I$ was 7 by 7 or larger. The convergence was extremely slow when $\boldsymbol{\Sigma}_I$ was 6 by 6. This held for both the power basis and the more computationally stable Bspline basis. In contrast, calculation of estimates with the restricted covariances always converged and converged very quickly, even when ten knots were used to model g_i. The small number of parameters in the restricted model can turn a computationally impossible analysis into a possible analysis.

The techniques we use in our calculations are not new. Many calculations in the linear mixed effects model appear in Demidenko [3] and Ruppert *et al* [28]. 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 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 $\mathrm{E}(\mu)$.

3. Non-random $\boldsymbol{\mu}$

3.1. Estimation of $\boldsymbol{\mu}$ under model B

Consider data generated according to (2.1) through (2.4) under either model A or B. Since $\boldsymbol{\delta}$ 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 $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\delta}')'$ under the assumptions of model B is the generalized least squares estimate, minimizing

$$\sum (\mathbf{Y}_i - \mathbf{C}_i\boldsymbol{\theta})'(\boldsymbol{\Sigma}_i^{*R})^{-1}(\mathbf{Y}_i - \mathbf{C}_i\boldsymbol{\theta}),$$

with $\boldsymbol{\Sigma}_i^{*R}$ denoting the variance of $\boldsymbol{\epsilon}_i^*$ under the restricted model B,

$$\boldsymbol{\Sigma}_i^{*R} = \mathbf{C}_{Ii}\boldsymbol{\Sigma}_I^R\mathbf{C}_{Ii}' \ + \ \sigma_\epsilon^2\mathbf{I}. \tag{3.1}$$

Therefore the estimator of $\boldsymbol{\theta}$ for known variance parameters is

$$\begin{aligned}\tilde{\boldsymbol{\theta}} = \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\delta}} \end{pmatrix} &= \textstyle\sum_i \left[\left(\sum_j \mathbf{C}_j'(\boldsymbol{\Sigma}_j^{*R})^{-1}\mathbf{C}_j \right)^{-1} \mathbf{C}_i'(\boldsymbol{\Sigma}_i^{*R})^{-1} \right] \mathbf{Y}_i \\ &\equiv \textstyle\sum \mathbf{H}_i(\boldsymbol{\Sigma}_1^{*R}, \ldots, \boldsymbol{\Sigma}_N^{*R})\ \mathbf{Y}_i \equiv \sum \mathbf{H}_i\mathbf{Y}_i. \end{aligned} \tag{3.2}$$

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

A linear mixed effects model fit of model B yields (restricted) maximum likelihood variance estimators $\hat{\boldsymbol{\Sigma}}_I^R$ and $\hat{\sigma}_\epsilon^2$, and thus yields $\hat{\mathbf{H}}_i = \mathbf{H}_i(\hat{\boldsymbol{\Sigma}}_1^{*R}, \ldots, \hat{\boldsymbol{\Sigma}}_N^{*R})$, an estimator of $\mathbf{H}_i$. The (restricted) maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is then equal to $\sum \hat{\mathbf{H}}_i\mathbf{Y}_i$ and the (restricted) maximum likelihood estimator of $\mu(t)$, $\hat{\mu}(t)$, is gotten in the obvious way from $\hat{\boldsymbol{\theta}}$. The method also provides estimators, $\hat{\boldsymbol{\theta}}_i$ $i = 1, \ldots, N$, of the BLUPs of the $\boldsymbol{\theta}_i$'s. These estimators, commonly called the estimated best linear unbiased predictors or EBLUPs, are gotten by substituting covariance estimates into the expressions for the best linear unbiased predictors. We use a tilde when an estimator or predictor is based on known covariance parameters and a hat when estimated covariance parameters are used. In particular, we use a tilde to denote a BLUP and a hat to denote an EBLUP.

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 $\hat{\mu}$, but acknowledge that doing so is likely to produce standard errors that may be small. This variability could be accounted for by, e.g., methods of Kackar and Harville [16].

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

$$\text{var}(\tilde{\boldsymbol{\theta}}) = \sum \mathbf{H}_i \ \text{var}(\boldsymbol{\epsilon}_i^*) \ \mathbf{H}_i'. \tag{3.3}$$

Keep in mind that $\mathbf{H}_i$ contains model B variance parameters while $\text{var}(\boldsymbol{\epsilon}_i^*)$ contains model A variance parameters.

If the restricted model B holds, then the covariance matrix of $\boldsymbol{\epsilon}_i^*$ is equal to $\boldsymbol{\Sigma}_i^{*R}$ as in (3.1), and $\text{var}(\tilde{\boldsymbol{\theta}})$ simplifies to $(\sum \mathbf{C}_i'(\boldsymbol{\Sigma}_i^{*R})^{-1}\mathbf{C}_i)^{-1}$, which we can estimate by

$$\widehat{\text{var}}_{\text{B}}(\tilde{\boldsymbol{\theta}}) = \left(\sum \mathbf{C}_i'(\hat{\boldsymbol{\Sigma}}_i^{*R})^{-1}\mathbf{C}_i\right)^{-1} \tag{3.4}$$

where $\hat{\boldsymbol{\Sigma}}_i^{*R}$ is obtained by fitting model B. Expression (3.4) was used to calculate the standard errors shown in panels b), c) and d) of Figure 4. Clearly, we do not want to use the model-based covariances which produced panels b) and c), as doing so gives unrealistic standard errors for our estimate of μ. The problems in panel d) are not as striking, but we still do see the effects of the assumed "time neutral" covariance structure.

To construct better standard errors, we require an estimator of $\text{var}(\boldsymbol{\epsilon}_i^*)$ in (3.3) that is valid under model A. The variance of $\boldsymbol{\epsilon}_i^*$ under model A is

$$\text{var}(\boldsymbol{\epsilon}_i^*) = \mathbf{C}_{Ii}\boldsymbol{\Sigma}_I\mathbf{C}_{Ii}' + \sigma_\epsilon^2\mathbf{I}$$

and thus we require an estimator of σ_ϵ^2 and an unrestricted estimator of $\boldsymbol{\Sigma}_I = \text{var}(\boldsymbol{\theta}_i)$. We estimate $\boldsymbol{\Sigma}_I$ by $\mathbf{S}_{\hat{\theta}}$, the sample covariance matrix of the $\hat{\boldsymbol{\theta}}_i$'s, our estimators of the BLUPs from fitting model B:

$$\mathbf{S}_{\hat{\theta}} = \frac{1}{N-1}\sum_i \left(\hat{\boldsymbol{\theta}}_i - \sum \hat{\boldsymbol{\theta}}_j/N\right)\left(\hat{\boldsymbol{\theta}}_i - \sum \hat{\boldsymbol{\theta}}_j/N\right)'. \tag{3.5}$$

We estimate σ_ϵ^2 by

$$\hat{\sigma}_\epsilon^2 = \frac{1}{\text{df}} \sum_i (\mathbf{Y}_i - \mathbf{C}_i\hat{\boldsymbol{\theta}} - \mathbf{C}_{Ii}\hat{\boldsymbol{\theta}}_i)'(\mathbf{Y}_i - \mathbf{C}_i\hat{\boldsymbol{\theta}} - \mathbf{C}_{Ii}\hat{\boldsymbol{\theta}}_i). \tag{3.6}$$

where

$$\text{df} = \sum_1^N n_i - \text{length}(\boldsymbol{\theta}) + \text{df}_{\text{adj}} - \sum_1^N \text{length}(\boldsymbol{\theta}_i)$$

and df_{adj} corrects for parameter over-counting. For instance, when using power bases at both the population level and the individual level, $\text{df}_{\text{adj}} = 2 +$ the number of common population and individual interior knots. In the special case 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 $n = n_i$, $K =$ the number of interior knots, $\text{df} = Nn - N(K+2)$. The resulting estimates of σ_ϵ^2 and $\text{var}(\hat{\boldsymbol{\theta}})$ agree with Demidenko's (pp 61 ff [3]).

Other estimates of σ_ϵ^2 and $\boldsymbol{\Sigma}_I = \text{var}(\boldsymbol{\theta}_i)$ are possible. Since $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_i$ are shrinkage estimators one could adjust (3.5) and (3.6) by adjusting the degrees of freedom to account for shrinkage. See Hodges and Sargent [14] for a discussion of a variety of suggestions. On the other hand, a sensible estimate of σ_ϵ^2 can be gotten by ordinary least squares, with no shrinkage in estimation of any basis function coefficient. Alternatively, we might consider a method of moments approach, as follows. Let $\mathbf{S}_{\tilde{\theta}_i}$ and $\tilde{\sigma}_\epsilon^2$ be the analogues of (3.5) and (3.6) but with

$\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_i$ calculated using the true variance parameters. The expected values of $\mathbf{S}_{\tilde{\theta}}$ and $\tilde{\sigma}_\epsilon^2$ are easily calculated and are linear in both σ_ϵ^2 and $\boldsymbol{\Sigma}_I$. To estimate σ_ϵ^2 and $\boldsymbol{\Sigma}_I$, set $\mathrm{E}(\mathbf{S}_{\tilde{\theta}}) = \mathbf{S}_{\hat{\theta}}$, $\mathrm{E}(\tilde{\sigma}_\epsilon^2) = \hat{\sigma}_\epsilon^2$ and solve. We have not made a careful investigation of these possibilities.

