Please see the online version of this article for supplementary materials.

Linear Mixed-Effects Modeling by Parameter Cascading

J. CAO and J. O. RAMSAY

A linear mixed-effects model (LME) is a familiar example of a multilevel parameter structure involving nuisance and structural parameters, as well as parameters that essentially control the model's complexity. Marginalization over nuisance parameters, such as the restricted maximization likelihood method, has been the usual estimation strategy, but it can involve onerous and complex algorithms to achieve the integrations involved. Parameter cascading is described as a multicriterion optimization algorithm that is relatively simple to program and leads to fast and stable computation. The method is applied to LME, where well-developed marginalization methods are already available. Our results suggest that parameter cascading is at least as good as, if not better than, the available methods. We also extend the LME model to multicurve data smoothing by introducing a basis partitioning scheme and defining roughness penalty terms for both functional fixed effect and random effects. The results are substantially better than those obtained by using the previous LME methods. A supplemental document is available online.

KEY WORDS: Marginalization; Mixed-Effect smoothing; Model complexity; Nuisance parameters; Regularization; Variance components.

1. INTRODUCTION

This paper investigates the performance of a parameter estimation paradigm that we call *parameter cascading* for models having a multilayered parameter structure. Multilevel parameter structures arise when a large number of nuisance parameters are required to capture local effects, in addition to structural parameters with more global implications for the model whose values are of primary interest. The use of regularization or smoothing methods also introduces parameters that control the overall complexity of a model, and may therefore also be considered in a parameter layer of their own. Parameter cascading allows users to choose different optimization criteria for each parameter level, and the sequential optimization of these various criteria induces a set of functional relationships between parameters at lower layers and higher-level parameters.

We evaluate parameter cascading within the important context of linear mixed-effects (LME) structures, where a number of other methods are already available to provide performance benchmarks. Since marginalization is widely used for the elimination of nuisance parameters defining random effects for clustered data, the paper also considers whether parameter cascading brings important improvements at a computational level, and whether efficient sampling variance estimates that are not conditioned on covariance, smoothing or bandwidth estimates are possible.

Many authors have proposed that nonparametric and semiparametric smoothing methods may be recast as linear mixed-effects models, and therefore applied to data using linear mixed-effects software. Perhaps the most substantial early application of this approach to challenging data was Brumback and Rice (1998), and subsequent treatments include Rice and Wu (2001), Guo (2002), Ruppert, Wand, and Carroll (2003), Wand (2003), Morris and Carroll (2006), and Welham et al. (2006). We also look at the use of parameter cascading as an approach to this large class of problems, even though they do not strictly fall within the random effects variance components estimation framework that first give rise to LME technology. We propose a partial or fractional variance components estimation in this high-dimensional context.

Parameter cascading was first used by Ramsay et al. (2007) for the estimation of parameters defining dynamical systems, a context where an explicit expression of the model is usually not possible. Cao and Ramsay (2007, 2009) applied parameter cascading to data smoothing with adaptive regularization in order to adjust confidence intervals to take into account the uncertainty in data-determined smoothing parameters. But the central idea shows up in a wide range of older problems, including the use of profiling in nonlinear least squares problems discussed in Bates and Watts (1988), as well as many other texts, and also in the large literature on the nuisance parameter estimation problem of Neyman and Scott (1948). Parameter cascading is also obvious in regularized data smoothing and functional parameter estimation in functional data analysis (Ramsay and Silverman 2005).

The paper is organized as follows. Section 2 applies parameter cascading to variance components estimation using the linear mixed-effects model, discusses interval estimation for this situation, and compares the performance of parameter cascading to four existing methods with simulations. Section 3 develops a linear mixed-effects model with partial variance components estimation for the multiple-curve data smoothing situation. Estimating separate levels of smoothing for the fixed and random effects implies a four-level parameter cascade. A theorem is developed to capture the close formal connection between the marginalization strategy and parameter cascading. Section 4 provides a real-data analysis involving daily temperatures for 34 years of weather at Vancouver. Section 5 compares the parameter cascading method with the popular method in estimating LME smoothing models using Monte Carlo simulations. Conclusions and discussion are provided in Section 6.

Jiguo Cao is an Assistant Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A1S6 (E-mail: *jca76@sfu.ca*). James O. Ramsay is a Professor Emeritus, Psychology Department, McGill University, Montreal, Quebec, Canada H3A1B1 (E-mail: *ramsay@psych.mcgill.ca*). This paper has benefitted substantially from the comments of the editor, an associate editor, and three referees. This research was supported by grants from the Natural Science and Engineering Research Council of Canada to both authors.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Theory and Methods
DOI: 10.1198/jasa.2009.tm09124

2. PARAMETER CASCADING FOR VARIANCE COMPONENTS

We assume the two-level Gaussian linear mixed-effects model, that is, for $i = 1, \ldots, N$,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,$$
$$\mathbf{b}_i \sim \mathrm{N}(0, \Delta),\ \boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \sigma_e^2\boldsymbol{\Sigma}_i), \tag{1}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$ is a vector of measurements associated with the ith randomly sampled unit. The order q matrix Δ defines the structure and size of the between-curve variation in the random-effect vectors $\mathbf{b}_i$. Order n_i matrix $\boldsymbol{\Sigma}_i$ contains known variances and covariances among residuals. Let p and q be the lengths of $\boldsymbol{\beta}$ and $\mathbf{b}_i$, respectively. In order to simplify notation, but without loss of generality, we assume $n_i = n$ and $\boldsymbol{\Sigma}_i = \mathbf{I}$ for $i = 1, \ldots, N$.

