Electronic Journal of Statistics
Vol. 4 (2010) 1202–1224
ISSN: 1935-7524
DOI: 10.1214/10-EJS583

Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data

Viani A.B. Djeundje and Iain D. Currie

Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK
e-mail: vad5@hw.ac.uk; I.D.Currie@hw.ac.uk

Abstract: A popular approach to smooth models for longitudinal data is to express the model as a mixed model, since this often leads to immediate model fitting with standard procedures. This approach is particularly appealing when truncated polynomials are used as a basis for the smoothing, as the mixed model representation is almost immediate. We show that this approach can lead to a severely biased estimate of the overall population effect and to confidence intervals with undesirable properties. We use penalization to investigate an alternative approach with either B-spline or truncated polynomial bases and show that this new approach does not suffer from the same defects. Our models are defined in terms of B-splines or truncated polynomials with appropriate penalties, but can be expressed as mixed models; this also gives access to fitting with standard procedures. We illustrate our methods with an analysis of two data sets: (a) a balanced data set on Canadian weather and (b) an unbalanced data set on the growth of children.

AMS 2000 subject classifications: Primary 62G08; secondary 62J07.
Keywords and phrases: B-splines, longitudinal data, mixed models, penalties, smoothing, truncated lines.

Received August 2010.

1. Introduction

Mixed effects models are a powerful tool for analyzing longitudinal or more generally grouped data. In its general form, a mixed model consists of expressing some linear predictor as a sum of two components: (a) the fixed effect, originally interpreted as the population/overall effect; and (b) the random effects, which result from the units drawn at random from the population. Searle *et al.* [20] investigate the basic concepts and theoretical aspects of mixed models, while Pinheiro & Bates [13] mainly look at computational issues. A key assumption for a mixed model is the structure of the covariance matrix of the random effects since its specification has important fitting and inferential consequences.

Smoothing methods are known for their flexibility in describing complex patterns, and their connection with mixed models has been of interest to many authors. As a result, modelling longitudinal data with smooth curves has gained much attention and become an area of intensive research. Early work in this area

is based on the mixed model representation of smoothing splines (see Brumback & Rice [1], Verbyla *et al.* [22] among others); however, smoothing splines can be computationally intensive. Alternatively, low rank smoothing methods have been expressed as mixed models. With truncated polynomial bases, a mixed model representation is almost immediate from the form of the basis and the penalty function; Ruppert *et al.* [18] is a comprehensive reference for this approach. Eilers [7], in the discussion to the Verbyla *et al.* [22] paper, pointed out that the P-splines system (Eilers & Marx [8]) with B-spline bases could also be expressed in the mixed model framework. Wood (p191) [24] used the singular value decomposition of the penalty matrix to express a smoothing model in a Bayesian way; Currie *et al.* [3] also used the singular value decomposition to give a mixed model representation of P-splines.

Nonetheless, the representation of smoothing models as mixed models is not without controversy. Green [9] commented on the Verbyla *et al.* [22] paper as follows: "Formulating spline smoothing as a mixed model is simply a mathematical device; the suggested logical distinction between the *fixed* linear trend and the *random* smooth variation is artificial". Green's point is that the randomness in the mixed model representation of smoothers is not assigned to units in a clear way as in the mixed models described in Searle *et al.* (chap 1) [20]. Thus, smoothers usually have a "mixed model representation" but not a "mixed model interpretation" in the original sense. Nowadays, one motivation behind the insertion of smoothing into the mixed model framework is the availability of computer packages for mixed models; two typical examples are the libraries `lme` and `nlme` in the software package R (R Development Core Team [15]) which are designed to fit and compare Gaussian linear and nonlinear mixed models; Pinheiro *et al.* [14].

In practice, however, the real effect of the bases and the covariance structure on the estimated effects in low rank smoothing methods for longitudinal data has received very little attention. In this paper we will consider two commonly used bases: the truncated polynomial bases (Ruppert *et al.* [18]) and the B-spline bases (Eilers & Marx [8]). We discuss first the unfortunate consequences that can occur when we use the standard mixed model with truncated lines for longitudinal data; and second, we discuss the resolution of these problems with appropriate penalties, whether truncated polynomial or B-spline bases are used. Our aim is to present a smooth mixed model for longitudinal data with a natural, ie, non arbitrary, covariance structure and an immediately interpretable fixed effect; this covariance structure is derived from the penalties used to design the model.

The plan of the paper is as follows. Section 2 presents the standard mixed model approach to longitudinal data using truncated lines for balanced data, by which we mean that the same number of observations are made on each unit at the same time points. We encounter some difficulties with this approach and use this to motivate a penalty approach which we examine in section 3. Section 4 presents two ways of fitting the model: (a) with the penalized residual sum of squares, and (b) with the mixed model representation of the model. Confidence band calculation is described in section 5 and an extension to un-

balanced data is presented in section 6. We close with some concluding remarks in section 7.

2. Standard penalized splines mixed model for longitudinal data

In this section, we describe some aspects of a standard penalized spline mixed model for longitudinal data, and motivate the need for an alternative approach. For simplicity, we start with balanced data and so assume that we have longitudinal data $\boldsymbol{Y} = (Y_{i,j})$, $1 \leq i \leq n_1$, $1 \leq j \leq n_2$, stored in the form of a matrix in such a way that columns are classified by subjects (j) and rows by time (t_i); that is, the column data $\boldsymbol{Y}_{\bullet,j}$ are repeated measurements of the response variable Y on the j th unit during time periods, $\boldsymbol{t} = (t_1, \ldots, t_{n_1})'$. A typical example is the well-known `pig.weights` data set available in the library `SemiPar` in the R package. This data set presents the weight measurements on $n_2 = 48$ pigs (subjects) over a period of $n_1 = 9$ weeks (time); an overview of these data is shown in the left panel in Figure 1. The global effect looks linear even though the individual subject lines are quite variable, and so it makes sense to consider models of the form

$$Y_{i,j} = \{\delta_0 + \delta_1 t_i\} + \{\breve{\delta}_{j,0} + \breve{\delta}_{j,1} t_i\} + \varepsilon_{i,j}, \tag{2.1}$$

where $\delta_0 + \delta_1 t$ describes the linear population/overall effect, $\breve{\delta}_{j,0} + \breve{\delta}_{j,1} t$ measures the deviations/departures of the j th subject/pig from the overall effect, and $\varepsilon_{i,j}$ represents the noise.

Clearly, we are not interested only in the specific 48 pigs involved in this study. The main motivation of mixed models is to enable our inference from (2.1) to apply to some population of pigs, and mixed models provide an attractive solution to this problem. We suppose that our sample of pigs is drawn at random from some population of pigs and that the impact of this randomness on model (2.1) is that the subject effects $\boldsymbol{u}_j = (\breve{\delta}_{j,0}, \breve{\delta}_{j,1})'$ are themselves random. A common specification of this randomness is that the $\boldsymbol{u}_j$ are generated from two-dimensional normal distributions with zero means. An important point is that this normal assumption solves the problem of non-identifiability of model (2.1); this same point will arise in section 3, when we will see that penalties provide an alternative solution to the identifiability problem. Under the assumption of normality and homoskedasticity, model (2.1) can be written

$$\boldsymbol{Y}_{\bullet,j} \,|\, \boldsymbol{u}_j \sim \mathcal{N}\left(\boldsymbol{X}_1\boldsymbol{\delta} + \boldsymbol{Z}_1\boldsymbol{u}_j,\ \sigma^2\boldsymbol{I}_{n_1}\right), \quad \boldsymbol{u}_j \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Psi}\right),$$

where $\boldsymbol{X}_1 = \boldsymbol{Z}_1 = [\boldsymbol{1}_{n_1} : \boldsymbol{t}]$, $\boldsymbol{\delta} = (\delta_0, \delta_1)'$, $\boldsymbol{u}_j = (\breve{\delta}_{j,0}, \breve{\delta}_{j,1})'$, $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, $\boldsymbol{1}_n$ is the vector of ones of length n, and $\boldsymbol{\Psi}$ is a 2×2 symmetric, positive definite matrix. This leads to the standard mixed model representation

$$\mathcal{Y} \,|\, \boldsymbol{u} \ \sim\ \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\delta} \ +\ \boldsymbol{Z}\boldsymbol{u},\ \sigma^2\boldsymbol{I}_{n_1 n_2}\right), \quad \boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Phi}\right),$$

with

$$\mathcal{Y} = vec(\boldsymbol{Y}),\ \boldsymbol{X} = \boldsymbol{1}_{n_2} \otimes \boldsymbol{X}_1,\ \boldsymbol{Z} = \boldsymbol{I}_{n_2} \otimes \boldsymbol{Z}_1,\ \boldsymbol{u} = vec(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_2}),\ \boldsymbol{\Phi} = \boldsymbol{I}_{n_2} \otimes \boldsymbol{\Psi}.$$

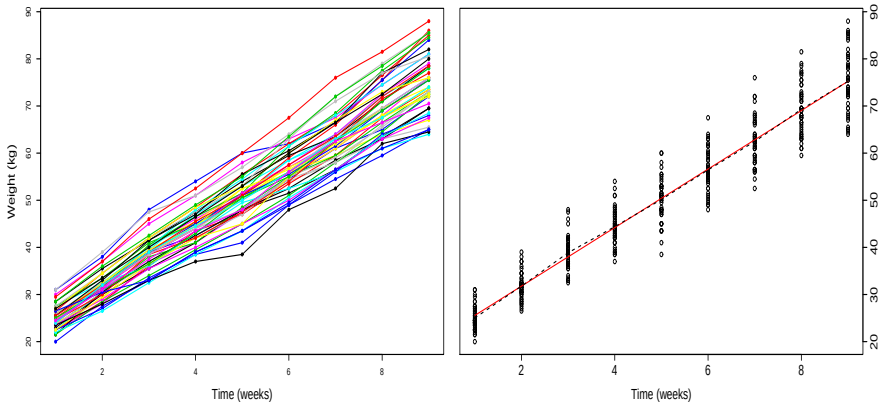

FIG 1. *Left: repeated measurements of the weight of* 48 *pigs over a period of* 9 *successive weeks (each continuous line refers to observations on the same pig). Right: fitted overall/population effect (red line) together with the observed average per week (black dashed line).*

Here, $\otimes$ represents the Kronecker product and $vec(\cdot)$ is the operator which stacks the elements of matrices/vectors into a single vector. In the literature, $\boldsymbol{\delta}$ is known as the fixed effect and $\boldsymbol{u}$ as the random effect. The right panel in Figure 1 illustrates the estimated overall line fitted with the `R` function `lme`. Sub-models of (2.1) for the `pig.weights` data have been investigated by many authors; Ruppert *et al.* [18] among others implemented the case $\breve{\delta}_{j,1} = 0$, meaning that the subject departures from the overall effect are parallel. Such sub-models can be tested against model (2.1) to investigate the significance of this parallelism. However, this sort of test needs to be treated with care since the null hypothesis, $H_0 : \boldsymbol{\Psi}_{1,2} = \boldsymbol{\Psi}_{2,2} = 0$, specifies that the non-negative $\boldsymbol{\Psi}_{2,2}$ is zero, and so sits on the boundary of the parameter space; see Self & Liang [21], Ruppert *et al.* (chap 4) [18].