Our estimator of $\mathrm{var}(\boldsymbol{\epsilon}_i^*)$ is then $\widehat{\mathrm{var}}(\boldsymbol{\epsilon}_i^*) = \mathbf{C}_{Ii}\mathbf{S}_{\hat{\boldsymbol{\theta}}}\mathbf{C}_{Ii}' + \hat{\sigma}_\epsilon^2\mathbf{I}$, and we use this in (3.3) to estimate the variance of $\hat{\theta}$ under model A:

$$\widehat{\mathrm{var}}_{\mathrm{A}}(\hat{\boldsymbol{\theta}}) = \sum \hat{\mathbf{H}}_i \, (\mathbf{C}_{Ii}\mathbf{S}_{\hat{\boldsymbol{\theta}}}\mathbf{C}_{Ii}' + \hat{\sigma}_\epsilon^2\mathbf{I}) \, \hat{\mathbf{H}}_i'. \tag{3.7}$$

The variance estimator in (3.7) relies on the assumed form of the variance of $\boldsymbol{\epsilon}_i^*$ given in model A, and so we call a standard error for $\hat{\mu}$ based on this variance estimator a *half sandwich* standard error. If this form of the variance is suspect, if, for instance, the covariance matrix of $\boldsymbol{\epsilon}_i$ is not a constant times the identity, then the following general sandwich estimator of the variance of $\hat{\boldsymbol{\theta}}$ might be preferred:

$$\widehat{\mathrm{var}}_s(\hat{\boldsymbol{\theta}}) = \sum \hat{\mathbf{H}}_i \, (\mathbf{Y}_i - \mathbf{C}_i\hat{\boldsymbol{\theta}})(\mathbf{Y}_i - \mathbf{C}_i\hat{\theta})' \, \hat{\mathbf{H}}_i'. \tag{3.8}$$

We call a standard error for $\hat{\mu}$ based on this variance estimator a *full sandwich* standard error. Robert-Granié, Heude and Foulle [25] 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 (2.4) balanced if $\mathbf{C}_i \equiv \mathbf{C}$ and $\mathbf{C}_{Ii} \equiv \mathbf{C}_I$. While many designs are not balanced, considering the balanced design can provide us with insight into linear mixed effects analysis. For a balanced design, the variance of $\boldsymbol{\epsilon}_i^*$ does not depend on i. In this case, the model B estimator of $\boldsymbol{\theta}$ in (3.2) simplifies to $\tilde{\boldsymbol{\theta}} = \{\mathbf{C}'[\boldsymbol{\Sigma}_1^{*R}]^{-1}\mathbf{C}\}^{-1}\mathbf{C}'[\boldsymbol{\Sigma}_1^{*R}]^{-1}\bar{\mathbf{Y}}$ which only depends on the data via $\bar{\mathbf{Y}} = \sum \mathbf{Y}_i/N$. Since $\hat{\mathbf{H}}_i \equiv \hat{\mathbf{H}} = N^{-1}\{\mathbf{C}'[\boldsymbol{\Sigma}_1^{*R}]^{-1}\mathbf{C}\}^{-1}\mathbf{C}'[\boldsymbol{\Sigma}_1^{*R}]^{-1}$ does not depend on i, the sandwich variance estimator in (3.8) simplifies as follows: using the facts that $\hat{\mathbf{H}}\mathbf{C} = N^{-1}\mathbf{I}$ and $\hat{\boldsymbol{\theta}} = \hat{\mathbf{H}}N\bar{\mathbf{Y}}$, we see that $\hat{\mathbf{H}}(\mathbf{Y}_i - \mathbf{C}\hat{\boldsymbol{\theta}}) = \hat{\mathbf{H}}(\mathbf{Y}_i - \bar{\mathbf{Y}})$ and so $\widehat{\mathrm{var}}_s(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{H}} \, \sum(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})' \, \hat{\mathbf{H}}'$. Thus, we see that estimating the variance of the $\boldsymbol{\epsilon}_i^*$'s in (3.3) via model B based residuals is equivalent to estimating the variance using the sample variance of the $\mathbf{Y}_i$'s.

Suppose that the design is balanced and that model (2.4) holds with $\boldsymbol{\Sigma}_\theta$ denoting the possibly restricted covariance matrix of $\boldsymbol{\theta}_i$. Then the maximum likelihood estimator of $\boldsymbol{\theta}$ when variance parameters are known is the generalized least squares estimate

$$\tilde{\boldsymbol{\theta}}_G = \{\mathbf{C}'[\mathrm{var}(\boldsymbol{\epsilon}_i^*)]^{-1}\mathbf{C}\}^{-1}\mathbf{C}'[\mathrm{var}(\boldsymbol{\epsilon}_i^*)]^{-1}\bar{\mathbf{Y}} \tag{3.9}$$

with $\mathrm{var}(\boldsymbol{\epsilon}_i^*) = \mathbf{C}_I\boldsymbol{\Sigma}_\theta\mathbf{C}_I' + \sigma_\epsilon^2\mathbf{I}$.

Under an additional condition on $\mathbf{C}$ and $\mathbf{C}_I$, given in the following Theorem, the estimator $\tilde{\boldsymbol{\theta}}_G$ is equal to the ordinary least squares estimator and thus does not depend on the assumed covariance matrix. Under the same condition explicit formulae for the maximum likelihood and restricted maximum likelihood estimators for Model A covariance parameters can be given; see the end of this section. The proof of the Theorem appears in the Appendix. By considering the proof, we see that Theorem 3.1 holds for general matrices $\mathbf{C}$ and $\mathbf{C}_I$, that is, the Theorem does not require that $\mathbf{C}$ and $\mathbf{C}_I$ depend on basis functions.

Theorem 3.1. *Suppose that model (2.4) holds with $\mathbf{C}_i \equiv \mathbf{C}$ and $\mathbf{C}_{Ii} \equiv \mathbf{C}_I$. If the column space of $\mathbf{C}_I$ is contained in the column space of $\mathbf{C}$, then the $\tilde{\boldsymbol{\theta}}_G$ in (3.9) and $\hat{\boldsymbol{\theta}}$, the corresponding maximum likelihood estimator when variance parameters are unknown, are equal to the ordinary least squares estimate $\hat{\boldsymbol{\theta}}_O \equiv (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\bar{\mathbf{Y}}$.*

The conditions of the Theorem relate to the choice of function spaces in modelling in (2.1) – (2.4). Suppose that the design is balanced, that the function space modelling the g_i's is a subspace of the function space modelling μ and that $\boldsymbol{\theta}$ is non-random. Then the column space of $\mathbf{C}_I$ is contained in the column space of $\mathbf{C}$ and the Theorem states that the MLE for $\boldsymbol{\theta}$ is the ordinary least squares estimate, not depending on the covariance structure of the $\boldsymbol{\epsilon}_i^*$'s. Translating this to our models with piecewise linear functions, if the knots for the g_i's are a subset of the knots for the population curve μ, $\tilde{\boldsymbol{\theta}}$ in (3.2) simplifies to the ordinary least squares estimate of $\boldsymbol{\theta}$ and so $\tilde{\mu}$ does not depend on the specific basis functions or on the assumed covariance structure of the station-specific random effects. Consequently our "forward time", "backward time" and "time neutral" estimates of μ are the same.

Demidenko [3] considers model (2.4) but with general matrices $\mathbf{C}_i$ and $\mathbf{C}_{Ii}$ not necessarily derived from basis functions. He establishes the conclusion of Theorem 3.1 in a balanced design under the stronger condition $\mathbf{C} = \mathbf{C}_I$; he then gives, under this same condition, explicit formulae for the maximum likelihood and restricted maximum likelihood estimates of $\boldsymbol{\Sigma}_I$ and σ_ϵ^2 when the population $\boldsymbol{\delta}$ is not random and when $\boldsymbol{\Sigma}_I$ is unrestricted (pp 61ff). Careful reading of his proof shows that these formulae remain valid whenever generalized least squares reduces to ordinary least squares. Thus under the conditions of Theorem 3.1 we find that the restricted and unrestricted maximum likelihood estimators of σ_ϵ^2 under model A are equal and given by

$$\hat{\sigma}_\epsilon^2 = \sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{C}\hat{\boldsymbol{\theta}}_0)' \left\{ \mathbf{I} - \mathbf{C}_I(\mathbf{C}_I'\mathbf{C}_I)^{-1}\mathbf{C}_I' \right\} (\mathbf{Y}_i - \mathbf{C}\hat{\boldsymbol{\theta}}_0) / \left\{ N(n-L) \right\}$$

where $\mathbf{C}_I$ is n by L. The maximum likelihood estimator of $\boldsymbol{\Sigma}_I$ is

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

where

$$\mathbf{S} = \sum (\mathbf{Y}_i - \mathbf{C}\hat{\boldsymbol{\theta}}_O)(\mathbf{Y}_i - \mathbf{C}\hat{\boldsymbol{\theta}}_O)'/N.$$

To get the restricted maximum likelihood estimator of $\boldsymbol{\Sigma}_I$ replace the N in the denominator of $\mathbf{S}$ by $N-1$. See Demidenko, p 63 [3].

3.4. Temperature data analysis with μ nonrandom

In all of the temperature data analyses, we model functions as splines of degree $p = 1$ using either the power basis or the B-spline representation, as described in Section 2. Knots are equi-spaced with equal distances from the "edges" of 1 and 365: a sequence of K interior knots is constructed with $\mathcal{K}_j = 1 + 364\, j/(K + 1)$, $j = 1, \ldots, K$.

For the analysis of Figure 3, we use 41 interior population knots and 7 interior individual knots, all equispaced, so the model doesn't satisfy the conditions of Theorem 1. As we see, our "time running forward" and "time running backward" and "time neutral" estimates of μ are not the same.