Let $\mathbf{X}$ denote the super-matrix constructed by stacking the matrices $\mathbf{X}_i$ on top of one another, and $\mathbf{y}$ and $\boldsymbol{\epsilon}$ denote the corresponding vectors for the $\mathbf{y}_i$'s and $\boldsymbol{\epsilon}_i$'s, respectively. Let $\mathbf{Z}$ denote the diagonal super-matrix with the $\mathbf{Z}_i$ matrices in the diagonal and having Nq columns, and vector $\mathbf{b}$ denote the stacked $\mathbf{b}_i$'s. In this notation, we may replace the first equation by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$.

It is a common but perhaps confusing practice to refer to two quite different situations as a linear mixed-effects model. In the *variance components* context, Δ must be estimated from the data, implying that $q(q+1)/2$ variances and covariances are to be estimated for Δ if it is unconstrained. From a computational perspective, it is preferable to estimate the Choleski factor $\mathbf{U}$ of either Δ or Δ^{-1} in order to ensure at least positive semidefiniteness. Pinheiro and Bates (2000) opted for the decomposition of Δ^{-1}, and we follow their lead in this paper. It is not always appreciated by appliers of LME that reasonable precision in the estimates of these variance components requires that N be roughly at least 10 times $q(q+1)/2$, implying in turn that in most experimental situations q will not exceed five or so. Special structures for Δ, such as being diagonal or band-structured, permit only slightly larger values of q in practice. It is the classic variance components estimation problem that is the focus of this section.

The second LME context, which can be called *smoothing*, involves rather larger values for q, and typically arises when $\mathbf{b}_i$ is a set of coefficients for a basis function expansion of a function designed to smooth a functional observation $\mathbf{y}_i$. One is then obliged to assume that Δ is a known function of a small number of parameters. A common example is $\Delta = \mathbf{D}\sigma_b^2$, where matrix $\mathbf{D}$ is known, so that only scalar variances σ_b^2 and σ_e^2 require estimation. We turn to this context in Section 3, where we also introduce a hybrid model for variation in $\mathbf{b}_i$ where Δ is estimated within a low-dimensional subspace, and is considered known in the orthogonal complement space to within a scale factor.

From the parameter cascading perspective, the random-effect vector $\mathbf{b}$ is a nuisance parameter and the fixed-effect vector $\boldsymbol{\beta}$ is a structural parameter. However, the complexity parameter $\boldsymbol{\gamma}$ can take various forms. In the variance components situation, elements of Δ^{-1} or its Choleski factor $\mathbf{U}$ are considered complexity parameters. In the smoothing context where $\Delta = \sigma_b^2\mathbf{D}$, the complexity parameter is $\gamma = \sigma_e^2/\sigma_b^2$ or, in a form more suitable for computation, $\gamma = \log(\sigma_e^2/\sigma_b^2)$. In the hybrid model for the smoothing context that we will use in Section 3, a more complex four-level parameter cascade will be introduced involving two levels of complexity parameters.

We now assume the variance components estimation context, where Δ requires estimation, and where it is either unconstrained, or structured in such a way that multiplication by a scalar preserves the structure. We define the Choleski factorization

$$\sigma_e^2\Delta^{-1} = \mathbf{U}^T\mathbf{U}. \tag{2}$$

One referee points out that the Choleski factorization is not uniquely defined for a given positive-definite matrix, but can be made unique by requiring its diagonal elements to be all positive. So here the diagonal elements in $\mathbf{U}$ are parameterized in terms of their logarithms. The complexity parameter is the vector $\boldsymbol{\gamma}$ of length $q(q+1)/2$, which contains the logarithms of the diagonal elements of $\mathbf{U}$ and the upper off-diagonal elements of $\mathbf{U}$. The estimate of $\boldsymbol{\gamma}$ optimizes the complexity criterion defined below. Error variance σ_e^2 is estimated separately in a way specified below as a byproduct of this optimization.

2.1 The Low-Level Nuisance Criterion J

The low-level criterion J is defined as the regularized error sum of squares

$$J(\mathbf{b}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_i^N [\|\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i\|^2 + \sigma_e^2\mathbf{b}_i^T\Delta^{-1}\mathbf{b}_i]. \tag{3}$$

Criterion J is minimized by

$$\widehat{\mathbf{b}}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{Z}_i^T\mathbf{Z}_i + \sigma_e^2\Delta^{-1})^{-1}\mathbf{Z}_i^T(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

The above estimate is identical to the best linear unbiased prediction (BLUP) for random effects (Henderson 1973 and Robinson 1991). Defining the symmetric b-hat matrix $\mathbf{P}_{bi}(\boldsymbol{\gamma})$ and its complement $\mathbf{Q}_{bi}(\boldsymbol{\gamma})$ as

$$\mathbf{P}_{bi}(\boldsymbol{\gamma}) = \mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i + \sigma_e^2\Delta^{-1})^{-1}\mathbf{Z}_i^T \quad \text{and}$$
$$\mathbf{Q}_{bi}(\boldsymbol{\gamma}) = \mathbf{I} - \mathbf{P}_{bi}(\boldsymbol{\gamma}),$$

respectively, leads to the *predicted* random effects conditional on β and γ

$$\mathbf{Z}_i\widehat{\mathbf{b}}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{P}_{bi}(\boldsymbol{\gamma})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

2.2 The Midlevel Structural Criterion H

We define the midlevel criterion H as the least squares. It drops the second term in the criterion J, because the second term in J regularizes the random effects $\mathbf{b}_i$, and this regulation information is passed to the midlevel criterion H by using $\widehat{\mathbf{b}}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

$$H(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \sum_i^N \|\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\widehat{\mathbf{b}}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})\|^2$$
$$= \sum_i^N \|\mathbf{Q}_{bi}(\boldsymbol{\gamma})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\|^2.$$