We have assumed that both the overall and the subject effects can be captured linearly; this assumption suits the pigs data well. However, this assumption is not tenable in general. Consider for example the left panel in Figure 2, which shows the daily average temperature (the averages are taken over the period 1960-1994) in 35 Canadian cities/subjects; these data can be found in the list `CanadianWeather` available in the library `fda` in `R`. Clearly, the linear assumption (at least for the overall effect) fails and more flexibility is required to model the observed effects. In order to account for flexibility in such situations, both linear components in (2.1), ie, the population and the subject effects, are often extended using truncated lines as follows:

$$Y_{i,j} = \left\{ \delta_0 + \delta_1 t_i + \sum_{r=1}^{q} \xi_r (t_i - \tau_r)_+ \right\} + \left\{ \breve{\delta}_{j,0} + \breve{\delta}_{j,1} t_i + \sum_{k=1}^{\breve{q}} \breve{\xi}_{j,k} (t_i - \breve{\tau}_k)_+ \right\} + \varepsilon_{i,j}, \tag{2.2}$$

where $x_+ = max\{x, 0\}$, and $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_q\}$ and $\breve{\boldsymbol{\tau}} = \{\breve{\tau}_1, \ldots, \breve{\tau}_{\breve{q}}\}$ are sets of equally spaced internal knots at the population and subject levels respectively.

To be precise, let $\Delta = (t_{n_1} - t_1)/(q+1)$ and $\breve{\Delta} = (t_{n_1} - t_1)/(\breve{q}+1)$ be the distance between the knots at the population and subject levels, then the τ_r and the $\breve{\tau}_k$ are defined by $\tau_r = t_1 + r\Delta$, $r = 1, \ldots, q$, and $\breve{\tau}_k = t_1 + k\breve{\Delta}$, $k = 1, \ldots, \breve{q}$.

Ruppert *et al.* (sect 9.3) [18] is an early reference to such subject-specific curves, although these authors reported there that "more work needs to be done on implementation". Model (2.2) can be expressed compactly as

$$\boldsymbol{Y}_{\bullet,j} = [\mathbf{1}_{n_1} : \boldsymbol{t}]\,\boldsymbol{\delta} + \boldsymbol{T}_1\boldsymbol{\xi} \;+\; [\mathbf{1}_{n_1} : \boldsymbol{t}]\,\breve{\boldsymbol{\delta}}_j + \breve{\boldsymbol{T}}_1\breve{\boldsymbol{\xi}}_j \;+\; \boldsymbol{\varepsilon}_{\bullet,j} \tag{2.3}$$

where $\boldsymbol{T}_1$ and $\breve{\boldsymbol{T}}_1$ are the matrices of truncated lines at the population and subject levels. Within this setting, the smoothness of the estimates, as well as the identifiability of model (2.2), is frequently achieved by imposing the following normal constraints on the coefficients (Coull *et al.* [2], Ruppert *et al.* (sect 9.3) [18], Durban *et al.* [5], etc):

$$\boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{0}, \sigma_P^2 \boldsymbol{I}_q\right), \quad \breve{\boldsymbol{\delta}}_j \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \quad \breve{\boldsymbol{\xi}}_j \sim \mathcal{N}\left(\mathbf{0}, \sigma_S^2 \boldsymbol{I}_{\breve{q}}\right). \tag{2.4}$$

In (2.4), σ_P^2 is the variance parameter driving the smoothness of the overall effect, $\boldsymbol{\Sigma}$ is a 2×2 symmetric, positive definite matrix, and σ_S^2 is the variance parameter driving the smoothness/randomness at the subject level.

The investigation of the covariance structure (2.4) in terms of modelling effects has not been of great concern. When we are dealing with smoothing at a single level, truncated lines with a ridge penalty produce satisfactory results; with smoothing at two levels, (ie, population and subject levels), the standard covariance specification (2.4) is problematic. Here we illustrate some of its unfortunate consequences.

We use the `lme` function as described in Durban *et al.* [5] to fit this model to `CanadianWeather`. The output of the `lme` function not only gives the estimates for the population and subjects effects, but also provides estimates for the variance parameters in (2.4), which we use to compute the bias corrected confidence intervals (see Ruppert *et al.* (sect 6.4) [18]) for the population and subjects effects. To illustrate our point, we first consider two knot-scenarios at the subject level. Guided by Ruppert [17], we use $q = 39$ equi-spaced knots $\boldsymbol{\tau}$ at the population level in both scenarios.

- Scenario 1: we use $\breve{q} = 19$ equi-spaced knots $\breve{\boldsymbol{\tau}}$ at the subject level; in this case, $\breve{\boldsymbol{\tau}} \subset \boldsymbol{\tau}$.
- Scenario 2: we use $\breve{q} = 21$ equi-spaced knots $\breve{\boldsymbol{\tau}}$ at the subject level.

The right panel in Figure 2 shows the fitted cities (obtained by adding the estimated population effect to the city effects) for both scenarios. As we can see from this graphic, the fits from both scenarios are almost identical and they look very satisfactory with regard to the data. One may be tempted to argue that this goodness of fit at the subject level induces a satisfactory fit at the population level. However, Figure 3 shows the fitted population effect for the two scenarios; we confirm the two observations of Heckman *et al.* [10]:

- the fitted population effect is very sensitive to the knot locations at the subject level, and

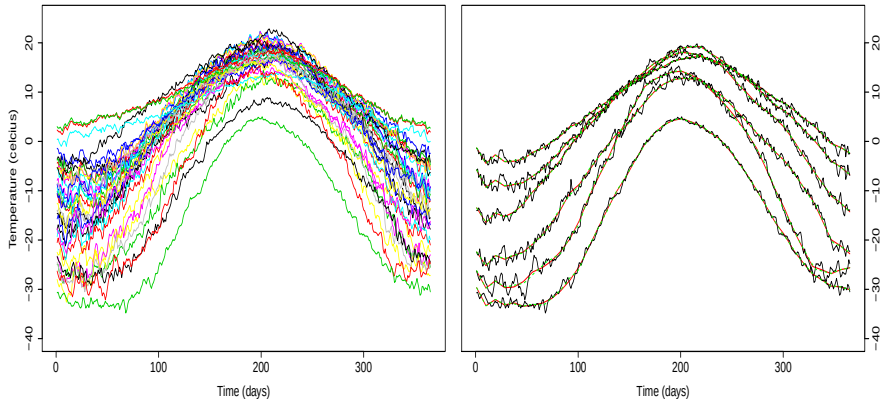

FIG 2. *Left: daily averages of temperature in 35 Canadian cities (each continuous line refers to observations on the same city). Right: the wiggly black lines are the observed values for selected cities; the red (smooth) lines are model (2.4) fitted with* `lme` *under scenario 1; the green (dashed) lines are model (2.4) fitted with* `lme` *under scenario 2 (largely hidden under the red lines).*

- the confidence bands exhibit a widening fan effect as we move from left to right.

Further, for a third scenario (not shown) with $\breve{q} = 20$, we observe upward bias, the opposite of that observed with $\breve{q} = 21$; for a fourth scenario, with $q = \breve{q} = 39$, we observe both severe bias and widening of the confidence interval. In all these scenarios, the behaviour (of the fitted population effect) is balanced by similar behaviour of the subject effects, in such a way that the global effect is recovered appropriately, as illustrated in the right panel of Figure 2. The reason for such behaviour is the mis-specification of the covariance structure. There are two main reasons for the choice in (2.4): first, the ridge penalty on a truncated lines basis works well when smoothing is at a single level and second, the simplicity of (2.4) is attractive; however, it does not appear capable of capturing appropriately the overall effect observed in the left panel in Figure 2. We also remark that if the truncated lines run from right-to-left, ie, with slope -1, as opposed to left-to-right, ie, with slope 1 as in (2.2), then the bias and the fanning effect in Figure 3 are reversed. We refer to these bases as the *forward* (slope 1) and *backward* (slope -1) bases.

One possibility is to use a full covariance matrix in (2.4) in place of $\sigma_S^2 \boldsymbol{I}_{\breve{q}}$. This approach is not attractive since it has no obvious interpretation; it is also computationally very intensive. Thus, we are faced with one of the common challenges in mixed models: the appropriate specification of the covariance structure of the random effect.