The unacceptably widening pointwise standard error bands in panels b) and c) of Figure 4 were computed using the model-based variance estimate in (3.4). As noted in Section 2, this widening is caused by the covariance assumptions of model B. In panel d), we see that the model-based pointwise standard errors using the "time neutral" covariance structure do not show sufficient heteroscedasticity, also as explained in Section 2.

In Figure 5, estimation of μ involved 41 interior population knots and 6 interior individual knots. When using these knots, by Theorem 3.1, the "time running forward", the "time running backward" and the "time neutral" estimates of μ are all the same. Indeed, when using these knots, the covariance structure of the random effects does not influence the estimate of μ, as the estimate is the ordinary least squares estimate. Panel a) shows the estimate of μ, along with the pointwise averages of the temperatures for comparison. Panel b) displays standard errors based on the sandwich estimators (3.7) and (3.8), using the "time neutral" covariance structure. Panel c) shows that the difference between sandwich standard errors using the "time running forward" and using the "time neutral" covariance structures is negligible. Sandwich estimation appears to have greatly reduced the bias in the standard error estimates, a bias caused by model mis-specification.

We have not plotted the model-based standard errors for this choice of knots, but they exhibit the same undesirable behaviour shown in Figure 4. Indeed, model-based standard errors exhibit undesirable behavior for a wide range of choices of number of knots.

It is important to remember that both model-based and sandwich standard errors for $\hat{\mu}$ are affected by the assumed covariance structure. For instance, 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 time" estimator will, in general, be different from those of the "backward time" estimator. Even sandwich standard errors can be affected by the covariance structure: in plots not shown here, in the "time running forward" and "time running backward" models, the standard errors using (3.7) or (3.8) also exhibit fanning, albeit mild, if the conditions of Theorem 3.1 do not hold. Even if the knot conditions implied by the assumptions of the Theorem do hold, some amount of fanning may occur (see Figure 8, where the fly data are analyzed assuming that μ is random).

In an analysis not shown here, we fit model B with the individual random effect g_i equal to a line with the assumed covariance matrix of the slope and

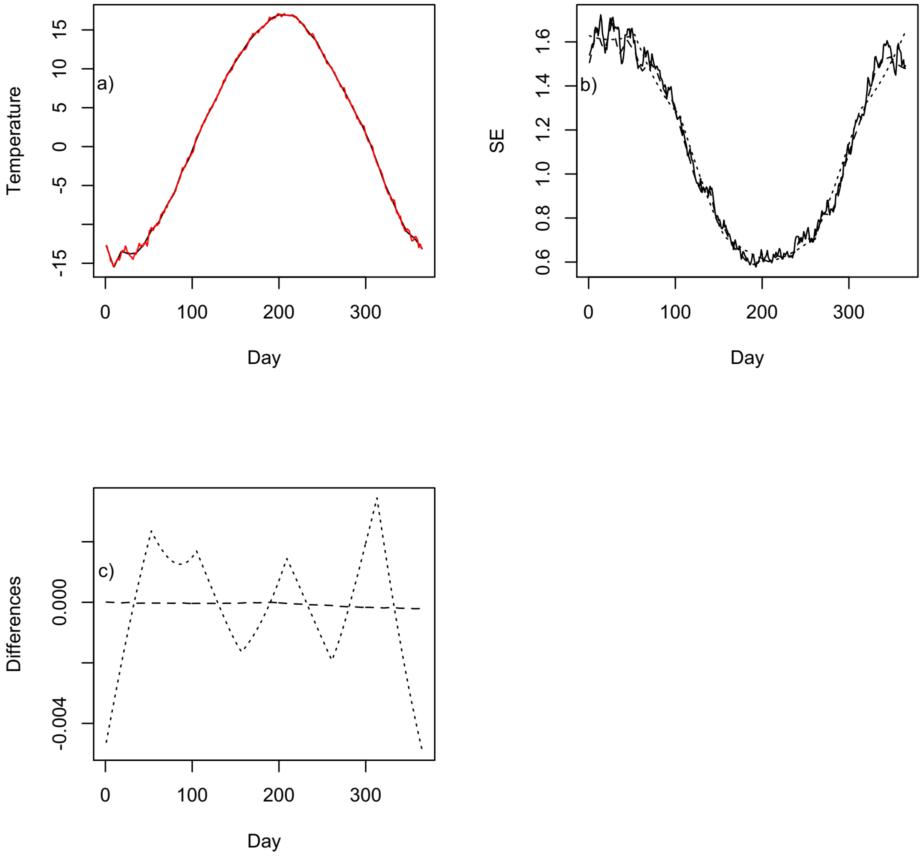

FIG 5. *Analysis of the Canadian weather data using the methods of Section 3, with 41 interior population knots and 6 interior individual knots. Panel a) shows the mixed model estimate of μ, which by Theorem 3.1 does not depend on the covariance structure of the individual random effects. Panel a) also shows pointwise averages (red line). Panel b) contains pointwise sandwich standard errors of $\hat{\mu}$ using the "time neutral" covariance structure for individual random effects. The dotted line is the half-sandwich standard error based on (3.7) and the dashed line is the full sandwich standard error based on (3.8). Panel c) shows the difference between sandwich standard errors using the "time running forward" and using the "time neutral" covariance structures. The dotted line is the difference for half-sandwich standard errors and the dashed line is the difference for full sandwich standard errors.*

intercept unrestricted. The widening of the standard error bands didn't occur, a result that agrees with the analysis in Smith and Wand [29].

4. Random μ

4.1. Prediction of μ under model D

Suppose data are generated according to (2.1) through (2.4) under either model C or D, so that the population effect $\boldsymbol{\delta}$ is random. Under either model we write

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_N \end{pmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \vdots \\ \mathbf{C}_N \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\delta} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1^* \\ \boldsymbol{\epsilon}_2^* \\ \vdots \\ \boldsymbol{\epsilon}_N^* \end{pmatrix} \equiv \boldsymbol{\mathcal{C}}\,\boldsymbol{\theta} + \boldsymbol{\epsilon}^*. \tag{4.1}$$

To calculate the estimators of $\boldsymbol{\beta}$ and the BLUP of $\boldsymbol{\delta}$, we assume that the restricted model D holds. The variance of $\boldsymbol{\epsilon}_i^*$ under model D is $\boldsymbol{\Sigma}_i^{*R}$ as defined in (3.1). Let $\mathbf{0}_{J,K}$ denote a $J \times K$ matrix of zeroes and let

$$\mathbf{I}_\delta = \begin{bmatrix} \mathbf{0}_{J_P,J_P} & \mathbf{0}_{J_P,K_P} \\ \mathbf{0}_{K_P,J_P} & \boldsymbol{\Sigma}_\delta^{-1} \end{bmatrix},$$

$$\boldsymbol{\mathcal{S}} = \mathrm{diag}(\boldsymbol{\Sigma}_1^{*R}, \boldsymbol{\Sigma}_2^{*R}, \ldots, \boldsymbol{\Sigma}_N^{*R})$$

and

$$\mathbf{A} = \boldsymbol{\mathcal{C}}'\boldsymbol{\mathcal{S}}^{-1}\boldsymbol{\mathcal{C}} + \frac{1}{\sigma_P^2}\mathbf{I}_\delta = \sum \boldsymbol{\mathcal{C}}_i'[\boldsymbol{\Sigma}_i^{*R}]^{-1}\boldsymbol{\mathcal{C}}_i + \frac{1}{\sigma_P^2}\,\mathbf{I}_\delta. \tag{4.2}$$

Then, for known variance parameters, under model D, $\tilde{\boldsymbol{\beta}}$, the maximum likelihood estimator of $\boldsymbol{\beta}$, and $\tilde{\boldsymbol{\delta}}$, the BLUP of $\boldsymbol{\delta}$, can be found via Henderson's justification (see Robinson [26]), as the minimizers of

$$(\mathbf{Y}-\boldsymbol{\mathcal{C}}\boldsymbol{\theta})'\boldsymbol{\mathcal{S}}^{-1}(\mathbf{Y}-\boldsymbol{\mathcal{C}}\boldsymbol{\theta}) \;+\; \frac{1}{\sigma_P^2}\boldsymbol{\delta}'\boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\delta} = (\mathbf{Y}-\boldsymbol{\mathcal{C}}\boldsymbol{\theta})'\boldsymbol{\mathcal{S}}^{-1}(\mathbf{Y}-\boldsymbol{\mathcal{C}}\boldsymbol{\theta}) \;+\; \frac{1}{\sigma_P^2}\boldsymbol{\theta}'\mathbf{I}_\delta\boldsymbol{\theta}$$

and so

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= \mathbf{A}^{-1}\boldsymbol{\mathcal{C}}'\boldsymbol{\mathcal{S}}^{-1}\mathbf{Y} = \mathbf{A}^{-1}\;\sum \mathbf{C}_i'(\boldsymbol{\Sigma}_i^{*R})^{-1}\mathbf{Y}_i \\ &\equiv \sum \boldsymbol{\mathcal{H}}_i(\boldsymbol{\Sigma}_1^{*R}, \ldots, \boldsymbol{\Sigma}_N^{*R}, \sigma_P^2)\mathbf{Y}_i \equiv \sum \boldsymbol{\mathcal{H}}_i\mathbf{Y}_i. \end{aligned} \tag{4.3}$$

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

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

A linear mixed effects model fit of model D yields (restricted) maximum likelihood estimators of σ_P^2, $\boldsymbol{\Sigma}_I^R$, and σ_ϵ^2, and thus estimators of $\boldsymbol{\mathcal{H}}_i$, denoted $\hat{\boldsymbol{\mathcal{H}}}_i$. The fit also produces estimators of the BLUPs of $\boldsymbol{\theta}$ and the $\boldsymbol{\theta}_i$'s. The estimator of the BLUP of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \sum \hat{\boldsymbol{\mathcal{H}}}_i\mathbf{Y}_i$. The predictor of $\mu(t)$, denoted $\hat{\mu}(t)$, is gotten in the obvious way from $\hat{\boldsymbol{\theta}}$.