The minimizer is

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\gamma}) = \mathbf{C}^{-1}(\boldsymbol{\gamma})\mathbf{d}(\boldsymbol{\gamma}),$$

where

$$\mathbf{C}(\boldsymbol{\gamma}) = \sum_i^N \mathbf{X}_i^T \mathbf{Q}_{bi}^2(\boldsymbol{\gamma})\mathbf{X}_i \qquad \text{and}$$

$$\mathbf{d}(\boldsymbol{\gamma}) = \sum_i^N \mathbf{X}_i^T \mathbf{Q}_{bi}^2(\boldsymbol{\gamma})\mathbf{y}_i.$$

This estimate of $\widehat{\boldsymbol{\beta}}$ is equivalent to the best linear unbiased estimate (BLUE) for fixed effect obtained from Henderson's equations (Henderson 1973). The fit to $\mathbf{y}_i$ is

$$\begin{aligned}\widehat{\mathbf{y}}_i &= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ &= \mathbf{Q}_{\beta i}(\boldsymbol{\gamma})\mathbf{Q}_{bi}(\boldsymbol{\gamma})\mathbf{y}_i + \mathbf{Q}_{bi}(\boldsymbol{\gamma})\mathbf{X}_i\mathbf{C}^{-1}(\boldsymbol{\gamma}) \sum_{j\neq i}^N \mathbf{X}_j^T \mathbf{Q}_{bj}^2(\boldsymbol{\gamma})\mathbf{y}_j,\end{aligned}$$

where

$$\begin{aligned}\mathbf{P}_{\beta i}(\boldsymbol{\gamma}) &= \mathbf{Q}_{bi}(\boldsymbol{\gamma})\mathbf{X}_i\mathbf{C}^{-1}(\boldsymbol{\gamma})\mathbf{X}_i^T\mathbf{Q}_{bi}(\boldsymbol{\gamma}) \qquad \text{and} \\ \mathbf{Q}_{\beta i} &= \mathbf{I} - \mathbf{P}_{\beta i}.\end{aligned} \tag{4}$$

We define the following matrices used to define the effective degrees of freedom of the model in Section 3.3.

$$\begin{aligned}\mathbf{P}_i(\boldsymbol{\gamma}) &= \mathbf{Q}_{\beta i}(\boldsymbol{\gamma})\mathbf{Q}_{bi}(\boldsymbol{\gamma}) \\ &= \mathbf{P}_{\beta i}(\boldsymbol{\gamma}) - \mathbf{P}_{\beta i}(\boldsymbol{\gamma})\mathbf{P}_{bi}(\boldsymbol{\gamma}) + \mathbf{P}_{bi}(\boldsymbol{\gamma}) \qquad \text{and} \\ \mathbf{Q}_i(\boldsymbol{\gamma}) &= \mathbf{I} - \mathbf{P}_i(\boldsymbol{\gamma}).\end{aligned}$$

2.3 The Top-Level Complexity Criterion *G*

Since fitting $\mathbf{y}$ is a linear operation indexed by $\boldsymbol{\gamma}$, it is natural to use the generalized cross-validation criterion

$$G(\boldsymbol{\gamma}) = \mathrm{GCV}(\boldsymbol{\gamma}) = nN\frac{\mathrm{SSE}(\boldsymbol{\gamma})}{\mathrm{dfe}(\boldsymbol{\gamma})^2}, \tag{5}$$

where $\mathrm{SSE} = \sum_i \|\widehat{\mathbf{y}}_i - \mathbf{y}_i\|^2$, and $\mathrm{dfe} = \sum_i \mathrm{trace}[\mathbf{Q}_i(\boldsymbol{\gamma})]$. The generalized cross-validation criterion is a measure of the trade-off between model fitting to the data and complexity of the model, and it can be also used to compare nested or nonnested LME models. Here numerical optimization is required, and we use Newton–Raphson iterations with the gradient worked out analytically.

Wahba (1985) and Gu (2002) suggest estimating the error variance σ_e^2 by

$$\hat{\sigma}_e^2 = \frac{\mathrm{SSE}(\widehat{\boldsymbol{\gamma}})}{\mathrm{dfe}(\widehat{\boldsymbol{\gamma}})}.$$

Equation (2) leads to the estimate for the variance–covariance matrix of random effects

$$\hat{\Delta} = \hat{\sigma}_e^2(\hat{\mathbf{U}}^T\hat{\mathbf{U}})^{-1}.$$

2.4 Estimation of Sampling Variances

We modify a method often used in nonlinear least squares problems for estimating the sampling variance–covariance matrix for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ (Bates and Watts 1988). This approach linearizes the mapping from the data vector $\mathbf{y}$ to the parameter space, estimates an assumed parametric distribution for the residuals $\boldsymbol{\epsilon} = \mathbf{y} - \mathrm{E}[\mathbf{y}]$, and then directly estimates the variance–covariance matrix for the parameter in question. The process might be referred to as a *linearized parametric bootstrap estimate* (Efron and Tibshirani 1993). The novel element here is the use of the Implicit Function Theorem to compute the linear approximation to the data-to-parameter map. Of course, linearization in this way is apt to break down if the actual map is severely nonlinear, and we recommend backing up interval and confidence regions estimated in this way with some selected simulations or the conventional bootstrap.