In the remainder of this paper, we do not rely on specifying the covariance structure directly; our plan is to work with appropriate penalties. The advantage of this approach is that we can discuss the modelling effects which we wish the

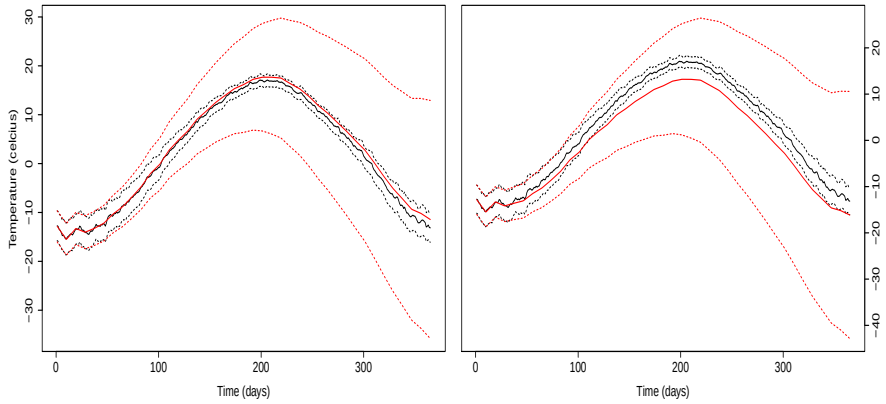

FIG 3. *Illustration of the sensitivity of the estimates of the population effect to the knot locations for the standard model (2.4). Left: scenario* 1, *ie,* 39 *and* 19 *inner knots at the population and subject levels respectively. Right: scenario* 2, *ie,* 39 *and* 21 *inner knots at the population and subject levels respectively. On both graphics, the black line is the observed mean effect with the associated empirical confidence band, while the red line is the fitted population effect.*

penalties to have; furthermore, we show how the penalty framework can be reformulated as a mixed model, and then the covariance structure follows naturally from the penalty structure and the bases. The supplementary materials contain R-code to reproduce Figures 2 and 3.

3. Penalty approach

We consider the general structure

$$Y_{i,j} = S(t_i) + S_j(t_i) + \varepsilon_{i,j}, \quad \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2), \tag{3.1}$$

for some functions $S(\cdot)$ and $S_j(\cdot)$ which quantify the population/overall effect and the deviations/departures of the j th unit from the population effect respectively. We view $S(\cdot)$ as a smooth function (assigned to the population effect) and the $S_j(\cdot)$ as random smooth functions (assigned to the cities). At this stage, we do not make any distributional assumptions as in (2.4); these will come naturally out of our approach. In this section, we present different approaches for modelling $S(\cdot)$ and $S_j(\cdot)$ and propose associated penalties for appropriate identification of these two components.

3.1. Penalties on B-spline bases

Here, we use B-spline bases to construct $S(\cdot)$ and $S_j(\cdot)$; we start with B-spline bases because our approach with B-splines will motivate our solution with truncated polynomials. In brief, with B-splines, we will place separate

penalties directly on the B-spline coefficients at the subject level, one to bring about smoothness (a difference penalty) and another to achieve identifiability by shrinkage (a ridge penalty). With truncated polynomials, it seems more difficult to achieve identifiability by direct shrinkage of the coefficients, as we have seen with the results of (2.2) in Figure 3. Indeed, with truncated polynomial bases we shall see in the next subsection that one way to achieve both smoothness and identifiability is to introduce a second penalty, as we do with B-spline bases. Two additional reasons for starting with B-splines are: first, the degree of the B-splines and the order of the penalty can be chosen independently; this gives additional flexibility to the modeller. Second, B-splines have good numerical properties. Hence, if we denote by $S(\boldsymbol{t})$ the element-wise action of $S(\cdot)$ on the time vector $\boldsymbol{t} = (t_1, \ldots, t_{n_1})'$, with a similar meaning for $S_j(\boldsymbol{t})$, then we write

$$S(\boldsymbol{t}) = \boldsymbol{B}\boldsymbol{a} \quad \text{and} \quad S_j(\boldsymbol{t}) = \breve{\boldsymbol{B}}\breve{\boldsymbol{a}}_j, \tag{3.2}$$

where $\boldsymbol{B}$, $n_1 \times c$, and $\breve{\boldsymbol{B}}$, $n_1 \times \breve{c}$, are regression matrices of B-splines evaluated along $\boldsymbol{t}$, $\boldsymbol{a}$ is a vector of coefficients specifying the population effect, and the $\breve{\boldsymbol{a}}_j$ are random vectors of coefficients related to the subjects. We will refer to (3.1) and (3.2) as model $\texttt{M1} = \texttt{M1}(\boldsymbol{B}, \breve{\boldsymbol{B}})$.

Note that `M1` is not identifiable; indeed, if we add (for example) a constant to $S(\cdot)$ and subtract the same constant from the $S_j(\cdot)$, then the predictor $S(\cdot) + S_j(\cdot)$ remains unchanged. Thus, two issues need to be clarified in `M1`: smoothness and identifiability. In the context of nested curves, Brumback & Rice [1] achieved the smoothness via the smoothing spline approach (which can be very time consuming, specifically in the presence of a large data set); from the mixed model representation, they suggested using ANOVA-like identifiability constraints by requiring that the fixed effects sum to zero at each level except the topmost level. Here we address smoothness and identifiability simultaneously via penalties as follows.

Let us first consider the overall effect $S(\boldsymbol{t})$. For this component we take a sufficiently "rich" set of B-splines as a basis and we apply a roughness penalty (Eilers & Marx, [8]) to the wiggliness of the components of $\boldsymbol{a}$ to achieve the smoothness. Thus the estimation of $\boldsymbol{a}$ will be subject to the constraint

$$\|\boldsymbol{\Delta}_d \boldsymbol{a}\|^2 < z,$$

where $\boldsymbol{\Delta}_d$ represents the difference matrix operator of order d, and z quantifies the amount of smoothness applied to $\boldsymbol{a}$. If $d = 3$ for example, $\boldsymbol{\Delta}_d$ has the four-diagonal structure

$$\boldsymbol{\Delta}_3 = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 3 & -3 & 1 & 0 \\ 0 & \cdots & 0 & 0 & -1 & 3 & -3 & 1 \end{bmatrix}.$$

Given the above specification on the overall effect, we solve the identifiability problem by shrinking the coefficients $\breve{\boldsymbol{a}}_j$ towards $\boldsymbol{0}$. It seems reasonable to apply

the same amount of shrinkage, $\breve{z}_2$, to each of the city effects. Thus, we submit the $\breve{\boldsymbol{a}}_j$ to the constraint

$$\|\breve{\boldsymbol{a}}_j\|^2 < \breve{z}_2,\ j = 1, \ldots, n_2.$$

The problem of smoothness of the city effects remains. Two possibilities are available:

(a) work with fewer B-splines (for the city effects) and only the ridge penalty, or
(b) take a rich set of B-splines as a basis (as for the overall effect) and apply a roughness penalty (together with the ridge penalty) on the $\breve{\boldsymbol{a}}_j$; hence we further penalize the roughness of the $\breve{\boldsymbol{a}}_j$, ie,

$$\|\boldsymbol{\Delta}_2\breve{\boldsymbol{a}}_j\|^2 < \breve{z}_1,\ j = 1, \ldots, n_2.$$

Clearly, (b) is computationally more intensive than (a) since each city has its own (large) set of coefficients, while in comparison with (b), (a) is economical; however, (a) is open to the criticism that the selection of the number of B-splines is manual and artificial. Nonetheless, both approaches produce similar results (at least for `CanadianWeather`), provided that a judicious choice of the number of B-splines is made at the subject level under (a). From now on, we will consider approach (b) only.

We remark that we have used a d-order penalty for smoothing at the population level since we may wish to have a specific fixed effect at this level; for instance, the `CanadianWeather` data in Figure 2 suggest a quadratic fixed effect at the population level, in which case we take $d = 3$. We have no particular form in mind for the city effects, and so we simply use a second order ($d = 2$) penalty to smooth these effects.

In summary for `M1`, (i) smoothing of the population effect is achieved by d-order penalization of the population coefficients, (ii) smoothing of the city effects is achieved by second order penalization of the city coefficients and (iii) identifiability is achieved by a ridge penalty on the city coefficients. These three points are summarized as follows:

$$\texttt{C1} : \|\boldsymbol{\Delta}_d\boldsymbol{a}\|^2 < z, \quad \|\boldsymbol{\Delta}_2\breve{\boldsymbol{a}}_j\|^2 < \breve{z}_1, \quad \|\breve{\boldsymbol{a}}_j\|^2 < \breve{z}_2; \tag{3.3}$$

these constraints apply to the model `M1` and we refer to (3.3) as the constraints `C1`.

3.2. Penalties on truncated polynomial bases

Here we express $S(\cdot)$ and $S_j(\cdot)$ in terms of a truncated polynomial and a truncated line basis respectively; ie, we set

$$S(\boldsymbol{t}) = [\boldsymbol{1}_{n_1} : \boldsymbol{t} : \cdots : \boldsymbol{t}^p]\boldsymbol{\delta} + \boldsymbol{T}_p\boldsymbol{\xi} \ = \ \boldsymbol{X}_p\boldsymbol{\delta} + \boldsymbol{T}_p\boldsymbol{\xi} \ = \ \boldsymbol{L}_p\boldsymbol{b}, \tag{3.4}$$

$$S_j(\boldsymbol{t}) = [\boldsymbol{1}_{n_1} : \boldsymbol{t}]\breve{\boldsymbol{\delta}}_j + \breve{\boldsymbol{T}}_1\breve{\boldsymbol{\xi}}_j \ = \ \boldsymbol{X}_1\breve{\boldsymbol{\delta}}_j + \breve{\boldsymbol{T}}_1\breve{\boldsymbol{\xi}}_j \ = \ \breve{\boldsymbol{L}}_1\breve{\boldsymbol{b}}_j, \tag{3.5}$$

say, where $\boldsymbol{T}_r$ and $\breve{\boldsymbol{T}}_r$ are matrices of truncated polynomials of degree r at the population and subject levels respectively. We will refer to (3.1), (3.4) and (3.5) as $\texttt{M2} = \texttt{M2}(\boldsymbol{T}, \breve{\boldsymbol{T}})$.