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

Historically, linear mixed effects models have been used to estimate fixed effects. Using them to predict random effects, as in the prediction of μ, raises conceptual problems in the interpretation of μ and in how one should construct prediction intervals. The appropriate method for assessing variability of $\hat{\mu}(t)$ is not obvious when μ is modelled as random.

To model the variability of $\hat{\mu}$, we could take a strictly Bayesian approach, since our estimate of $\mu(t)$ is based on the BLUP $\tilde{\mu}(t)$, which is equal to $\mathrm{E}(\mu(t)|$ the data). The Bayesian viewpoint would have us construct a credible interval for $\mu(t)$ using the posterior variance of $\mu(t)$ given the data. However, one must be confident in the choice of prior. In our context, the prior is based on a smoothing trick and not reflective of prior information about μ. That is, the prior and the view that μ is random do not arise from Bayesian principles. Therefore, we prefer a frequentist approach, assessing variability based on mean squared error. 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 $\tilde{\mu}(t)$ $\pm[\mathrm{var}\{\tilde{\mu}(t)\}]^{1/2}$ since $\mathrm{var}\{\tilde{\mu}(t)\} = \mathrm{var}\{\tilde{\mu}(t) - \sum \beta_j \psi_{Pj}(t)\}$ measures variability of $\tilde{\mu}(t)$ about the population fixed effect, not about $\mu(t)$.

We study two measures of the magnitude of $\tilde{\mu}(t) - \mu(t)$: $e^2_\delta(t) = \mathrm{E}[\ \{\tilde{\mu}(t) - \mu(t)\}^2 \mid \boldsymbol{\delta}]$ 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 $e^2_\delta(t)$ 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. Although our arguments here are in a Bayesian framework with μ random, the conditional approach would appeal to the frequentist, who views μ as fixed and considers 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.

We calculate $e^2_\delta(t)$ and $e^2(t)$, using (4.1), (4.3) and some algebra:

$$\begin{aligned}
\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} &= \mathbf{A}^{-1}\boldsymbol{\mathcal{C}}'\boldsymbol{\mathcal{S}}^{-1}\mathbf{Y} - \boldsymbol{\theta} \\
&= -\frac{1}{\sigma_P^2}\ \mathbf{A}^{-1}\mathbf{I}_\delta\boldsymbol{\theta} + \mathbf{A}^{-1}\boldsymbol{\mathcal{C}}'\boldsymbol{\mathcal{S}}^{-1}\boldsymbol{\epsilon}^* \qquad (4.4)\\
&= -\frac{1}{\sigma_P^2}\ \mathbf{A}^{-1}\begin{pmatrix} 0 \\ \boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\delta}\end{pmatrix} + \mathbf{A}^{-1}\boldsymbol{\mathcal{C}}'\boldsymbol{\mathcal{S}}^{-1}\boldsymbol{\epsilon}^*.
\end{aligned}$$

Consider the first measure:

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

where, by (4.4),

$$\mathbf{B}_{\theta|\delta}\ =\ \mathrm{E}(\ \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\ |\boldsymbol{\delta})\ =\ -\frac{1}{\sigma_P^2}\ \mathbf{A}^{-1}\begin{pmatrix} \mathbf{0} \\ \boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\delta}\end{pmatrix}$$

and

$$\begin{aligned} \mathbf{V}_{\theta|\delta} &= \operatorname{var}(\tilde{\boldsymbol\theta}|\boldsymbol\delta) = \operatorname{var}(\tilde{\boldsymbol\theta}-\boldsymbol\theta|\boldsymbol\delta)\\ &= \mathbf{A}^{-1}\ \boldsymbol{\mathcal{C}}'\boldsymbol{\mathcal{S}}^{-1}\ \operatorname{var}(\boldsymbol\epsilon^*)\ \boldsymbol{\mathcal{S}}^{-1}\boldsymbol{\mathcal{C}}\ \ \mathbf{A}^{-1}\\ &= \mathbf{A}^{-1}\ \sum \mathbf{C}_i'(\boldsymbol\Sigma_i^{*R})^{-1}\ \operatorname{var}(\boldsymbol\epsilon_i^*)\ (\boldsymbol\Sigma_i^{*R})^{-1}\mathbf{C}_i\ \ \mathbf{A}^{-1}\\ &\equiv \sum \boldsymbol{\mathcal{H}}_i\ \operatorname{var}(\boldsymbol\epsilon_i^*)\ \boldsymbol{\mathcal{H}}_i' \equiv \mathbf{V}_\theta, \end{aligned} \tag{4.5}$$

since $\mathbf{V}_{\theta|\delta}$ doesn't depend on $\boldsymbol\delta$ and therefore is not random. So

$$e_\delta^2(t) = \boldsymbol{f}(t)'\left(\frac{1}{\sigma_P^4}\ \mathbf{A}^{-1}\begin{bmatrix} 0 & 0\\ 0 & \boldsymbol\Sigma_\delta^{-1}\boldsymbol\delta\boldsymbol\delta'\boldsymbol\Sigma_\delta^{-1}\end{bmatrix}\mathbf{A}^{-1} + \mathbf{V}_\theta\right)\boldsymbol{f}(t). \tag{4.6}$$

In the Appendix, we show that $\operatorname{pr}(|\tilde\mu(t)-\mu(t)| \le z_{\alpha/2}\ e_\delta(t) \mid \delta) \ge 1-\alpha$ 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 $\boldsymbol\delta'\boldsymbol\delta$ is large.

Now consider the second measure, $e^2(t)$. To calculate $e^2(t)$, we take the expectation of (4.6) under model C:

$$e^2(t) = \boldsymbol{f}(t)'\left(\ \frac{\sigma_{P,C}^2}{\sigma_P^4}\ \mathbf{A}^{-1}\mathbf{I}_\delta\mathbf{A}^{-1} + \ \ \mathbf{V}_{\boldsymbol\theta}\ \right)\boldsymbol{f}(t). \tag{4.7}$$

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 (4.4), $\operatorname{E}(\tilde{\boldsymbol\theta}-\boldsymbol\theta) = 0$ and so $\operatorname{E}\{\tilde\mu(t)-\mu(t)\} = 0$. Thus $e^2(t) = \operatorname{var}\{\tilde\mu(t)-\mu(t)\}$ and so $\operatorname{pr}(\mu(t) \in \tilde\mu(t) \pm z_{\alpha/2}\ e(t)) = 1-\alpha$.

If the variance model is correctly specified, that is, if model D holds, then $e^2(t)$ is equal to the posterior variance of μ given the data. To see this, write $\tilde\mu(t) = E\{\mu(t)|\mathbf{Y}_1,\ldots,\mathbf{Y}_N\}$ and $\operatorname{var}\{\mu(t)|\mathbf{Y}_1,\ldots,\mathbf{Y}_N\} = \operatorname{E}[\{\mu(t)-\tilde\mu(t)\}^2|\mathbf{Y}_1,\ldots,\mathbf{Y}_N]$, which, in the normal model, does not depend on $\mathbf{Y}_1,\ldots,\mathbf{Y}_N$. So $\operatorname{var}\{\mu(t)|\mathbf{Y}_1,\ldots,\mathbf{Y}_N\} = \operatorname{E}[\operatorname{var}\{\mu(t)|\mathbf{Y}_1,\ldots,\mathbf{Y}_N\}] = e^2(t)$. As noted, this posterior variance is commonly used in a Bayesian approach to assess variability of the posterior mean.

The pointwise standard errors proposed by Djeundje and Currie [4] are equivalent to the square roots of estimates of pointwise posterior variances, under a prior chosen to reduce frequentist bias. However, in our simulation study in Section 6, when we applied their method to data generated under a different, but reasonable, prior, the standard errors proved to be far too narrow.

Standard errors based on the assumed prior are discussed by Ruppert *et al* [28], but only in the case that $N = 1$, that is, in the case of P-spline smoothing regression. In Section 6.4 of their book, they discuss the analogues of $\mathbf{V}_\theta$, $e_\delta^2(t)$ and $e^2(t)$ for this single-curve case. To calculate confidence intervals for $\mu(t)$, they compare $e^2(t)$ and $\boldsymbol{f}(t)'\mathbf{V}_\theta\boldsymbol{f}(t)$, and state they prefer the former. They don't consider confidence intervals based on $e^2_\delta(t)$. Their calculations assume that the smoothing model D holds, that is, that $\boldsymbol{\Sigma}_I$ is as in (2.5), while our calculations hold under the more general model C. The simulation results in Section 6 show that this method, with its reliance on the assumed covariance structure, doesn't extend well to more than one curve.

Other authors have proposed alternatives to model-based standard errors. Sun, Zhang and Tong [31] study a linear mixed effects model with time-varying coefficients at the population level, proposing a two step estimation procedure to reduce computational cost. The first step is local least squares regression ignoring the covariance structure, to estimate the population mean parameters. This step produces individual level residuals which are then used in a method of moments procedure to estimate variance parameters. For the case that $N = 1$, Crainiceanu *et al* [2] take a Bayesian approach to modelling $\boldsymbol{\Sigma}_\delta$.