Let $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$. Since the estimate $\widehat{\boldsymbol{\gamma}}$ is an implicit functions of $\mathbf{y}$, we can approximate $\widehat{\boldsymbol{\gamma}}(\mathbf{y})$ with the first-order Taylor expansion:

$$\widehat{\boldsymbol{\gamma}}(\mathbf{y}) \approx \widehat{\boldsymbol{\gamma}}(\boldsymbol{\mu}) + \left[\frac{d\widehat{\boldsymbol{\gamma}}}{d\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}} (\mathbf{y} - \boldsymbol{\mu}). \tag{6}$$

We can take variance on both sides of (6), and obtain the approximation for the variance of $\widehat{\boldsymbol{\gamma}}$:

$$\mathrm{Var}[\widehat{\boldsymbol{\gamma}}(\mathbf{y})] \approx \left[\frac{d\widehat{\boldsymbol{\gamma}}}{d\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}} \mathrm{Var}(\mathbf{y}) \left[\frac{d\widehat{\boldsymbol{\gamma}}}{d\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}}^T.$$

Since $\widehat{\boldsymbol{\gamma}}$ is an implicit function of $\mathbf{y}$, and if

$$\mathrm{Det}\left(\frac{\partial^2 G}{\partial \boldsymbol{\gamma}^2}\bigg|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}}\right) \neq 0,$$

we can apply the Implicit Function Theorem to obtain

$$\frac{d\widehat{\boldsymbol{\gamma}}}{d\mathbf{y}} = -\left[\frac{\partial^2 G}{\partial \boldsymbol{\gamma}^2}\bigg|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}}\right]^{-1} \left[\frac{\partial^2 G}{\partial \boldsymbol{\gamma}\,\partial \mathbf{y}}\bigg|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}}\right].$$

The covariance matrix of data, $\mathrm{Var}(\mathbf{y})$, can be estimated as a block diagonal matrix with the ith diagonal block

$$\mathrm{Var}(\mathbf{y}_i) \approx \hat{\sigma}_e^2\mathbf{I} + \mathbf{Z}_i\hat{\Delta}\mathbf{Z}_i^T.$$

Since the fixed effect $\widehat{\boldsymbol{\beta}}$ is an explicit function of $\boldsymbol{\gamma}$ and $\mathbf{y}$, the unconditional variance estimate for $\widehat{\boldsymbol{\beta}}$ can also be obtained by

$$\mathrm{Var}[\widehat{\boldsymbol{\beta}}(\mathbf{y})] = \left[\frac{d\widehat{\boldsymbol{\beta}}}{d\mathbf{y}}\right] \mathrm{Var}(\mathbf{y}) \left[\frac{d\widehat{\boldsymbol{\beta}}}{d\mathbf{y}}\right]^T,$$

where the total derivative of $\widehat{\boldsymbol{\beta}}$ with respect to $\mathbf{y}$ is

$$\frac{d\widehat{\boldsymbol{\beta}}}{d\mathbf{y}} = \frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\widehat{\boldsymbol{\gamma}}}\frac{d\widehat{\boldsymbol{\gamma}}}{d\mathbf{y}} + \frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\mathbf{y}}.$$

Note that the unconditional variance estimate for $\widehat{\boldsymbol{\beta}}$ takes into account the uncertainty resulting from the estimation of $\widehat{\boldsymbol{\gamma}}$ and the data $\mathbf{y}$, in contrast to the usual practice of plugging the estimate $\widehat{\boldsymbol{\gamma}}$ into sampling variance estimates as if $\widehat{\boldsymbol{\gamma}}$ was not random.

2.5 Simulation Results for a Typical Design

We now use simulations to compare the performance of parameter cascading (PC) with the positive-definiteness-constrained restricted maximum likelihood estimates (LME-REML) provided by the `lme()` function in both R and S-PLUS, and with the unconstrained maximum likelihood (UNC-ML) and unconstrained restricted likelihood estimation (UNC-REML) for the linear mixed effects models. Since the design is balanced ($p = q$ and n_i constant), there are closed forms for unconstrained UNC-ML estimates and UNC-REML estimates, but these offer no protection against violation of positive semidefiniteness. We used default convergence criteria for `lme()` so as to assess the performance of `lme()` as it would be used in most applications.

It is important to note that we, along with Demidenko (2004), experienced an unacceptable rate of failures to convergence (36% when $N = 10$ in our case) with the default optimization function `nlminb` in the R version of the function `lme`, but when we followed the suggestion of a referee and specified the `optim()` function, this problem disappeared. We cannot recommend the use of the default optimizer in R, and hope that this serious issue is rectified soon. The S-PLUS version, on the other hand, performed perfectly.

Demidenko (2004) offered some simulation results for a linear mixed-effects model defined in (1) with a balanced design. His parameter values were $\beta_1 = 1$, $\beta_2 = 0.1$, $\sigma_e^2 = 1$, $\mathbf{\Sigma}_i = \mathbf{I}$ for $i = 1, \ldots, N$,

$$\mathbf{X}_i = \mathbf{Z}_i = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix}^T \quad \text{for } i = 1, \ldots, N$$

and

$$\Delta = \begin{pmatrix} 1.0 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}.$$

These settings resembled the types and values of linear mixed models often reported in the applied literature. We repeated this experiment using $N = 10$, 20, 50, and 100.

The point estimates for the fixed effects with respect to mean, standard deviation, and mean squared error differed by no more than one in the third decimal place for all methods and for all sample sizes N. The differences among methods for the estimation of σ_e were also trivial for the three methods assuring positive semidefiniteness of $\hat{\Delta}$. But UNC-REML and UNC-MLE $\hat{\sigma}_e$ values had better mean bias (mean 0.988 as opposed to 0.963) and slightly worse mean squared errors (0.026 versus 0.022) when $N = 10$, but even these differences disappeared by $N = 20$.