With a roughness penalty on B-splines bases, a polynomial fixed effect of degree $(d-1)$ at the population level was captured by choosing a difference penalty of order d. Here, we achieve the same thing in (3.4) by choosing the corresponding degree $p = d-1$ of the polynomial basis. Since the subject effects are likely (at least for `CanadianWeather`) to be quite different from one another, we simply capture them with truncated lines.

With a B-spline basis as in the previous section, the behaviour of $\breve{\boldsymbol{B}}\breve{\boldsymbol{a}}_j$ is very similar to that of $\breve{\boldsymbol{a}}_j$ in the sense that smoothness on $\breve{\boldsymbol{a}}_j$ implies the smoothness of $\breve{\boldsymbol{B}}\breve{\boldsymbol{a}}_j$, and shrinkage on $\breve{\boldsymbol{a}}_j$ implies shrinkage of $\breve{\boldsymbol{B}}\breve{\boldsymbol{a}}_j$. This is not entirely clear with truncated polynomial bases of degree p. For the latter, the coefficient vector $\breve{\boldsymbol{\xi}}_j$ reflects the jumps in the derivatives of order p at the corresponding knots and so the smoothness of the population and subject effects is usually obtained by applying a ridge penalty on $\boldsymbol{\xi}$ and $\breve{\boldsymbol{\xi}}_j$, ie,

$$\|\boldsymbol{\xi}\|^2 < z \quad \text{and} \quad \|\breve{\boldsymbol{\xi}}_j\| < \breve{z}_1,\ j = 1, \ldots, n_2.$$

With B-splines, we have two penalties at the subject level, one for smoothness and one for identifiability. With this in mind, we achieve identifiability by the introduction of a second penalty in `M2` at the subject level which shrinks each subject effect $S_j(\boldsymbol{t}) = \breve{\boldsymbol{L}}_1\breve{\boldsymbol{b}}_j$ towards $\boldsymbol{0}$:

$$\|\breve{\boldsymbol{L}}_1\breve{\boldsymbol{b}}_j\|^2 < \breve{z}_2,\ j = 1, \ldots, n_2.$$

In summary for `M2`, smoothness at the population and subject level is obtained by applying a ridge penalty on the truncated polynomial coefficients, and identifiability is achieved by shrinking all the subject effects towards 0. We summarize these constraints in

$$\texttt{C2} : \|\boldsymbol{\xi}\|^2 < z, \quad \|\breve{\boldsymbol{\xi}}_j\| < \breve{z}_1, \quad \|\breve{\boldsymbol{L}}_1\breve{\boldsymbol{b}}_j\|^2 < \breve{z}_2. \tag{3.6}$$

3.3. Penalties on a mixture of B-spline and truncated polynomial bases

Here we consider a mixture of B-splines and truncated polynomials. We start with

$$S(\boldsymbol{t}) = \boldsymbol{B}\boldsymbol{a} \quad \text{and} \quad S_j(\boldsymbol{t}) = \breve{\boldsymbol{L}}_1\breve{\boldsymbol{b}}_j \ = \ \boldsymbol{X}_1\breve{\boldsymbol{\delta}}_j + \breve{\boldsymbol{T}}_1\breve{\boldsymbol{\xi}}_j, \tag{3.7}$$

say, where the components are defined in previous subsections. We refer to (3.1) and (3.7) as $\texttt{M3} = \texttt{M3}(\boldsymbol{B}, \breve{\boldsymbol{T}})$; for the same reasons detailed previously, smoothness and identifiability constraints on `M3` are

$$\texttt{C3} : \|\boldsymbol{\Delta}_d\boldsymbol{a}\|^2 < z, \quad \|\breve{\boldsymbol{\xi}}_j\| < \breve{z}_1, \quad \|\breve{\boldsymbol{L}}_1\breve{\boldsymbol{b}}_j\|^2 < \breve{z}_2. \tag{3.8}$$

Similarly, we consider the representation

$$S(\boldsymbol{t}) = \boldsymbol{L}_p\boldsymbol{b} \ = \ \boldsymbol{X}_p\boldsymbol{\delta} + \boldsymbol{T}_p\boldsymbol{\xi} \quad \text{and} \quad S_j(\boldsymbol{t}) = \breve{\boldsymbol{B}}\breve{\boldsymbol{a}}_j; \tag{3.9}$$

we refer to (3.1) and (3.9) as $\texttt{M4} = \texttt{M4}(\boldsymbol{T}, \breve{\boldsymbol{B}})$; smoothness and identifiability constraints on M4 are as follows:

$$\texttt{C4}: \|\boldsymbol{\xi}\|^2 < z, \quad \|\boldsymbol{\Delta}_2 \breve{\boldsymbol{a}}_j\|^2 < \breve{z}_1, \quad \|\breve{\boldsymbol{a}}_j\|^2 < \breve{z}_2. \tag{3.10}$$

3.4. Further possibilities

Models M1, M2, M3 and M4 with the associated constraints C1, C2, C3 and C4 are the main models that we investigate here. In all of these cases, we achieve smoothness either by a roughness (difference) penalty on the B-spline coefficients or a ridge penalty on the truncated polynomial coefficients. An alternative might be to smooth by applying a roughness (difference) penalty directly on the estimates, ie, on $S(\boldsymbol{t})$ and $S_j(\boldsymbol{t})$, (whether a B-spline or a truncated polynomial basis is used). Note also that solving the identifiability problem by shrinking the subject effects $S_j(\boldsymbol{t})$ is also applicable with a B-spline basis at the subject level. Furthermore, instead of applying shrinkage to $S_j(\boldsymbol{t})$, one can solve the identifiability problem by applying the shrinkage at the knots only, ie, a ridge penalty to $S_j(\breve{\boldsymbol{\tau}})$. These are topics for further research.

4. Inference and applications

In the previous section, we presented four formulations of model (3.1) with penalized splines. Each of these formulations has the form

$$\boldsymbol{Y}_{\bullet,j} = \boldsymbol{G}\boldsymbol{\alpha} + \breve{\boldsymbol{G}}\breve{\boldsymbol{\alpha}}_j + \boldsymbol{\varepsilon}_{\bullet,j}, \quad \boldsymbol{\varepsilon}_{\bullet,j} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n_1}) \tag{4.1}$$

which can be expressed compactly as

$$\mathcal{Y} = \boldsymbol{\Omega}\boldsymbol{\theta} + vec(\boldsymbol{\varepsilon}), \quad vec(\boldsymbol{\varepsilon}) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n_1 n_2}) \tag{4.2}$$

where $\boldsymbol{\theta} = vec(\boldsymbol{\alpha}, \breve{\boldsymbol{\alpha}})$, $\breve{\boldsymbol{\alpha}} = vec(\breve{\boldsymbol{\alpha}}_1, \ldots, \breve{\boldsymbol{\alpha}}_{n_2})$, is the joint vector of coefficients, $\boldsymbol{\Omega} = [\boldsymbol{1}_{n_2} \otimes \boldsymbol{G} : \boldsymbol{I}_{n_2} \otimes \breve{\boldsymbol{G}}]$ is the full regression matrix, and $\boldsymbol{G}$, $n_1 \times c$, and $\breve{\boldsymbol{G}}$, $n_1 \times \breve{c}$, are regression matrices at the population and subject levels; specifically, we have:

$$(\boldsymbol{G}, \boldsymbol{\alpha}) = \begin{cases} (\boldsymbol{B}, \boldsymbol{a}) & \text{under M1 or M3} \\ (\boldsymbol{L}_p, \boldsymbol{b}) & \text{under M2 or M4} \end{cases} \tag{4.3}$$

$$(\breve{\boldsymbol{G}}, \breve{\boldsymbol{\alpha}}_j) = \begin{cases} (\breve{\boldsymbol{B}}, \breve{\boldsymbol{a}}_j) & \text{under M1 or M4} \\ (\breve{\boldsymbol{L}}_1, \breve{\boldsymbol{b}}_j) & \text{under M2 or M3.} \end{cases} \tag{4.4}$$

We will present two ways of fitting model (4.2) (with reference to the data CanadianWeather) under the associated constraints C1, C2, C3 or C4. The first approach will be based on the penalized residual sum of squares, while the second approach will use the mixed model representation of the model.

4.1. Inference with penalized residual sum of squares

Using Lagrange arguments, the penalized residual sum of squares (PRSS) of (4.2), ie, the residual sum of squares (RSS) under constraints C1, C2, C3 or C4,

can be expressed as

$$\texttt{PRSS} = \texttt{RSS} + \boldsymbol{\theta}'\boldsymbol{P}\boldsymbol{\theta}, \quad \text{with} \quad \texttt{RSS} = (\mathcal{Y} - \boldsymbol{\Omega}\boldsymbol{\theta})'(\mathcal{Y} - \boldsymbol{\Omega}\boldsymbol{\theta}), \tag{4.5}$$

where

$$\boldsymbol{P} = blockdiag(\boldsymbol{P}_{\boldsymbol{\alpha}}, \boldsymbol{I}_{n_2} \otimes \boldsymbol{P}_{\breve{\boldsymbol{\alpha}}}) \tag{4.6}$$

is the block diagonal penalty matrix, with

$$\boldsymbol{P}_{\boldsymbol{\alpha}} = \begin{cases} \lambda \boldsymbol{\Delta}_d' \boldsymbol{\Delta}_d & \text{for } \texttt{M1}(\boldsymbol{B}, \breve{\boldsymbol{B}}) \text{ under } \texttt{C1} \text{ or } \texttt{M3}(\boldsymbol{B}, \breve{\boldsymbol{T}}) \text{ under } \texttt{C3} \\ \lambda \boldsymbol{J}_d & \text{for } \texttt{M2}(\boldsymbol{T}, \breve{\boldsymbol{T}}) \text{ under } \texttt{C2} \text{ or } \texttt{M4}(\boldsymbol{T}, \breve{\boldsymbol{B}}) \text{ under } \texttt{C4} \end{cases} \tag{4.7}$$

$$\boldsymbol{P}_{\breve{\boldsymbol{\alpha}}} = \begin{cases} \breve{\lambda}_1 \boldsymbol{\Delta}_2' \boldsymbol{\Delta}_2 + \breve{\lambda}_2 \boldsymbol{I}_{\breve{c}} & \text{for } \texttt{M1}(\boldsymbol{B}, \breve{\boldsymbol{B}}) \text{ under } \texttt{C1} \text{ or } \texttt{M4}(\boldsymbol{T}, \breve{\boldsymbol{B}}) \text{ under } \texttt{C4} \\ \breve{\lambda}_1 \boldsymbol{J}_2 + \breve{\lambda}_2 \breve{\boldsymbol{L}}_1' \breve{\boldsymbol{L}}_1 & \text{for } \texttt{M2}(\boldsymbol{T}, \breve{\boldsymbol{T}}) \text{ under } \texttt{C2} \text{ or } \texttt{M3}(\boldsymbol{B}, \breve{\boldsymbol{T}}) \text{ under } \texttt{C3}. \end{cases} \tag{4.8}$$