4.3. Estimating $\boldsymbol{e^2_\delta(t)}$ *and* $\boldsymbol{e^2(t)}$

We have defined $\tilde\mu(t)$ and $\hat\mu(t)$, predictors of $\mu(t)$, in the restricted model D and we have defined two measures of the variability of $\tilde\mu(t)$, $e^2_\delta(t)$ in (4.6) and $e^2(t)$ in (4.7). Our calculations for $e^2_\delta(t)$ and $e^2(t)$ are valid under the unrestricted model C. In this section, we define estimators of $e^2_\delta(t)$ 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 $\hat\mu(t)$.

Our estimates of e^2_δ and e^2 rely on estimators of $\mathbf{A}$ in equation (4.2) and the $\boldsymbol{\mathcal{H}}_i$'s in equation (4.3). We estimate $\mathbf{A}$ and the $\boldsymbol{\mathcal{H}}_i$'s by "plugging in" the parameter estimates from our fit of the restricted model D. Our fit of this restricted model also yields $\hat\sigma^2_P$, the estimate of σ^2_P, and $\hat{\boldsymbol{\delta}}$, the BLUP of $\boldsymbol{\delta}$. Using all of these in (4.5) and (4.6), we define the full sandwich estimate of e^2_δ as

$$\hat e^2_{\delta,\text{full}}(t) = \boldsymbol{f}(t)'\Bigg(\frac{1}{\hat\sigma^4_P}\,\hat{\mathbf{A}}^{-1}\begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{\Sigma}_\delta^{-1}\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_\delta^{-1}\end{bmatrix}\hat{\mathbf{A}}^{-1} + \sum \hat{\boldsymbol{\mathcal{H}}}_i(\mathbf{Y}_i - \boldsymbol{\mathcal{C}}_i\hat{\boldsymbol{\theta}})(\mathbf{Y}_i - \boldsymbol{\mathcal{C}}_i\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\mathcal{H}}}_i'\Bigg)\boldsymbol{f}(t.)$$

To define the full sandwich estimate of e^2 in (4.7), we let $\hat\sigma^2_{P,C}$ be the sample variance of the components of $\hat{\boldsymbol{\delta}}$:

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

and let

$$\hat e^2_{\text{full}}(t) = \boldsymbol{f}(t)'\left(\frac{\hat\sigma^2_{P,C}}{\hat\sigma^4_P}\,\hat{\mathbf{A}}^{-1}\mathbf{I}_\delta\hat{\mathbf{A}}^{-1} + \sum \hat{\boldsymbol{\mathcal{H}}}_i(\mathbf{Y}_i - \boldsymbol{\mathcal{C}}_i\hat{\boldsymbol{\theta}})(\mathbf{Y}_i - \boldsymbol{\mathcal{C}}_i\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\mathcal{H}}}_i'\right)\boldsymbol{f}(t).$$

We also define half sandwich estimates of e^2_δ and e^2, estimates that assume the covariance structure of the $\boldsymbol{\epsilon}^*_i$'s is as specified in model C. For these estimates, we simply replace $(\mathbf{Y}_i - \boldsymbol{\mathcal{C}}_i\hat{\boldsymbol{\theta}})(\mathbf{Y}_i - \boldsymbol{\mathcal{C}}_i\hat{\boldsymbol{\theta}})'$ in the full sandwich estimators with an estimate of $\text{var}(\boldsymbol{\epsilon}_i^*)$ valid under the unrestricted model C:

$$\widehat{\text{var}}(\boldsymbol{\epsilon}_i^*) = C_{Ii}\mathbf{S}_{\hat{\theta}}C_{Ii}' + \hat{\sigma}_\epsilon^2\mathbf{I}.$$

with $\mathbf{S}_{\hat{\theta}}$ and $\hat{\sigma}_\epsilon^2$ as in equations (3.5) and (3.6), but using the $\hat{\theta}_i$'s and $\hat{\theta}$ gotten from fitting model D instead of model B.

Thus, we have four different sandwich prediction errors. Using $e^2(t)$ yields a full sandwich error and a half sandwich error. Using $e_\delta^2(t)$ yields a full sandwich error given $\boldsymbol{\delta}$ and a half sandwich error given $\boldsymbol{\delta}$. Our simulation studies indicate that the performances of the four estimators are comparable. One might prefer the full sandwich estimators, as they require fewer model assumptions. In a frequentist approach in which the randomness of μ is simply a device for smoothing, one would prefer a conditional mean squared error. Combining these reasons yields $\hat{e}_{\delta,\text{full}}^2$ as the preferred estimator.

4.4. Temperature data analysis

We constructed figures (not shown) analogous to Figure 3 and 4, except based on model D with μ random, piecewise linear with "time running forward" covariance structure, that is with the "running forward" power basis and with Σ_δ equal to the identity. We constructed predictors of μ using different configurations of knots, and, for the individual level random effects, different covariance structures, that is, for "time running forward", "time running backward" and "time neutral". We also constructed the model-based standard errors and all types of pointwise sandwich standard errors. The results were, for the most part, qualitatively the same as those assuming μ fixed.

We found that, when knots weren't chosen according to the principles of Theorem 3.1, the "time running forward" estimate of μ did not track the pointwise average well. Surprisingly, the "time running backward" estimate did track fairly well. The "time neutral" estimate of μ tracked well. Choosing the individual knots to be a subset of the population knots resulted in all estimates tracking the mean.

No matter what the choice of knots, the model-based prediction errors for the "time running forward" and "time running backward" individual level covariance structures were unreasonably wide, just as in Figure 4. The "time neutral" covariance structure gave pointwise model-based standard error bands that were similar to those in Figure 4, not reflecting the fact that variability in temperatures between cities is much lower in the summer than in the winter.

Under all three individual level covariance structure assumptions, the sandwich standard errors provided substantial improvement over the model-based standard errors. This improvement led to satisfactory standard errors when the individual knots were a subset of the population knots. However, when the individual knots were not a subset of the population knots, the sandwich standard errors did not track the raw standard errors as closely as one would like. Indeed, in this case, the sandwich standard errors still showed some slight signs of the underlying assumed covariance structure. We see this phenomenon in the analysis of the fruit fly data set, in the next section.

For further details and plots, see the Supplementary Material [13].

5. Fruit fly data analysis

We analyzed the fruit fly data using all of the techniques described in Sections 3 and 4. Here, we only present some of the Section 4 analyses, based on random μ. Complete analyses are contained in the Supplementary Material. As always, for μ random, we assumed the "time running forward" covariance structure for μ. Throughout this section, analysis is based on piecewise linear functions with 29 equispaced interior knots for the population mean and 4 equispaced interior knots for the individual effects. Thus the individual knots are a subset of the population knots, and Theorem 3.1 would hold if the design were balanced and μ were modelled as a fixed function.

As an ad hoc method of estimating expected VO2 response within each group, we linearly interpolated the data from each individual and calculated pointwise means at a grid of 100 time points. We also calculated pointwise standard errors of these means, using the pointwise standard deviation divided by the square root of the number of individuals. Note that these standard errors are probably too small, as we have used extra "data points" in their calculation. But the standard errors do give some indication of the pattern of variability since the number of observations is fairly evenly spread over the scaled time interval. If, on the other hand, there were few observations over a portion of the interval, our standard errors might be misleading due to the varying level of certainty in those standard errors.

Figure 6 contains results from an analysis using the individual level "time neutral" structure. The plots show the estimates of the expected VO2 response of fruit flies in the selection and control groups, along with pointwise full-sandwich standard errors based on $e^2(t)$. The Figure also contains the difference of these estimates, along with pointwise standard errors.

Figure 7 shows all types of "time neutral" standard errors for the selection and control groups. Note that the model-based standard errors indicate an approximate homoscedasticity that is not supported by the data. Note also how the sandwich standard errors have corrected this.

Figure 8 shows all types of "time running forward" standard errors. The model-based standard errors display extreme fanning for t near 1. The sandwich standard errors have corrected much but not all of the fanning problem. However, in plots (not shown), we found that estimates of μ under the individual level "time running forward" model seemed to be biased, drifting from the point-wise mean for t near 1. Perhaps the widening of the standard error bars compensates for this drifting of the estimate of μ.

Figures in the Supplementary Material indicate that, when the knots do not satisfy the conditions of Theorem 3.1, the assumed covariance structure has a stronger effect on the sandwich standard errors.