Summaries for point estimations for $N = 10$ for the random effects variance covariance matrix estimates $\hat{\Delta}$ are shown in Table 1. UNC-REML estimates have the best bias, but both UNC-REML and UNC-ML estimates have worse mean squared errors due to their larger sampling variances. Even so, the amount of bias associated with the other three methods is fairly small considering the small sample size. The results for $N = 20$, 50, 100 are given in Tables 2–4 in the supplementary file. The bias in the Δ estimates considerably improves by $N = 20$, and is negligible when $N = 50$ and 100.

However, along with Demidenko (2004), we note a large percentage of Δ estimates with negative eigenvalues for UNC-ML and UNC-REML. For example, percentages of negative eigenvalue cases were 55%, 50%, 35%, and 20% for $N = 10$, 20, 50, and 100, respectively, for UNC-REML; UNC-ML gives slightly larger percentages of negative eigenvalue cases.

Table 1. The means, standard deviations (STD), and mean squared errors (MSE) for point estimations of linear mixed-effects models over 1000 simulations using five methods when the number of subjects $N = 10$

	True		PC	R	S-Plus	UNC-REML	UNC-ML
Δ_{11}	1.000	Mean	1.228	1.232	1.227	0.974	0.726
		STD	1.029	1.030	1.030	1.219	1.115
		MSE	1.110	1.113	1.112	1.485	1.316
Δ_{12}	0.300	Mean	0.192	0.188	0.191	0.278	0.300
		STD	0.411	0.409	0.410	0.476	0.434
		MSE	0.180	0.179	0.180	0.226	0.188
Δ_{22}	0.500	Mean	0.549	0.551	0.550	0.518	0.446
		STD	0.306	0.326	0.326	0.340	0.308
		MSE	0.096	0.109	0.109	0.116	0.098

NOTE: "PC," "R," "S-Plus," "UNC-REML," and "UNC-ML" stand for parameter cascading, the `lme()` function in R using the "optim" optimization option, the `lme()` function in S-Plus, unconstrained restricted maximum likelihood, and unconstrained maximum likelihood, respectively.

Table 2 shows the performance of the standard error estimates for fixed effects with the five methods for $N = 10$. UNC-REML and UNC-ML have worse coverage probability and have a slight tendency to underestimate the standard error. The five methods have negligible differences when $N = 20$, 50, 100, which are displayed in Table 5 of the supplementary file.

In order to increase the range of design and parameter characteristics for which conclusions can be drawn from our simulations, we repeated the simulation study for the parameter cascading method by allowing the constants N and n and parameters $\beta_1, \beta_2, \sigma_e^2$ and Δ generated independently and randomly. The relative performance of the parameter cascading method remained very much as shown above. The result details are given in Section 1.2 of the supplementary file.

3. PARAMETER CASCADING FOR LME SMOOTHING

The LME model (1) can be adapted to a functional data context in which the data y_{ij} arise from a sample of n discrete and noisy observations of each of N realizations of smooth underlying Gaussian random processes. The LME smoothing model can be written as follows:

$$\mathbf{y}_i = \mu(\mathbf{t}_i) + r_i(\mathbf{t}_i) + \boldsymbol{\epsilon}_i, \tag{7}$$

where $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_i$ are defined as in (1), and $\mathbf{t}_i$ is the corresponding vector of measurement locations. The random effects are now defined as smooth variations $r_i(t)$ around a fixed-effect function $\mu(t)$, and these functions can be represented by the basis function expansions

$$\mu(t) = \boldsymbol{\beta}^T \boldsymbol{\varphi}(t) \qquad \text{and} \qquad r_i(t) = \mathbf{b}_i^T \boldsymbol{\phi}(t),$$

where the coefficient vectors and corresponding vectors of basis functions are of length p for fixed-effect function μ and, for simplicity only, will be assumed here to be of length q for all random-effect functions r_i. Design matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ are now constructed by evaluating basis functions at the points $\mathbf{t}_i$ of observation associated with data vector $\mathbf{y}_i$. Hybrid designs may

Table 2. The means, standard deviations (STD), and mean squared errors (MSE) for standard error (SE) estimations of fixed effects over 1000 simulations when the number of subjects $N = 10$

	$\widehat{SE}(\hat{\beta}_1)$				$\widehat{SE}(\hat{\beta}_2)$			
Sample STD	0.4847				0.2742			
Method	Mean	STD	MSE	$CP(\beta_1)$	Mean	STD	MSE	$CP(\beta_1)$
PC	0.503	0.100	0.010	95.1%	0.271	0.054	0.003	93.1%
R	0.505	0.100	0.010	94.6%	0.266	0.059	0.004	91.9%
S-PLUS	0.505	0.100	0.010	94.5%	0.266	0.060	0.004	91.8%
UNC-REML	0.484	0.114	0.013	92.9%	0.261	0.062	0.004	90.9%
UNC-ML	0.460	0.108	0.012	91.4%	0.247	0.058	0.004	89.4%

NOTE: "PC," "R," "S-Plus," "UNC-REML," and "UNC-ML" stand for parameter cascading, the `lme()` function in R using the "optim" optimization option, the `lme()` function in S-Plus, restricted maximum likelihood, and maximum likelihood, respectively. $CP(\beta_1)$ and $CP(\beta_2)$ are the coverage probabilities of 95% confidence intervals for $\hat{\beta}_1$ and $\hat{\beta}_2$. The 95% confidence intervals is constructed as $[\hat{\beta}_j - 1.96 * \widehat{SE}(\hat{\beta}_j), \hat{\beta}_j + 1.96 * \widehat{SE}(\hat{\beta}_j)], j = 1, 2$.

also arise where these design matrices are augmented by conventional covariates and interaction terms, a situation that we may refer to semiparametric regression for samples of curves.