Here, $\boldsymbol{J}_r$ is the identity matrix (of appropriate size) where the upper r diagonal elements have been replaced by 0's, while λ and $\breve{\lambda}_1$ are the smoothing parameters at the population and subject level respectively, and $\breve{\lambda}_2$ is the shrinkage parameter of the subject effects; $(\lambda, \breve{\lambda}_1, \breve{\lambda}_2)$ plays (inversely) the equivalent role as $(z, \breve{z}_1, \breve{z}_2)$ used throughout section 3. More precisely, increasing values of λ and $\breve{\lambda}_1$, (ie, decreasing the values of z and $\breve{z}_1$) induces more smoothness on the population and subject effects, while increasing the values of $\breve{\lambda}_2$, (ie, decreasing values of $\breve{z}_2$) corresponds to heavier shrinkage on the subject effects. At the limit, ie, $\breve{\lambda}_2 \to \infty$ (or equivalently $\breve{z}_2 \to 0$), we have $S_j(\cdot) \to 0$; this reduces the linear predictor of the model to the population effect. Given values of these parameters, we obtain

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{\Omega}'\boldsymbol{\Omega} + \boldsymbol{P})^{-1} \boldsymbol{\Omega}'\mathcal{Y}$$

on minimizing the PRSS in (4.5). For illustration we choose the smoothing/shrinkage parameters by minimizing the Bayesian Information Criterion (BIC) Schwarz [19]

$$\texttt{BIC}(\lambda, \breve{\lambda}_1, \breve{\lambda}_2) = n_1 n_2 \times \log(\texttt{RSS}) \; + \; tr(\boldsymbol{H}) \times \log(n_1 n_2); \tag{4.9}$$

in (4.9), $tr(\boldsymbol{H})$, the trace of the hat matrix $\boldsymbol{H}$, is the *effective dimension* of the fitted model, where $\boldsymbol{H} = \boldsymbol{\Omega}(\boldsymbol{\Omega}'\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\boldsymbol{\Omega}'$ maps the observations to the fitted values. Nonetheless, since the BIC mostly addresses model choice at the global level, a more formal criterion for the selection of the shrinkage parameter may be derived by some partition of the BIC into a population and a subject component. Alternatively, one may prefer to choose a certain fixed amount of shrinkage, or to study (as a function of the shrinkage parameter) the departure of the fitted population effect from the observed mean, and then choose the amount of shrinkage that minimizes this departure.

Finally, in line with the familiar unbiased estimate of variance in linear regression (Wood (p171) [24]), we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\texttt{RSS}}{n_1 n_2 - tr(\boldsymbol{H})},$$

although with penalization this estimate is only approximately unbiased.

TABLE 1
Summary table for four models applied to Canadian weather data

	M1$(\boldsymbol{B}, \breve{\boldsymbol{B}})$	M2$(\boldsymbol{T}, \breve{\boldsymbol{T}})$	M3$(\boldsymbol{B}, \breve{\boldsymbol{T}})$	M4$(\boldsymbol{T}, \breve{\boldsymbol{B}})$
$(\lambda, \breve{\lambda}_1, \breve{\lambda}_2)$	$(0.035, 20, 0.023)$	$(250, 1097, 7 \times 10^{-4})$	$(0.083, 1097, 5 \times 10^{-4})$	$(250, 20, 0.023)$
RSS	6902	6624	6635	6902
$tr(\boldsymbol{H})$	450	514	513	450
BIC	117179	117261	117267	117182
σ^2	0.56	0.54	0.54	0.56

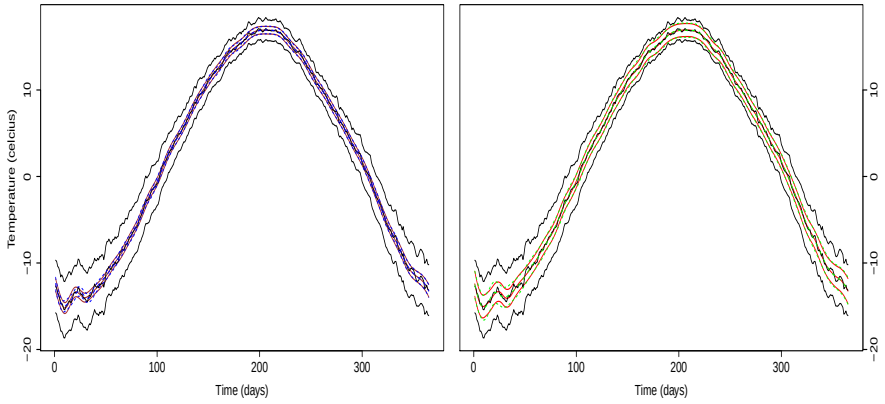

FIG 4. *Data and fitted population effect for our four models. The wiggly (black) line is the data (average) with the associated empirical pointwise confidence interval. Left:* M1 *(brown) and* M4 *(blue). Right:* M2 *(red) and* M3 *(green).*

We now apply this procedure to CanadianWeather. For all our applications, we will use cubic B-splines. For each of our four models, we follow Ruppert *et al.* [18] and so use 39 equi-spaced internal knots at the population and subject levels respectively. The results are summarized in Table 1. Figure 4 illustrates the fitted population effect with the associated confidence intervals for our four models; the confidence intervals have been computed using the Bayesian argument described in section 5. M1 and M4 are the best models (for CanadianWeather) both in terms of BIC and parsimony. Note that the confidence bands for models M1 and M4, the two models with $\breve{\boldsymbol{B}}$ as regression matrix at the subject level, are narrower than those of M2 and M3. The right panel in Figure 5 displays the city effects as estimated under model M1; these city effects are essentially identical for all four models. The supplementary materials contain R-code to reproduce Figure 4 and Table 1.

In summary, our four models all return essentially identical estimates of both the population and city effects; the width of the confidence intervals appears to depend on the basis used at the subject level. We will return to this point in our concluding remarks.

We have also considered using different knots scenarios at the subject level from that at the population level. While this generally produces consistent estimates of the population and subject effects, this may adversely affect the

confidence intervals at both levels for all four models, if the number of knots at the subject level is "too small" relative to that at the population level. Again, we will return to this point in our closing discussion.

Going back to the data in Figure 2, we see that the overall effect looks quadratic with some noise, mostly at the boundaries. Furthermore, our data present a mixed model structure; by this we mean that it is natural to suppose that our cities are a random selection from the population of Canadian cities. We discuss a mixed model approach for (3.1) or (4.1) in the next section.

4.2. Mixed model representation and interpretation

In the representation (4.1), it is natural to think of the coefficients $\breve{\boldsymbol{\alpha}}_j$ as random, since the subjects which they represent are randomly chosen from the population. The question is the following: from the smoothness and identifiability assumptions made so far, can we "naturally" derive the distributions which have generated these coefficients/subjects?

4.2.1. Mixed model representation for `M2` *and* `M4`

Recall that under `M2` or `M4`, we have a truncated polynomial basis at the population level, ie, $\boldsymbol{G} = \boldsymbol{L}_p = [\boldsymbol{X}_p : \boldsymbol{T}_p]$ and $\boldsymbol{\alpha} = \boldsymbol{b} = vec(\boldsymbol{\delta}, \boldsymbol{\xi})$. Here, the mixed model representation is straightforward; this follows from the structure of the truncated polynomial basis at the population level. Indeed, the minimization of `PRSS` under `M2` or `M4` (with the associated constraints) is equivalent to the maximization of the log-likelihood which arises from the triplet $(\mathcal{Y}, \boldsymbol{\xi}, \breve{\boldsymbol{\alpha}})$, where $\boldsymbol{\xi}$ and $\breve{\boldsymbol{\alpha}}$ are treated as a pair of independent random vectors under the distributional assumptions

$$\begin{aligned} \boldsymbol{Y}_{\bullet,j} \,|\, \boldsymbol{\xi}, \breve{\boldsymbol{\alpha}}_j &\sim \mathcal{N}\left(\boldsymbol{X}_p\boldsymbol{\delta} + \boldsymbol{T}_p\boldsymbol{\xi} \;+\; \breve{\boldsymbol{G}}\breve{\boldsymbol{\alpha}}_j,\;\; \sigma^2\,\boldsymbol{I}_{n_1}\right), \\ \boldsymbol{\xi} \sim \mathcal{N}\left(\boldsymbol{0},\; \tfrac{\sigma^2}{\lambda}\boldsymbol{I}_{c-d}\right),\;\; \breve{\boldsymbol{\alpha}}_j &\sim \mathcal{N}\left(\boldsymbol{0},\; \sigma^2\boldsymbol{P}_{\breve{\boldsymbol{\alpha}}}^{-1}\right) \end{aligned} \tag{4.10}$$

where $\boldsymbol{P}_{\breve{\boldsymbol{\alpha}}}$, defined in (4.8), depends on $\breve{\lambda}_1$ and $\breve{\lambda}_2$. We comment on this representation in subsection 4.2.3.