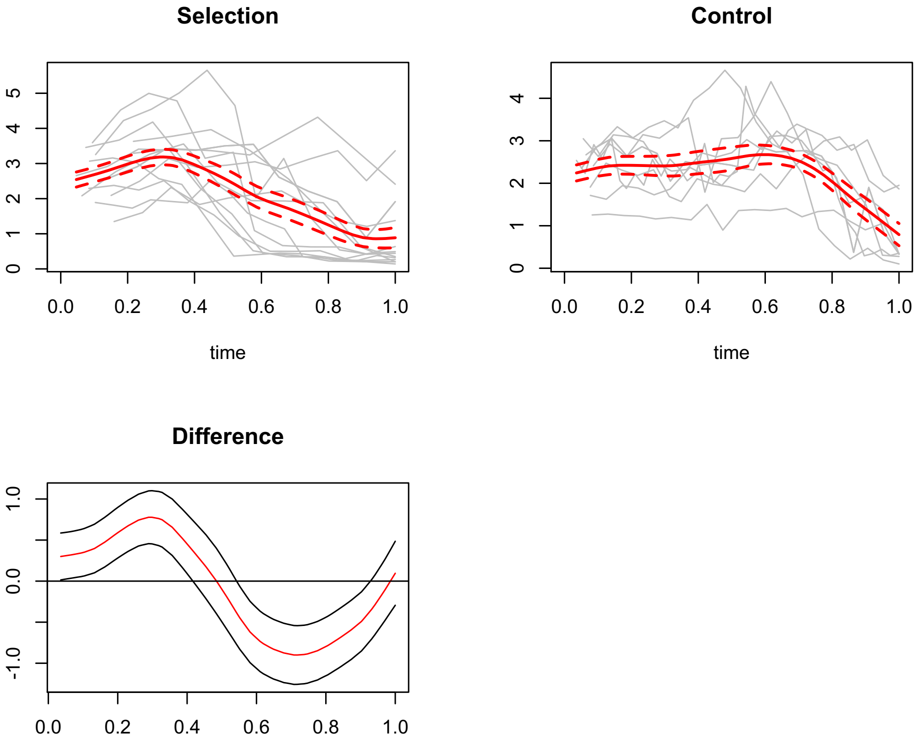

Fig 6. *Estimates and pointwise full-sandwich standard errors of the population mean V02 in the selection and control groups of flies, by techniques of Section 4, using 29 equispaced interior population knots and 4 equispaced interior individual knots. Analysis used the "time neutral" covariance structure for the individual level random effects and full sandwich standard errors are based on $\hat{e}^2_{\mathrm{full}}(t)$. Also shown is the difference of the estimates, with pointwise standard errors of the difference.*

6. Simulation study

We simulated and analyzed data sets in order to study the validity of the different pointwise standard errors. The model for the simulation was chosen to produce data that mimicked the Canadian weather data; station i's temperature on day j was $Y_{ij} = \mu(j) + a_i[\cos(2\pi j/365) + 2] + \epsilon_{ij}$. Here μ is of the form $\mu(t) = \gamma_1 + \gamma_2 t + \gamma_3 \cos(2\pi t/365) + \gamma_4 \sin(2\pi t/365) + \gamma_5 \cos(4\pi t/365) + \gamma_6 \sin(4\pi t/365)$ where the γ_k's minimize $\sum_1^{365}(\mu(j) - \bar{Y}_j)^2$, where $\bar{Y}_j = \sum_i Y_{ij}/35$. The random effect a_i is normal with mean 0 and standard deviation 3.5 and the error ϵ_{ij} is normal with mean 0 and standard deviation 0.842721, which is the estimated value of σ_ϵ from the data analysis. To compare the simulated data to the real data, Figure 9 shows one simulated data set, the Canadian weather data, the pointwise means for these two data sets, and the standard errors of these pointwise means.

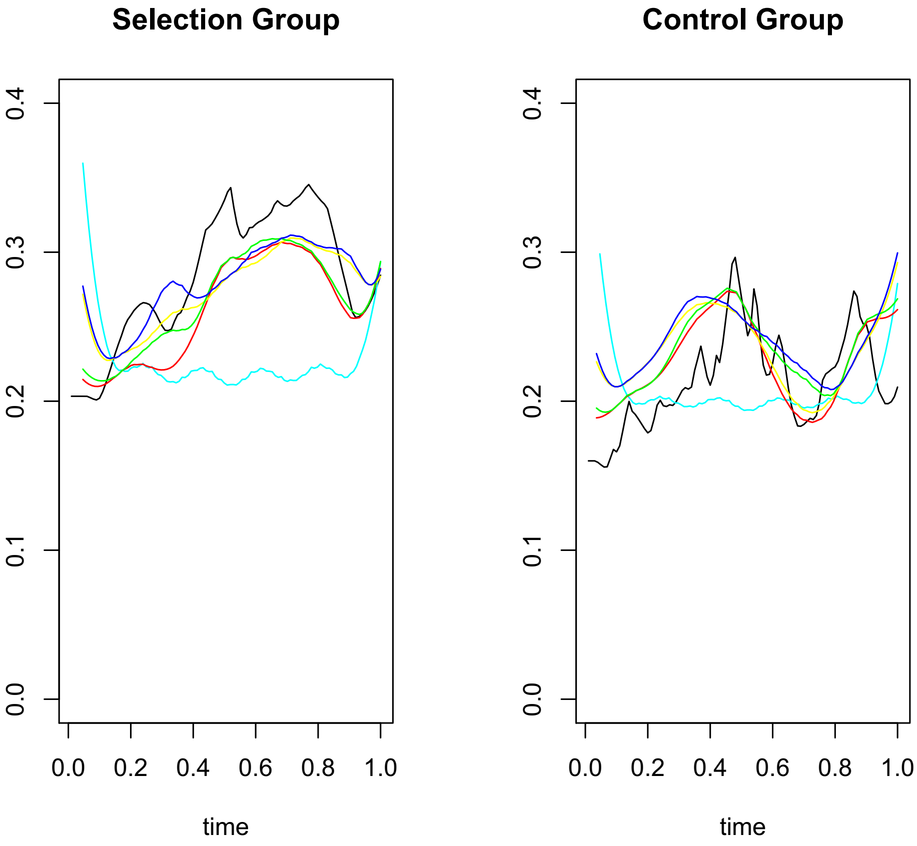

FIG 7. *Standard errors of estimates of the population mean V02 in the selection and control groups of flies by techniques of Section 4, using 29 equispaced interior population knots and 4 equispaced interior individual knots and the "time neutral" covariance structure for the individual level random effects. The standard errors portrayed are: the ad hoc using linear interpolation (black line), the full sandwich (red) using* $\hat{e}_{\text{full}}$*, the half sandwich (yellow), the full sandwich given* δ *(green) using* $\hat{e}_{\delta,\text{full}}$*, the half sandwich given* δ *(blue) and the model-based (cyan).*

We estimated μ in each simulated data set four ways: assuming μ is fixed and the individual random effects have a "time running forward" structure, assuming μ is fixed and the individual random effects have a "time neutral" covariance structure, assuming μ is random and the individual random effects have a "time running forward" structure, and assuming μ is random and the individual random effects have a "time neutral" covariance structure. For random μ, we assumed the "time running forward" covariance structure with $\mathbf{\Sigma}_\delta$ equal to the identity. All functions were modelled as piecewise linear. We analyzed the data using 39 interior population knots and 7 interior individual knots, so that the individual knots were a subset of the population knots.

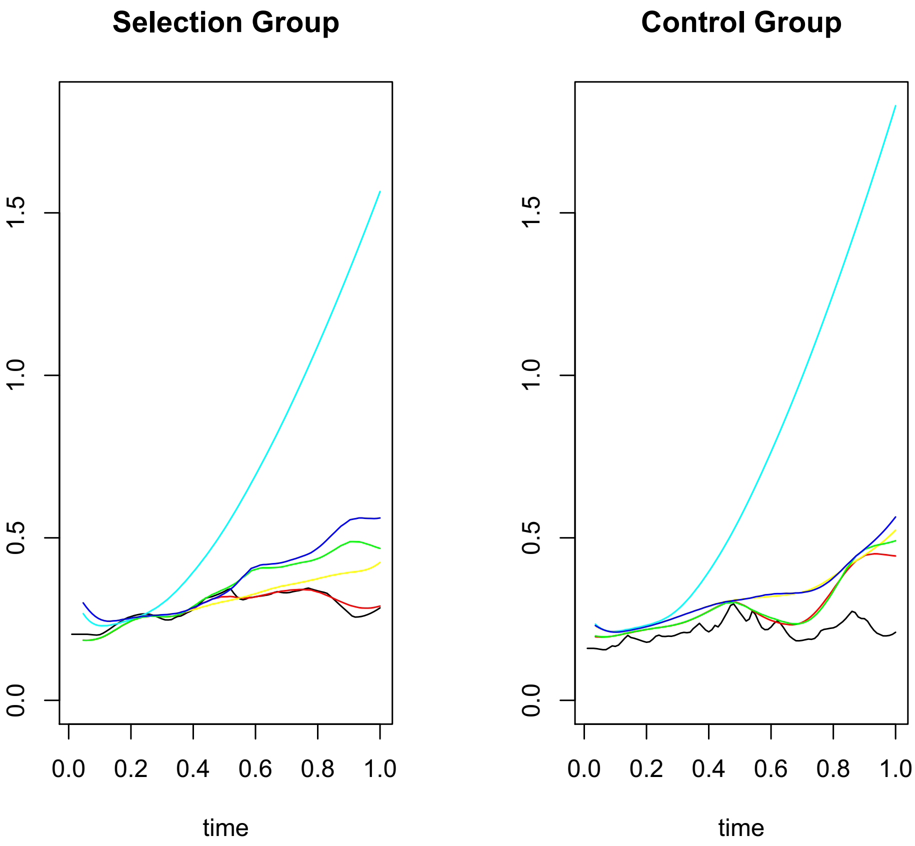

Fig 8. *Standard errors of estimates of the population mean V02 in the selection and control groups of flies by techniques of Section 4, using 29 equispaced interior population knots and 4 equispaced interior individual knots and the "time running forward" covariance structure for the individual level random effects. The standard errors portrayed are: the ad hoc using linear interpolation (black line), the full sandwich (red) using $\hat{e}_{\mathrm{full}}$, the half sandwich (yellow), the full sandwich given δ (green) using $\hat{e}_{\delta,\mathrm{full}}$, the half sandwich given δ (blue) and the model-based (cyan).*

We found that all methods of estimating μ worked well and that model-based standard errors always performed poorly. All sandwich standard errors performed well.