There is no need for $p = q$ or $\mathbf{X}_i = \mathbf{Z}_i$, although this will often be the case. The simple model used to simulate data in the previous section is in fact just such a situation, where $p = q = 2$ and the monomial basis is used for both sets of functions. However, identifiability issues require special treatment for the balanced design where $\mathbf{X}_i$ and $\mathbf{Z}_i$ do not vary over records and where, in addition, $p = q \approx n$.

In the functional context, moreover, it can be desirable and even essential to impose smoothness on the estimated fixed-effect function $\mu(t)$ as well as on the estimates of the functional random effects $r_i(t)$. It can be important to choose these two smoothing levels independently.

A special feature of LME smoothing is that the complexity of each curve, captured by the size of p and q, can imply that N will not likely be large enough relative to $q(q+1)/2$ to permit estimating an unrestricted intercurve variance–covariance matrix Δ. The usual practice in LME smoothing, illustrated in Brumback and Rice (1998), is to represent intercurve variation by a scalar multiplying a symmetric positive definite matrix assumed to be known, but we propose a compromise that allows for estimating a variance–covariance structure for low-dimensional functional variation combined with a data-defined level of smoothing of random functional effects orthogonal to this low-dimensional subspace. To enable this, we first consider some practical aspects of constructing partitioned basis systems.

3.1 Basis Partitioning

Because the LME smoothing context can involve large values of q, we have to settle for the more realistic objective of studying in detail the functional variation of the random-effect functions r_i within a low-dimensional subspace of dimension s, where the size of s will depend on the number N of curves that we have available to estimate this variation.

Much of the early theoretical work on smoothing, represented in Wahba (1990) and Gu (2002), assumes a basis system partitioned into a low-dimensional subspace defined as the kernel of a linear differential operator and a high dimensional complement with an inner product defined by this operator, thus defining a reproducing kernel Hilbert space. Although this is a powerful framework, daunting mathematical detail and numerical analysis issues have hindered its application. We offer here a more convenient approach involving a high-dimensional q-vector of basis functions $\boldsymbol{\phi}(t)$ and a low-dimension s-vector $\boldsymbol{\theta}(t) = (\theta_1(t), \ldots, \theta_s(t))$, $s \ll q$ such that all functional variation of a specified character is concentrated with the span of $\boldsymbol{\theta}(t)$.

The basis *sweep* or left-division operator $\boldsymbol{\theta}\backslash\boldsymbol{\phi}$ is defined to be

$$\boldsymbol{\theta}\backslash\boldsymbol{\phi} = \boldsymbol{\phi} - \mathbf{G}\boldsymbol{\theta}, \quad \text{where} \quad \mathbf{G} = \left[\int \boldsymbol{\phi}(t)\boldsymbol{\theta}^T(t)\,dt\right]\left[\int \boldsymbol{\theta}(t)\boldsymbol{\theta}^T(t)\,dt\right]^{-1}. \tag{8}$$

The q by s matrix $\mathbf{G}$ is the matrix of regression coefficients for a functional regression of $\boldsymbol{\phi}$ on $\boldsymbol{\theta}$, and, consequently, $\int \boldsymbol{\theta}(t)[\boldsymbol{\theta}\backslash\boldsymbol{\phi}(t)]^T\,dt = 0$. Because the q-vector $\boldsymbol{\theta}\backslash\boldsymbol{\phi}$ now spans a space of dimension $q - s$, the reduced vector $\boldsymbol{\psi} = \mathbf{V}(\boldsymbol{\theta}\backslash\boldsymbol{\phi})$, where $q - s$ by q matrix $\mathbf{V}$ contains in its rows the first $q - s$ eigenvectors of the matrix $\int [\boldsymbol{\theta}\backslash\boldsymbol{\phi}(t)][\boldsymbol{\theta}\backslash\boldsymbol{\phi}(t)]^T\,dt$, is more useful. In other words, $\boldsymbol{\psi}$ is computed from $\boldsymbol{\theta}\backslash\boldsymbol{\phi}$ by principal components analysis, and the basis functions in $\boldsymbol{\psi}$ are defined in this way such that they are orthogonal to each other as well as to $\boldsymbol{\theta}$.

Let L be a linear differential operator that defines the total roughness $\int [Lr_i(t)]^2\,dt$ of the ith random-effect function. The small number of basis functions $\theta_\ell(t)$ in $\boldsymbol{\theta}(t)$ are chosen as the basis functions satisfying $L\theta_\ell(t) = 0$. For example, when the linear differential operator is the second differentiator, that is, $L\theta_\ell(t) = d^2\theta_\ell(t)/dt^2$, the two basis functions in $\boldsymbol{\theta}(t)$ are the monomials $\theta_1(t) = 1$ and $\theta_2(t) = t$. Figure 1 shows in the top panel the consequences of sweeping the monomials $\theta_1(t) = 1$ and $\theta_2(t) = t$ from eight B-spline basis functions defined by equally spaced knots. The bottom panel shows the six basis functions $\boldsymbol{\psi}(t)$ resulting from the PCA compression.

The integrals in (8) can be approximated by using the `inprod` function provided in the "fda" package in R, and also available in Matlab from website *www.functionaldata.org*. The basis defined by Demmler and Reinsch (1975) and used by Brumback and Rice (1998) is closely related, but differs in terms of being defined by matrix operations applied to matrices of basis values defined at data observation points $\mathbf{t}_i$, and therefore is specific to the data design.