4.2.2. Mixed model representation for `M1` *and* `M3`

Under both `M1` and `M3`, we have a B-spline basis at the population level and so $\boldsymbol{G} = \boldsymbol{B}$ and $\boldsymbol{\alpha} = \boldsymbol{a}$. In this case, the minimization of `PRSS` (with the associated constraints) is equivalent to maximizing the log-likelihood which arises from the triplet $(\mathcal{Y}, \boldsymbol{a}, \breve{\boldsymbol{\alpha}})$, where $\boldsymbol{a}$ and $\breve{\boldsymbol{\alpha}}$ are treated as a pair of independent random vectors under the (improper for $\boldsymbol{a}$) distributional assumptions

$$\begin{aligned} \boldsymbol{Y}_{\bullet,j} \,|\, \boldsymbol{a}, \breve{\boldsymbol{\alpha}}_j &\sim \mathcal{N}\left(\boldsymbol{B}\boldsymbol{a} \;+\; \breve{\boldsymbol{G}}\breve{\boldsymbol{\alpha}}_j,\;\; \sigma^2\,\boldsymbol{I}_{n_1}\right), \\ \boldsymbol{a} \sim \mathcal{N}\left(\boldsymbol{0},\; \sigma^2(\lambda\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d)^{-}\right),\;\; \breve{\boldsymbol{\alpha}}_j &\sim \mathcal{N}\left(\boldsymbol{0},\; \sigma^2\boldsymbol{P}_{\breve{\boldsymbol{\alpha}}}^{-1}\right). \end{aligned} \tag{4.11}$$

The roughness matrix $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$, which gives rise to the improper prior distribution for $\boldsymbol{a}$ in (4.11), is singular, symmetric and has rank $c-d$. Now the singular value decomposition of $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$ is of the form $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ where $\boldsymbol{\Lambda} = diag(\rho_1, \ldots, \rho_{c-d}, 0, \ldots, 0)$ is the $c \times c$ diagonal matrix of eigenvalues arranged in non-increasing order; and $\boldsymbol{U}$ is the matrix with columns given by the eigenvectors of $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$. We will denote $diag(\rho_1, \ldots, \rho_{c-d})$ by $\boldsymbol{\Lambda}_+$. With this notation, the smoother $\boldsymbol{B}\boldsymbol{a}$ can be expressed as

$$\boldsymbol{B}\boldsymbol{a} = \boldsymbol{X}_p\boldsymbol{a}_1 + \boldsymbol{R}\boldsymbol{a}_2, \quad \text{with } \boldsymbol{R} = \boldsymbol{B}\boldsymbol{U}_+\boldsymbol{\Lambda}_+^{-1/2}, \quad \boldsymbol{a}_2 = \boldsymbol{\Lambda}_+^{1/2}\boldsymbol{U}_+'\boldsymbol{a}. \tag{4.12}$$

In (4.12), $\boldsymbol{U}_+$ is the sub-matrix of $\boldsymbol{U}$ which corresponds to the positive eigenvalues of $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$. The (improper) normal assumption about $\boldsymbol{a}$ in (4.11) reduces to $\boldsymbol{a}_2 \sim \mathcal{N}\left(\boldsymbol{0},\ \frac{\sigma^2}{\lambda}\boldsymbol{I}_{c-d}\right)$. Finally, minimizing the `PRSS` of `M1` or `M3` yields the mixed model representation

$$\begin{aligned} \boldsymbol{Y}_{\bullet,j} \,|\, \boldsymbol{a}_2, \breve{\boldsymbol{\alpha}}_j &\sim \mathcal{N}\left(\boldsymbol{X}_p\boldsymbol{a}_1 + \boldsymbol{R}\boldsymbol{a}_2 + \breve{\boldsymbol{G}}\breve{\boldsymbol{\alpha}}_j,\ \sigma^2\,\boldsymbol{I}_{n_1}\right), \\ \boldsymbol{a}_2 \sim \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{\lambda}\boldsymbol{I}_{c-d}\right), &\quad \breve{\boldsymbol{\alpha}}_j \sim \mathcal{N}\left(\boldsymbol{0},\ \sigma^2\boldsymbol{P}_{\breve{\alpha}}^{-1}\right); \end{aligned} \tag{4.13}$$

we comment on this representation in the next subsection.

4.2.3. Interpretation of the components

Clearly from (4.10) and (4.13), the mixed model representation of (4.1) for our four models `M1`, `M2`, `M3` and `M4` with the associated constraints has the form

$$\begin{aligned} \boldsymbol{Y}_{\bullet,j} \,|\, \boldsymbol{\gamma}, \breve{\boldsymbol{\alpha}}_j &\sim \mathcal{N}\left(\boldsymbol{X}_p\boldsymbol{\beta} + \boldsymbol{Z}_p\boldsymbol{\gamma} + \breve{\boldsymbol{G}}\breve{\boldsymbol{\alpha}}_j,\ \sigma^2\,\boldsymbol{I}_{n_1}\right), \\ \boldsymbol{\gamma} \sim \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{\lambda}\boldsymbol{I}_{c-d}\right), &\quad \breve{\boldsymbol{\alpha}}_j \sim \mathcal{N}\left(\boldsymbol{0},\ \sigma^2\boldsymbol{P}_{\breve{\alpha}}^{-1}\right), \end{aligned} \tag{4.14}$$

for appropriate $\boldsymbol{\beta}$, $\boldsymbol{Z}_p$, and $\boldsymbol{\gamma}$. Hence, the model predictor is made up of three components:

- The first component, $\boldsymbol{X}_p\boldsymbol{\beta}$, represents the fixed overall effect. Motivated by the overview of the data in Figure 2, we require this component to be quadratic for `CanadianWeather`; this justifies the use of the third order difference penalty in `M1` and `M3`. An illustration of this first component under `M1` is shown by the continuous line in the left panel of Figure 5.
- The second component, $\boldsymbol{Z}_p\boldsymbol{\gamma}$, which is shrunk towards $\boldsymbol{0}$, accounts for the flexibility of the population effect, and smoothly captures the deviation of the population effect from a simple quadratic curve. We do not view the normal constraint on this component as random behaviour, but just as a smoothing device. This component is illustrated (under `M1`) for `CanadianWeather` by the dashed line in the left panel in Figure 5.
- The third/random component, $\breve{\boldsymbol{G}}\breve{\boldsymbol{\alpha}}_j$, measures the random departure of the subjects from the overall effect. The normal constraint on this component incorporates the random behaviour of the cities (controlled by $\breve{\lambda}_2$) as well as the smoothness of the city effects (as measured by $\breve{\lambda}_1$); these are shown for `CanadianWeather` (under `M1`) on the right panel in Figure 5.

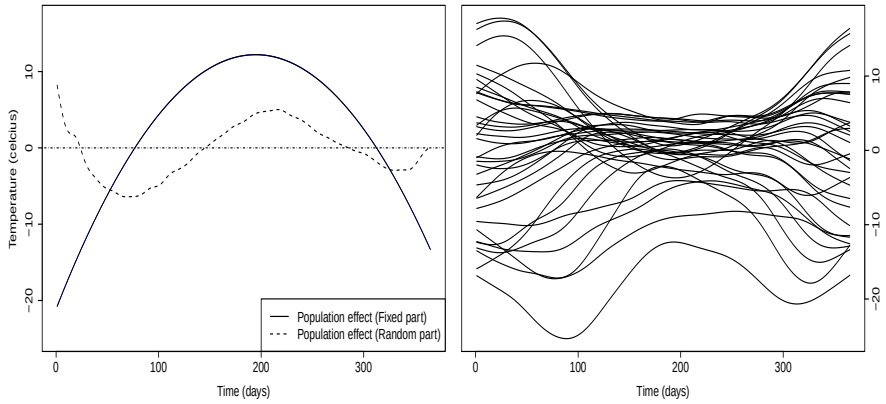

FIG 5. *The three components of model* `M1` *applied to the data* `CanadianWeather`. *Left: decomposition of the fitted population effect into the fixed (quadratic) component (continuous line) and the random component (dashed line). Right: fitted subject effects.*

4.2.4. Mixed model fitting

Since the smoothness/shrinkage constraints on the second component, $\boldsymbol{Z}_p\boldsymbol{\gamma}$ in (4.14), are expressed in terms of a distribution, it may also be treated as some fictive part of the random component; this will then allow the full insertion of the model into the mixed model framework, where we can take advantage of the estimation methodology of restricted maximum likelihood. In line with this motivation, setting

$$\begin{gathered}\boldsymbol{X} = \mathbf{1}_{n_2} \otimes \boldsymbol{X}_p, \quad \boldsymbol{Z} = \left[\mathbf{1}_{n_2} \otimes \boldsymbol{Z}_p \,:\, \boldsymbol{I}_{n_2} \otimes \breve{\boldsymbol{G}}\right], \quad \boldsymbol{u} = vec\left(\boldsymbol{\gamma}, \breve{\boldsymbol{\alpha}}_1, \ldots, \breve{\boldsymbol{\alpha}}_{n_2}\right), \\ \boldsymbol{\Phi} = \boldsymbol{\Phi}_{\sigma^2,\lambda,\breve{\lambda}_1,\breve{\lambda}_2} = \sigma^2 \times blockdiag\left(\lambda^{-1}\boldsymbol{I}_{c-d},\ \boldsymbol{I}_{n_2} \otimes \boldsymbol{P}_{\breve{\boldsymbol{\alpha}}}^{-1}\right)\end{gathered} \tag{4.15}$$

reduces (4.14) to

$$\mathcal{Y} \,|\, \boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\ \ \sigma^2\, \boldsymbol{I}_{n_1 n_2}\right), \quad \boldsymbol{u} \sim \mathcal{N}\left(\mathbf{0},\ \boldsymbol{\Phi}\right). \tag{4.16}$$

Hence, model (3.1), together with the smoothness and identifiability of the model achieved via the penalties in section 3, is now expressed as a mixed model with fixed effects $\boldsymbol{\beta}$, random effects $\boldsymbol{u}$ and covariance matrix $\boldsymbol{\Phi}$. The larger the smoothing/shrinkage parameters are, the flatter the random component $\boldsymbol{Z}\boldsymbol{u}$ is, and the closer the predictor $\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$ approaches the fixed component $\boldsymbol{X}\boldsymbol{\beta}$. With a truncated polynomial basis at the population level, the mixed model representation was straightforward, because of the form of the basis; with a B-spline basis however, additional effort was required; in this case, the structure of the resulting regression matrix $\boldsymbol{R}$ (as defined in (4.12)) is far from obvious, but it does arise naturally as a function of the positive eigenvalues and corresponding eigenvectors of $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$, and the B-spline basis.