We present some plots here from the random μ "time neutral" analysis, with more plots available in the Supplementary Material. Figure 10 shows pointwise quantiles of the model-based standard errors (in black) and the full-sandwich standard errors based on $\hat{e}_{\mathrm{full}}(t)$ (in red). The dashed blue line is the pointwise standard deviation of the 200 estimates of μ - this is a simulated estimate of the target of our standard errors. We see that the sandwich estimator performs well and the model-based errors do not. Figure 11 shows the proportion of times that a supposed 95% or 98% confidence interval for μ actually contains μ. We

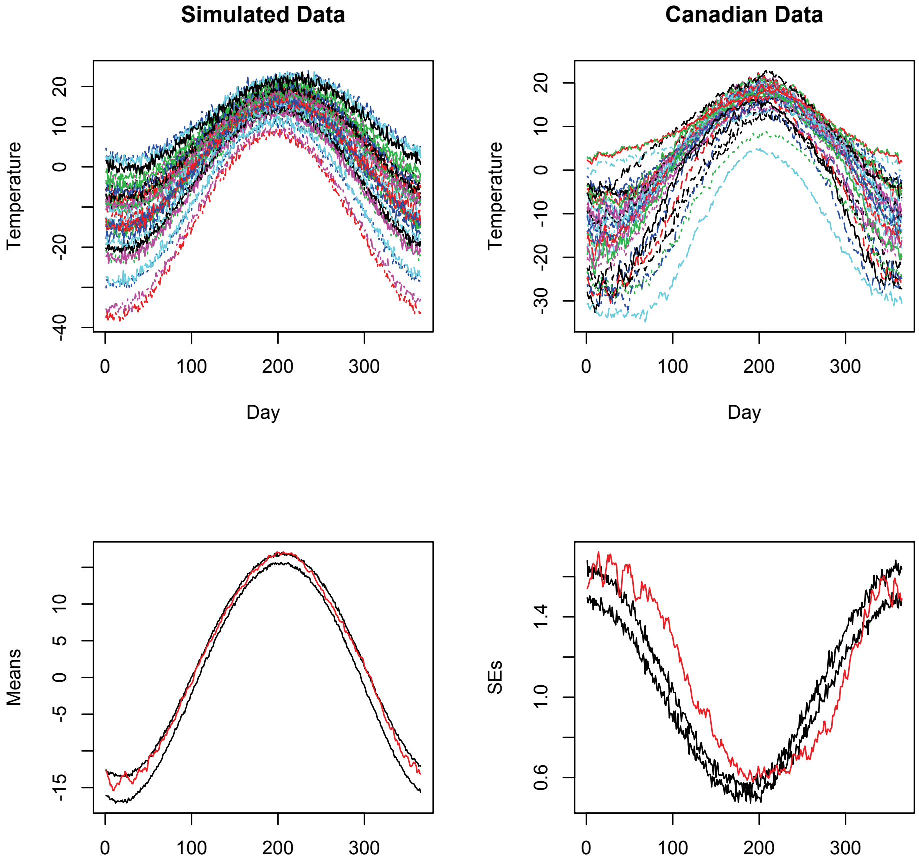

FIG 9. *Comparing the Canadian weather data to data simulated as described in Section 6. The upper left plot shows one of the simulated data sets and the upper right plot contains the Canadian data. The lower left plot shows pointwise means of two simulated data sets (black lines) and of the Canadian data (red line). The lower right plot shows the pointwise standard errors of the means depicted in the lower left.*

calculated intervals as $\hat{\mu}(t) \pm z_{\alpha/2}$ SE(t), with $z_{\alpha/2}$ from the normal distribution. The model-based standard error under-covers during much of the mid-year, as expected from Figure 10. The sandwich standard error has moderate undercoverage, but it may be possible to correct this using a t-distribution instead of the normal when constructing confidence limits.

We also compared our method to that of Djeundje and Currie [4], using their posted R-code (doi: 10.1214/10-EJS583SUPP) in their files Bases.r and Masters_3.r, with basis set to c("B", "B"), corresponding to fitting their model M1, as described in and near their equation (3.3). In their paper, they reported unusually small standard errors. Standard errors for our simulated data were also small, much smaller than the pointwise standard deviations of the estimates of μ.

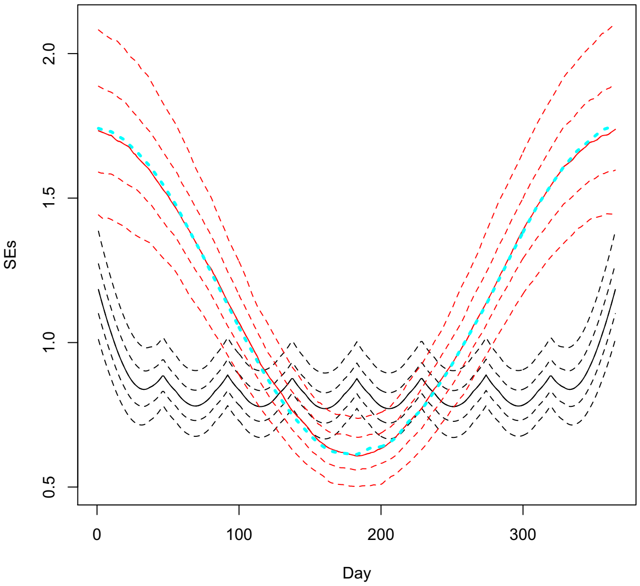

Fig 10. *The pointwise 5th, 25th, 75th 95th quantiles (dashed lines) and 50th quantiles (solid line) of the 200 standard errors from the simulation study of Section 6, with 39 interior population knots and 7 interior individual knots. The analysis assumes that μ is random and that the individual random effects have a "time neutral" covariance structure. The black lines are the quantiles of the model-based standard errors and the red lines are the quantiles of the full sandwich standard errors, using $\hat{e}_{\text{full}}(t)$. The dotted cyan line is the empirical standard error, that is, the pointwise standard deviation of the estimates of μ.*

The resulting pointwise 95% confidence intervals for μ had abut 50% covererage. Discussion of these results is in the Supplementary Material.

7. Discussion

The use of linear mixed effects modelling as a smoothing tool in the analysis of longitudinal data has increased, with many researchers taking advantage of readily available mixed effects model software. Incorporating spline functions into the analysis at both the population level and the individual levels allows more flexible estimators than those from traditional parametric methods. Additionally, using smoothing-based models B or D for the individual random effects results in fast computation. However, use of these spline models can lead to an incorrect population mean estimate and incorrect pointwise standard errors. Indeed, over-reliance on any particular covariance model for the individual random effects can cause problems.

The impact of the assumed covariance structure of the individual effects on the population mean estimate is sometimes serious but can often be remedied by ensuring that the function space for the individual effects is a subspace of the function space for the population mean curve. Therefore, we recommend that the

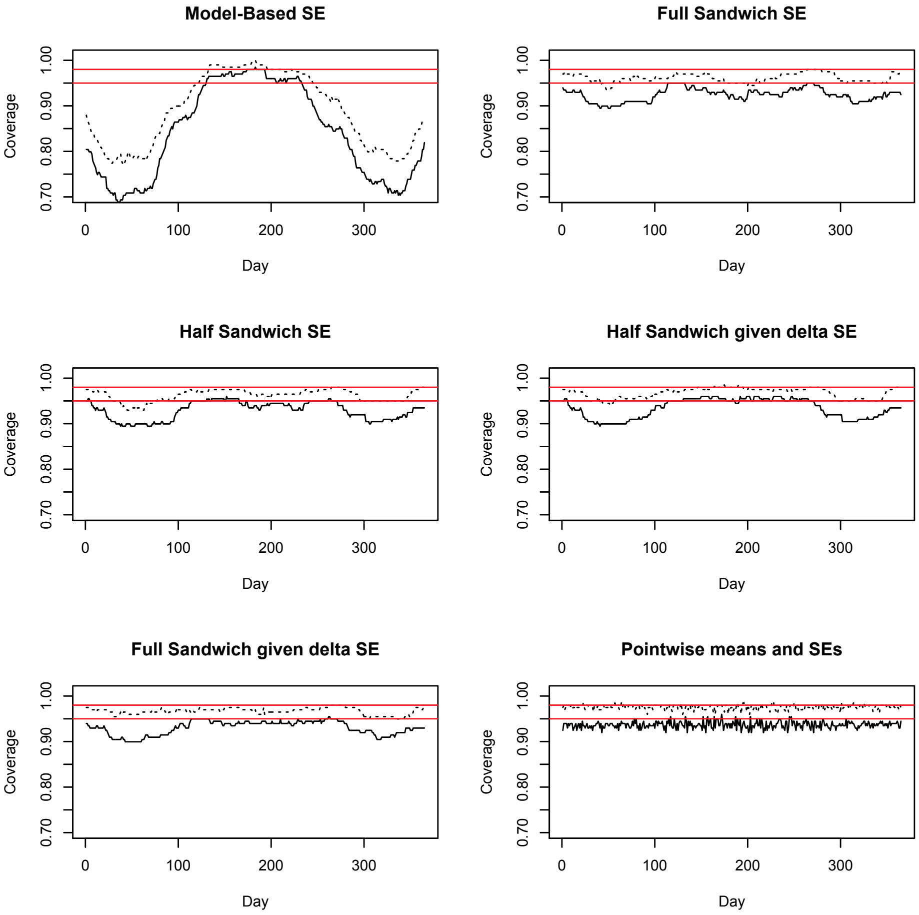

Fig 11. *Coverage of confidence intervals in the simulation study of Section 6, assuming that μ is random and using the time neutral covariance structure. The plots show the proportion of times that μ lies within the pointwise 95% confidence interval (solid line) and the 98th percent confidence interval (dashed line). Horizontal red lines are drawn at 0.95 and 0.98.*

individual level knots be a subset of the population level knots. This structure has the conceptual advantage of allowing us to readily view the random effects as deviations from the mean. It also has the appealing property of resulting in the ordinary least squares estimate of μ when the design is balanced for the case that μ is modelled as non-random (see Theorem 3.1). As we have seen, this choice of knots also improves performance of our proposed sandwich estimators of standard errors. We have seen from simulation studies that confidence interval coverage is good, without any further correction of bias in the estimate of μ.