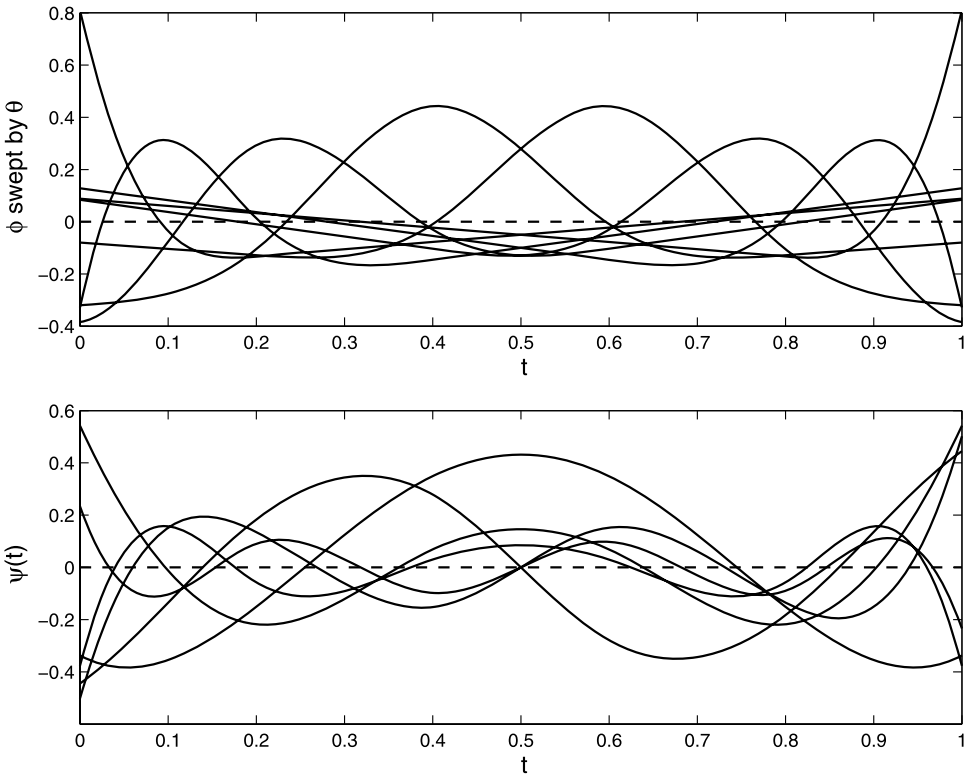

Figure 1. The top panel shows the eight B-splines $\boldsymbol{\phi}(t)$ defined by equally spaced knots over [0, 1] swept by functions $\theta_1(t) = 1$ and $\theta_2(t) = t$, and the bottom panel shows these swept functions compressed to six basis functions $\boldsymbol{\psi}(t)$ by a functional principal components analysis.

3.2 LME Smoothing Using a Partitioned Basis

We now assume a partitioned basis $\boldsymbol{\phi}(t) = (\boldsymbol{\theta}(t)^T, \boldsymbol{\psi}(t)^T)^T$ having components of dimension s and $q - s$, respectively. The corresponding order q coefficient vector $\mathbf{b}_i$ defining functional random component $r_i(t) = \mathbf{b}_i^T \boldsymbol{\phi}(t)$ is partitioned into corresponding components $\mathbf{b}_i = (\mathbf{b}_{i0}^T, \mathbf{b}_{i1}^T)^T$, as is the design matrix $\mathbf{Z}_i = [\mathbf{Z}_{i0}, \mathbf{Z}_{i1}]$.

Let L be a linear differential operator, such as D^2, that defines the total roughness $\int [Lr_i(t)]^2\, dt$ of the ith random-effect function. We assume that $L\boldsymbol{\theta}(t) = 0$ and define the order q roughness penalty matrix as $\mathbf{R}_b = \int L\boldsymbol{\phi}(t) L\boldsymbol{\phi}^T(t)\, dt$, so that $\int [Lr_i(t)]^2\, dt = \mathbf{b}_i^T \mathbf{R}_b \mathbf{b}_i = \mathbf{b}_{i1}^T \mathbf{R}_\psi \mathbf{b}_{i1}$, where the order $q - s$ roughness penalty matrix $\mathbf{R}_\psi = \int L\boldsymbol{\psi}(t) L\boldsymbol{\psi}^T(t)\, dt$. Because we assume that $L\boldsymbol{\theta}(t) = 0$, only the lower right order $q - s$ submatrix of $\mathbf{R}_b$ will be nonzero, which is $\mathbf{R}_\psi$. For example, suppose $\boldsymbol{\phi}(t)$ is a vector of 73 Fourier basis functions with the period 1. When the linear differential operator L is the harmonic acceleration operator $L\boldsymbol{\phi}(t) = (2\pi)^2 D\boldsymbol{\phi}(t)/dt + d^3\boldsymbol{\phi}(t)/dt^3$, the associated partition will be $\boldsymbol{\theta}(t) = (1, \sin(2\pi t), \cos(2\pi t))^T$ and $\boldsymbol{\psi}(t) = (\sin(2\pi kt), \cos(2\pi kt), k = 2, \ldots, 36)^T$.

We now define the inverse variance component matrix Δ^{-1} in Section 2 to be a block diagonal matrix

$$\Delta^{-1} = \begin{pmatrix} \mathbf{U}^T\mathbf{U} & \mathbf{0}_{s\times(q-s)} \\ \mathbf{0}_{(q-s)\times s} & \exp(\eta_b)\mathbf{R}_\psi \end{pmatrix}. \tag{9}$$

Now $\mathbf{U}^T\mathbf{U}$ represents the inverse intercurve variation in the low-dimension space spanned by $\boldsymbol{\theta}(t)$, while the scalar $\exp(\eta_b)$ captures only inverse scale variation in the complementary space spanned by $\boldsymbol{\psi}(t)$. Using this definition, the low-level fitting nuisance criterion J in (3) remains unchanged.