A standard approach (Searle *et al.* [20]) for fitting (4.16) consists of maximizing the joint distribution of $(\mathcal{Y}, \boldsymbol{u})$ over $(\boldsymbol{\beta}, \boldsymbol{u})$. This leads to a set of equations

that simultaneously yields an estimate for the fixed effect ($\boldsymbol{\beta}$) and a predictor for the random effect ($\boldsymbol{u}$):

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\mathcal{Y}\\ \tilde{\boldsymbol{u}} &= \boldsymbol{\Phi}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\mathcal{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right).\end{aligned} \tag{4.17}$$

These are the Best Linear Unbiased Estimator/Predictor (BLUE/BLUP) of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ respectively (Robinson, [16]). In (4.17), $\boldsymbol{V} = var(\mathcal{Y}) = \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}_{n_1n_2}$, where $\boldsymbol{\Phi} = \boldsymbol{\Phi}_{\sigma^2,\lambda,\breve{\lambda}_1,\breve{\lambda}_2}$ is specified in (4.15) and so, $(\sigma^2, \lambda, \breve{\lambda}_1, \breve{\lambda}_2)$ are all viewed here as variance parameters.

Clearly, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$ are numerically known (only) upon specification of the variance parameters. There is a variety of literature for estimating these parameters. One possibility is to maximize the unconditional likelihood of $\mathcal{Y}$, $\mathcal{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta},\ \boldsymbol{V})$. However, one unsatisfactory property of maximum likelihood in estimating the variance components is that it discards the information on the degrees of freedom involved in the estimation of the fixed effect. As a result, the maximum likelihood estimator of the variance parameters tends to be biased. This problem with maximum likelihood is corrected by the so-called restricted likelihood (Patterson & Thompson [12]), one justification of which consists in assuming a uniform prior distribution for the fixed effects, then integrating them out of the likelihood (Laird & Ware [11], Pinheiro & Bates [13]). After substituting $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}}$, minus twice the restricted log-likelihood of model (4.16) is given (up to an additive constant) by

$$\begin{aligned}-2\ell\left(\sigma,\lambda,\breve{\lambda}_1,\breve{\lambda}_2\right) &= \log(|\boldsymbol{V}|) + \log\left(|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|\right)\\ &\quad + \mathcal{Y}'\left(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right)\mathcal{Y}.\end{aligned} \tag{4.18}$$

The estimation machine then consists of estimating the variance parameters from maximization of ℓ in (4.18) and substituting them into (4.17) to derive estimates for the BLUE/BLUP. An interesting point is that the standard mixed model approach in (2.4) involves six variance parameters, while our penalty approach (4.16) involves only four such parameters; hence, besides producing satisfactory results, the penalty approach is computationally more economical compared to (2.4), when searching for optimal values of the smoothing/variance parameters.

5. Variability bands

So far, we have presented two approaches for fitting and interpreting model (3.1). Inference in this context is a delicate issue because it depends on whether the mixed model formulation is being used or not. We first hide the mixed model formulation, and we discuss inference from the Bayesian perspective by relying on the posterior distribution of the parameters. The main motivation for working in the Bayesian framework here stems from the fact that the estimator

of the parameter vector $\boldsymbol{\theta}$, defined in (4.2), is usually biased (unless the parameter vector itself vanishes). As a result, the central limit theorem cannot be used directly to compute approximate confidence intervals. In addition to that, there is evidence that the posterior confidence bands derived from the Bayesian perspective have good sampling properties; Wood (chap 4) [24].

We are motivated by the smoothness/identifiability criteria `C1`, `C2`, `C3`, or `C4` on the joint vector $\boldsymbol{\theta}$ and so assume the following (improper) prior distribution about $\boldsymbol{\theta}$: $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{P}^{-})$, where $\boldsymbol{P}$ is defined in (4.6); we then combine this prior information with the conditional distribution $\mathcal{Y}|\boldsymbol{\theta}$ in (4.14) to yield the posterior distribution

$$\boldsymbol{\theta} \,|\, \mathcal{Y} \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}},\ \sigma^2(\boldsymbol{\Omega}'\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\right). \tag{5.1}$$

The posterior distribution of the population coefficient $\boldsymbol{\alpha}$ (defined in (4.1)) follows immediately from the upper left $c \times c$ block of $var(\boldsymbol{\theta} \,|\, \mathcal{Y})$; it is this posterior distribution that we have used to compute the confidence band of the population effect shown in Figure 4. The computation of the confidence bands for the city effects follows similarly. One interesting point here is that the confidence bands obtained from (5.1) are identical to those provided by the bias adjusted confidence bands derived from the mixed model representation (provided the same values of the variance parameters are used in the two perspectives); see Ruppert *et al.* (chap 6) [18].

6. Extensions

For the data `CanadianWeather`, we have assumed that the 35 cities are similar in the sense that model (3.1) is expressed in terms of a common mean (smooth) curve and the subject/city departures. These cities can be classified into four regions: Arctic, Atlantic, Continental and Pacific, and plotting the data by region shows that the observed means (average) are different in shape and level from one region to the other. Model (3.1) can be extended to account for the region effects. Once again, the estimated mean effect per region is very sensitive to the knot locations if fitted with the standard approach similar to (2.2)–(2.4). In contrast, any of our models (similar to) `M1`, `M2`, `M3` or `M4` with the associated smoothness and identifiability constraints (similar to) `C1`, `C2`, `C3` or `C4` does not suffer from this defect; the results are not presented here.

For simplicity, we have presented our work so far in the special case of balanced data where measurements are made on the subjects at the same time. The extension to different timings for subjects as well as to unbalanced data is straightforward. For illustration, we consider the simulated data based on the model in Durban *et al.* [5] related to the heights of 197 children suffering from acute lymphoblastic leukaemia, and receiving three different treatments; these data are displayed in Figure 6. If we denote by $Y_{i,j,k}$ the ith measurement of the height of the jth child receiving treatment k, then its linear predictor can be expressed as

$$E\left[Y_{i,j,k}\right] = S_k(t_{i,j,k}) + S_{j,k}(t_{i,j,k}), \tag{6.1}$$

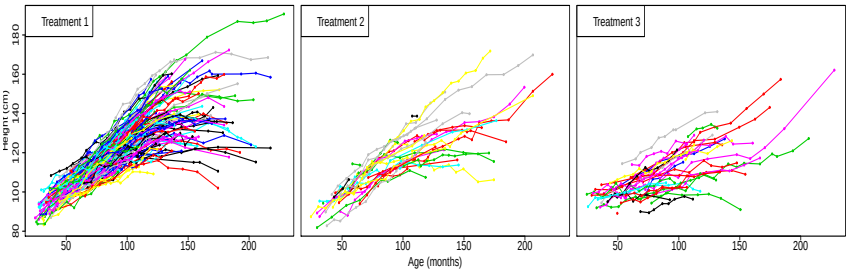

FIG 6. *Heights of* 197 *children suffering from acute lymphoblastic leukaemia, and receiving three different treatments.*

where $S_k(\cdot)$ and $S_{j,k}(\cdot)$ are smooth functions which quantify the treatment effect and the children deviations respectively. We can express (6.1) in matrix form as

$$E\left[\boldsymbol{Y}_{\bullet,j,k}\right] = S_k(\boldsymbol{t}_{\bullet,j,k}) + S_{j,k}(\boldsymbol{t}_{\bullet,j,k}), \tag{6.2}$$

where $\boldsymbol{t}_{\bullet,j,k}$ is the time vector for the j th child receiving treatment k. Details on the components of this model with truncated lines can be found in Coull *et al.* [2] and Durban *et al.* [5], among others; these authors referred to this model as a factor by curve interaction model. Even though the number of observations per child is very small (it varies from 1 to 21), the estimated treatments effects are biased if fitted with the standard approach similar to (2.4); this is shown in the left panel of Figure 7. Indeed, for the same knot locations (40 and 10 inner knots at the population and child levels respectively as used by Durban *et al.* [5]), this graphic illustrates a clear difference between the estimates derived from the two equivalent bases: (a) the forward truncated lines bases and (b) the backward truncated lines bases. The fanning effect seen on the right of the left panel of Figure 7 arises partly from data thinning. In addition, however, we observe the same fanning effects as seen in Figure 3.