For the choice of the number of population level knots, we agree with Ruppert *et al* (Section 5.5.3 [27]) who state that “the idea is to choose enough knots to resolve the essential structure of the underlying regression function". When we consider μ random, based on our experience, we also agree with the recommendation of Ruppert *et al* to choose 20 to 40 population knots or one-fourth the number of distinct t values, whichever is smaller. Note, however, that their recommendation was made in a different context, with only one regression data set. Like Ruppert *et al* in their context, we found that the exact number of population knots had little effect, provided the number was sufficiently large. This is in contrast to μ non-random, where one would typically use a much smaller number of population knots, as too many knots will result in a rough estimate of μ.

At the individual level, we do not recommend using only a random intercept or a random slope and intercept, as this typically is not rich enough to account for individual to individual variability. We have found that, to capture this variability, it suffices to use just a few individual knots.

As we have seen, 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, we recommend using the "time neutral" linear Bspline basis with independent and identically distributed coefficients to estimate μ. As even this covariance structure may be incorrect, we still recommend basing standard errors on an unrestricted covariance structure, using sandwich standard errors based on (3.7), (3.8), (4.6) or (4.7). We have found that these sandwich standard errors perform best when we have assumed a "time neutral" covariance structure.

The main contribution of this paper is in drawing attention to the problem of model-based standard errors in the smoothing formulation of random regression, and in presenting a better approach. 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.

Supplementary Material

Supplement to "Penalized regression, mixed effects models and appropriate modelling"

(doi: 10.1214/00-EJS809SUPP; .zip).

The Supplementary Material includes code to obtain estimates of μ and standard errors, as described in Sections 3 and 4. Also included are

- code to produce plots from the paper;
- code to generate simulated data and to run the simulation for our methods;
- code to simulate according to the methods of Djeundje and Currie [4];
- details of the results of the simulation study;

- a description of the analysis of the Canadian weather data, assuming that μ is random, and accompanying code;
- results of the analysis of the fruit fly data, for both μ non-random and μ random.

Appendix A: Appendix

A.1. Proof of Theorem 3.1

The proof of Theorem 3.1 uses the Lemma below, which is a modified, more general version of Theorem 2 of Section 2.3 of [3]. The proof of the Theorem is given after that of the Lemma.

Lemma A.1. *Suppose that* $\mathbf{G}$ *is a matrix of full column rank, that* $\mathbf{M}$ *is a symmetric matrix with* $\mathbf{M}+\mathbf{I}$ *invertible and that the column space of* $\mathbf{M}$ *is contained in the column space of* $\mathbf{G}$*. Then*

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

Proof. We take transposes and show that

$$\left(\mathbf{M}+\mathbf{I}\right)^{-1}\mathbf{G}\{\mathbf{G}'\left(\mathbf{M}+\mathbf{I}\right)^{-1}\mathbf{G}\})^{-1} - \mathbf{G}\left(\mathbf{G}'\mathbf{G}\right)^{-1} = \text{ the 0 matrix.}$$

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

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

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

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

□

Proof of Theorem 3.1. Write $\operatorname{var}(\boldsymbol{\epsilon}_i^*) \equiv \sigma_\epsilon^2(\mathbf{M}+\mathbf{I})$ where $\mathbf{M} = \mathbf{C}_I\boldsymbol{\Sigma}_\theta\mathbf{C}_I'/\sigma_\epsilon^2$. Then $\tilde{\boldsymbol{\theta}}_G = \{\mathbf{C}'\left(\mathbf{M}+\mathbf{I}\right)^{-1}\mathbf{C}\}^{-1}\mathbf{C}'\left(\mathbf{M}+\mathbf{I}\right)^{-1}\bar{\mathbf{Y}}$ and the Ordinary Least Squares estimator is $\hat{\boldsymbol{\theta}}_O = (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\bar{\mathbf{Y}}$. Since the column space of $\mathbf{C}_I$ lies in the column space of $\mathbf{C}$, the column space of $\mathbf{C}_I\mathbf{B}$ also lies in the column space of $\mathbf{C}$ for any matrix $\mathbf{B}$. Thus the column space of $\mathbf{M}$ lies in the column space of $\mathbf{C}$. The result follows directly from the Lemma. □

A.2. Confidence intervals

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 $e_\delta^2(t)$, 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

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

where Z follows a standard normal distribution. Similarly,

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

Consider the function

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

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

[1] B. A. Brumback, L. C. Brumback, and M. J. Lindstrom. *Longitudinal Data Analysis*, pages 291–318. Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G., eds. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC Press, Boca Raton, Florida, 2009. MR1500110

[2] Ciprian M. Crainiceanu, David Ruppert, Raymond J. Carroll, Adarsh Joshi, and Billy Goodner. Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 16(2):265–88, 2007. MR2370943

[3] Eugene Demidenko. *Mixed Models: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, 2004. MR2077875

[4] Viani A. B. Djeundje and Iain D. Currie. Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. *Electronic Journal of Statistics*, 4:1202–1224, 2010. MR2735884

[5] M. Durban, J. Harezlak, M. P. Wand, and R. J. Carroll. Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24(8):1153–67, 2005. MR2134571

[6] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, 1996. MR1435485

[7] Paul H. C. Eilers and Brian D. Marx. *Splines, knots and penalties.* Wiley Interdisciplinary Reviews: Computational Statistics. 2010.

[8] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, 2004. MR2063401

[9] D. G. Folk and T. J. Bradley. The evolution of recovery from desiccation stress in laboratory-selected populations of drosophila melanogaster. *The Journal of Experimental Biology*, 207:2671–2678, 2004.

[10] J. H. Friedman. Multivariate adaptive regression splines (with discussion). *Annals of Statististics*, 19:1–141, 1991. MR1091842

[11] A. Gilmour, B. Gogel, B. R. Cullis, and R. Thompson. *ASReml User Guide Release 2.0*. VSN International Ltd., Hemel Hempstead, U.K., 2006.

[12] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1994. MR1270012

[13] N. Heckman, R. Lockhart, and J. D. Nielsen, Supplementary Material to “Regression, Mixed Effects Models and Appropriate Modelling”. DOI: 10.1214/00-EJS809SUPP.

[14] J. S. Hodges and D. J. Sargent. Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88:367–79, 2001. MR1844837

[15] A. E. Huisman, R. F. Veerkamp, and J. A. M. Van Arendonk. Genetic parameters for various random regression models to describe the weight data of pigs. *Journal of Animal Science*, 80:575–82, 2002.

[16] Raghu N. Kackar and David A. Harville. Approximations for standard errors of estimators of fixed and random effect in mixed linear models. *Journal of the American Statistical Association*, 79:853–862, 1984. MR0770278

[17] George S. Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41:495–502, 1970. MR0254999

[18] Kung Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. MR0836430

[19] Karin Meyer. Random regression analyses using B-splines to model growth of Australian Angus cattle. *Genetics Selection Evolution*, 37(5):473–500, 2005.

[20] Karin Meyer. WOMBAT - a tool for mixed model analyses in quantitative genetics by REML. *Journal of Zheijang University Science B*, 8:815–21, 2007.

[21] L. Ngo and M. P. Wand. Smoothing with mixed model software. *Journal of Statistical Software*, 9:1–54, 2004.

[22] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005. MR2168993

[23] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. MR2514435

[24] John A. Rice and Colin O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57(1):253–9, 2001. MR1833314

[25] Christèle Robert-Granié, Barbara Heude, and Jean-Louis Foulley. Modelling the growth curve of Maine-Anjou beef cattle using heteroskedastic random coefficients models. *Genetics Selection Evolution*, 34(4):423–45, 2002.

[26] G. K. Robinson. That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6(1):15–51, 1991. MR1108815

[27] David Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2003. MR1998720

[28] David Ruppert, M. P. Wand, and R. J. Carroll. Semiparametric regression during 2003-2007. *Electronic Journal of Statistics*, 3:1192–1256, 2010. MR2566186

[29] Andrew D. A. C. Smith and M. P. Wand. Streamlined variance calculations for semiparametric mixed models. *Statistics in Medicine*, 27(3):435–48, 2008. MR2418454

[30] C. J. Stone, M. Hansen, C. Kooperberg, and Y. K. Truong. Polynomial splines and their tensor products in extended linear modeling. *Annals of Statististics*, 25:1371–1425, 1997. MR1463561

[31] Yan Sun, Wenyang Zhang, and Howell Tong. Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics*, 35(6):2795–2814, 2007. MR2382666

[32] A. A. Szpiro, K. M. Rice, and T. Lumley. Model-robust regression and Bayesian 'sandwich' estimator. *Annals of Applied Statistics*, to appear.

[33] A. P. Verbyla, B. R. Cullis, M. G. Kenward, and S. J. Welham. The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of The Royal Statistical Society Series C*, 48(3):269–311, 1999.

[34] Sue J. Welham, Brian R. Cullis, Michael G. Kenward, and Robin Thompson. A comparison of mixed model splines for curve fitting. *Australian & New Zealand Journal of Statistics*, 49(1):1–23, 2007. MR2345406

[35] I. M. S. White, R. Thompson, and S. Brotherstone. Genetic and environmental smoothing of lactation curves with cubic splines. *Journal of Dairy Science*, 82:632–8, 1999.