Alternatively, each of our four formulations `M1`, `M2`, `M3` or `M4` is extendable to such unbalanced situations. More precisely, we set

$$S_k(\boldsymbol{t}_{\bullet,j,k}) = \boldsymbol{G}_{\boldsymbol{t}_{\bullet,j,k}}\boldsymbol{\alpha}_k \quad \text{and} \quad S_{j,k}(\boldsymbol{t}_{\bullet,j,k}) = \breve{\boldsymbol{G}}_{\boldsymbol{t}_{\bullet,j,k}}\breve{\boldsymbol{\alpha}}_{j,k}, \tag{6.3}$$

where $\boldsymbol{\alpha}_k$ and $\breve{\boldsymbol{\alpha}}_{j,k}$ are vectors of coefficients quantifying the treatment and children effects respectively; $\boldsymbol{G}_{\boldsymbol{t}_{\bullet,j,k}}$ and $\breve{\boldsymbol{G}}_{\boldsymbol{t}_{\bullet,j,k}}$ are matrices (of B-splines or truncated polynomials as the case may be) at the treatment and child level, constructed along the time vector $\boldsymbol{t}_{\bullet,j,k}$. Unlike the data `CanadianWeather`, there is no strong motivation to use the third order penalty or a second order truncated polynomial for these growth data (see Figure 6). Model (6.3) can be expressed compactly as in (4.2) with appropriate components, and then the coefficients can be estimated either by the penalized residual sum of squares (with appropriate penalty matrix) or by re-parameterization as a mixed model, as described in section 4.2. The fitted mean effects for the three treatments are shown on the right panel in Figure 7; this approach does not suffer from the

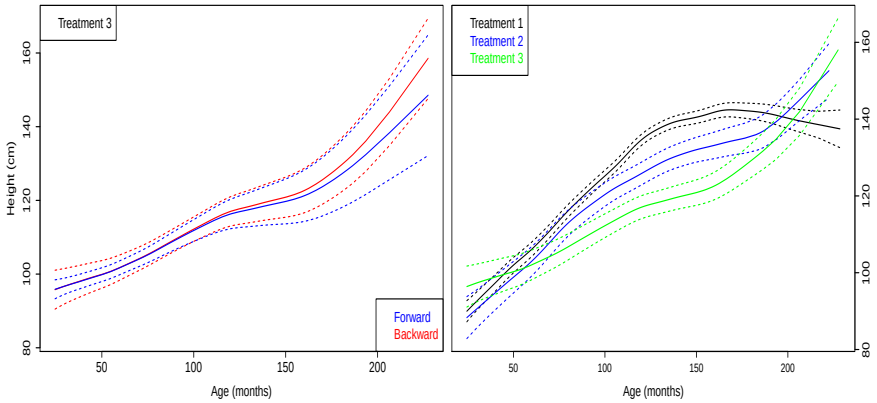

FIG 7. *Left: estimated mean height for children receiving treatment* 3*, using forward (blue) and backward (red) truncated lines bases respectively, with ridge penalties. Right: estimated mean height for the three treatments using B-splines with penalties at the treatment and child levels. The dashed lines represent the corresponding confidence bands.*

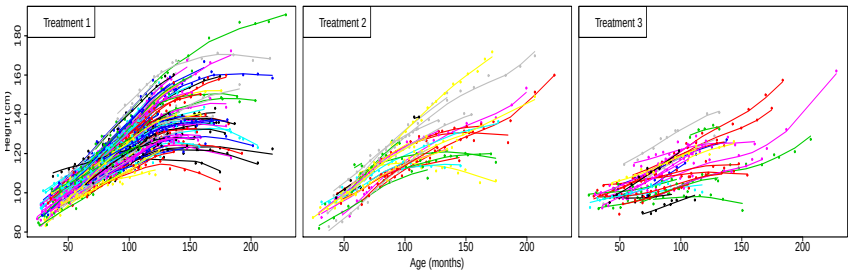

FIG 8. *Fitted model together with the data. Here we have used B-spline bases with associated penalties both at the treatment and children levels.*

instability illustrated in the left panel. Adding the fitted child effects to these treatment effects yields the child curves presented in Figure 8. More generally, the penalty approach is easily adapted to model multilevel nested hierarchical curves. The supplementary materials contain R-code to reproduce Figure 7.

7. Concluding remarks

In this work we first illustrated some consequences of the mis-specification of the standard covariance structure (2.4) in a mixed model (for longitudinal data) defined using truncated line bases. One simple way of demonstrating the problem is to fit only the population effect. Here, truncated lines with a ridge penalty and B-splines with a roughness penalty give almost identical answers, and both capture the population effect correctly with appropriate confidence intervals. However, when we add the city effects, the estimates of the population effect

and its associated confidence intervals are distorted when truncated lines with the covariance structure (2.4) are used, as shown in Figure 3. No such distortion occurs with the penalty approach presented in this paper; the estimates of the population effect are identical whether city effects are included or not.

For the penalty approach, we first specify the bases, and then we design the components of the model (population, subject, etc, effects). With the components in place, we use penalties to bring about the model effects we wish to achieve. One unlooked for bonus with the penalty approach is the reduction in the number of variance parameters to be estimated, from six with (2.4) to four with (4.16). Even though the B-spline and truncated polynomial bases produced satisfactory results in the applications presented in this paper, we have a preference for B-splines bases, because of the direct connection between the regression coefficients and the penalty which is applied to these coefficients; for instance, with the B-spline basis, we can easily adjust the penalty to link the start and end of the year via a circular penalty, or to account for a periodic effect (for example if we are interested in modelling the temperatures collected over many years) by using a harmonic penalty. In the case of the `CanadianWeather` data we have used neither a circular nor a harmonic penalty since we used these data to illustrate some general points of fitting nested smooth curves.

One obvious advantage that truncated polynomial bases appear to have over B-spline bases (at the population level) is that a mixed model representation is immediate; this allows simple fitting with standard methodology (restricted likelihood) upon model specification. With B-splines (at the population level) we must work a little harder to achieve a mixed model representation and so gain access to the standard methodology of restricted likelihood; the transformed bases are not intuitive but can be arrived at via a penalty argument. Of course, the B-spline approach can be expressed in terms of truncated polynomials, but the resulting covariance structure would again be all but impossible to guess without penalties; some details on transformations from B-splines to truncated polynomials can be found in de Boor [6] and Welham *et al.* [23].

We return to two issues which we raised at the end of section 4.1. First, our methods appear to be successful in recovering population and subject effects, and in solving the problem of the widening fan effect found with (2.4). However, the width of the associated confidence intervals arising from the use of `BIC` depends on whether a B-spline or a truncated lines basis is used at the subject level. Nonetheless, it is possible to "play" with the values of the smoothing/shrinkage parameters when truncated lines are used and to produce the confidence intervals obtained with B-splines. A second difficulty arises when selecting the smoothing/shrinkage parameters by optimizing a deviance-type criterion like `BIC`. We found that, for balanced data such as `CanadianWeather`, if the number of knots at the subject level is "too small" relative to the number at the population level, ie, $\breve{q} \ll q$, (for instance, $q = 39$ and $\breve{q} < 10$, for `CanadianWeather`), then the optimal values of the shrinkage/smoothing parameters as selected by `BIC` fall on the boundary of the parameters space; this can lead to unexpectedly wide confidence intervals at the population and subject levels. Hence, in practice, attention must be given to the choice of q and $\breve{q}$ (for

example by following Ruppert [17] both at the population and subject levels) as well as to the optimization criterion.

Acknowledgements

The second author is very grateful to Richard Lockhart of Simon Fraser University for introducing him to this topic and hosting a visit to Simon Fraser. Both authors benefited from discussions with Maria Durban, Paul Eilers and others which were funded by the Spanish Ministry of Science and Innovation (project MTM 2008-02901). The first author was funded by an EPSRC scholarship and a grant from the Continuous Mortality Investigation. Finally, we are most grateful to the Editor for his constructive and stimulating comments at all stages of the preparation of this paper.

Supplementary Material

Supplementary materials for "Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data" by Djeundje and Currie
(doi: 10.1214/10-EJS583SUPP). The suppementary materials contain R-code to reproduce Table 1 and various Figures in the paper. A guide to using this code is also included.

References

[1] Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, **93**, 961–976. MR1649194

[2] Coull, B. A., Ruppert, D. and Wand, M. P. (2001). Simple incorporation of interactions into additive models. *Biometrics*, **57**, 539–545. MR1855689

[3] Currie, I. D., Durban, M. and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259–280. MR2188985

[4] Djeundje, V. A. B. and Currie, I. D. (2010). Supplement to "Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data." DOI: 10.1214/10-EJS583SUPP

[5] Durban, M., Harezlak, J., Wand, M. P. and Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153–1167. MR2134571

[6] de Boor, C. (2001). *A practical guide to splines*. New York: Springer. MR1900298

[7] Eilers, P. H. C. (1999). Discussion on 'The analysis of designed experiments and longitudinal data by using smoothing splines' (by A. P. Verbyla, B. R. Cullis, M. G. Kenward and S. J. Whelam). *Journal of the Royal Statistical Society, Series C*, **48**, 307–308.

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

[9] Green, P. J. (1999). Discussion on 'The analysis of designed experiments and longitudinal data by using smoothing splines' (by A. P. Verbyla, B. R. Cullis, M. G. Kenward and S. J. Whelam). *Journal of the Royal Statistical Society, Series C*, **48**, 304–305.

[10] Heckman, N., Lockhart, R. and Nielson, J. D. (personal communication). Penalized regression, mixed effects models and appropriate modelling.

[11] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

[12] Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554. MR0319325

[13] Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects Models in S and S-plus.* New York: Springer.

[14] Pinheiro, J. C., Bates, D. M., DebRoy, S., Sarkar, D. and R Core team (2009). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-96.

[15] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[16] Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, **6**, 15–51. MR1108815

[17] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757. MR1944261

[18] Ruppert, D., Wand, M.P. and Carroll, R. J. (2003). *Semiparametric regression.* Cambridge: Cambridge University Press. MR1998720

[19] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464. MR0468014

[20] Searle, S. R., Casella, G. and McCulloch, C. E. (2006). *Variance components (2nd ed.).* New York: John Wiley & Sons. MR2298115

[21] Self, S. G. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610. MR0898365

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

[23] Welham, S. J., Cullis, B. R., Kenward, M. G. and Thompson R. (2007). A comparison of mixed model splines for curve fitting. *Australian and New Zealand Journal of Statistics*, **49**, 1–23. MR2345406

[24] Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* London: Chapman and Hall. MR2206355