It may also be desirable to control the smoothness of the fixed-effect function μ by also applying a roughness penalty. That is, we may elect to extend the midlevel criterion to

$$H(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\widehat{\mathbf{b}}\|^2 + \exp(\eta_\beta)\boldsymbol{\beta}^T\mathbf{R}_\beta\boldsymbol{\beta},$$

where $\mathbf{R}_\beta$ is an order p roughness penalty matrix, possibly defined in the same manner as $\mathbf{R}_b$.

Smoothing parameters η_b and η_β, along with triangular variance component $\mathbf{U}$ are all complexity controllers. However, we now see that η_β controls complexity at a higher or more global level than do η_b and $\mathbf{U}$. This suggests a fourth level in our parameter hierarchy, which is now $\mathbf{b}_i(\boldsymbol{\beta}, \eta_b(\eta_\beta), \mathbf{U}(\eta_\beta))(\mathbf{y})$.

This fourth level is optional, and it may be considered adequate to impose smoothness on the fixed-effect function by keeping p, the dimension of $\boldsymbol{\beta}$, small. Nevertheless, we consider the possibility that $p = q$ and, moreover, that p and q could approach, equal, or even exceed n, the number of observations per curve.

However, we have observed an identifiability issue in this case that interfered with stable numerical optimization and caused other computation problems. It is possible that the fixed-effect curve estimate $\widehat{\boldsymbol{\mu}}$ may bleed a certain amount of shape variation off to the random effect estimates $\hat{r}_i$, and therefore become seriously biased. To avoid this, we now extend the LME framework to permit the constraint $\sum_i \hat{r}_i(t) = 0$ for all t, as well as, in principle, other constraints.

Assuming the balanced design case where n, $\mathbf{X}_i$ and $\mathbf{Z}_i$ do not vary over i, let the q by N matrix $\mathbf{B}$ contain in its columns the

random coefficient vectors $\mathbf{b}_i$. Let the M by N matrix $\mathbf{W}$, where $M \le N$, be a design matrix representing a linearly constrained structure in $\mathbf{B}$ through the equation $\mathbf{B} = \mathbf{AW}$, where the q by M matrix $\mathbf{A}$ now contains unrestricted coefficient vectors. We specifically have in mind the constraint $\mathbf{W1}_N = 0$ where the N-vector $\mathbf{1}_N$ contains only one's. This ensures that $\sum_i b_{ij} = 0$ for all j, and therefore that the functional random effects $r_i(t)$ defined with the coefficient vectors $\mathbf{b}_i$ sum to 0 for any t.

Define the Mq by Nq super-matrix $\widetilde{\mathbf{W}} = (\mathbf{I}_q \otimes \mathbf{W})$. Writing Mq-vector $\mathbf{a} = \mathrm{vec}(\mathbf{A})$, we then have that $\mathbf{b} = \widetilde{\mathbf{W}}^T\mathbf{a}$. Finally, our notation is made more compact by defining Nn by Mq super-matrix $\widetilde{\mathbf{Z}} = (\mathbf{W}^T \otimes \mathbf{Z})$ and order Nq super-matrix $\tilde{\Delta}^{-1} = (\mathbf{1}_N \otimes \sigma_e^2\Delta^{-1})$.

We can now reexpress the low-level criterion (3) as

$$J(\mathbf{a}|\boldsymbol{\beta}, \eta_b, \eta_\beta, \mathbf{U}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \widetilde{\mathbf{Z}}\mathbf{a}\|^2 + \mathbf{a}^T\widetilde{\mathbf{W}}\tilde{\Delta}^{-1}\widetilde{\mathbf{W}}^T\mathbf{a}.$$

Its minimizer is

$$\widehat{\mathbf{a}}(\boldsymbol{\beta}, \eta_b, \eta_\beta, \mathbf{U}) = [\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}^T + \widetilde{\mathbf{W}}\tilde{\Delta}^{-1}\widetilde{\mathbf{W}}^T]^{-1}\widetilde{\mathbf{Z}}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Define the order nN random-effect smoothing operator $\mathbf{P}_b$ and its complement $\mathbf{Q}_b$ as

$$\mathbf{P}_b = \widetilde{\mathbf{Z}}[\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}^T + \widetilde{\mathbf{W}}\tilde{\Delta}^{-1}\widetilde{\mathbf{W}}^T]^{-1}\widetilde{\mathbf{Z}}^T \qquad \text{and}$$

$$\mathbf{Q}_b = \mathbf{I}_{nN} - \mathbf{P}_b,$$

respectively. We now have

$$\min_{\mathbf{a}}\{J(\mathbf{a}|\boldsymbol{\beta}, \eta_b, \eta_\beta, \mathbf{U})\} = \|\mathbf{Q}_b(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 + \widehat{\mathbf{a}}^T\widetilde{\mathbf{W}}\tilde{\Delta}^{-1}\widetilde{\mathbf{W}}^T\widehat{\mathbf{a}}.$$

The midlevel criterion $\mathbf{H}(\boldsymbol{\beta})$ is now

$$\begin{aligned}\mathbf{H}(\boldsymbol{\beta}|\eta_b, \eta_\beta, \mathbf{U}) &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \widetilde{\mathbf{Z}}\widehat{\mathbf{a}}\|^2 + \exp(\eta_\beta)\boldsymbol{\beta}^T\mathbf{R}_\beta\boldsymbol{\beta} \\ &= \|\mathbf{Q}_b(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 + \exp(\eta_\beta)\boldsymbol{\beta}^T\mathbf{R}_\beta\boldsymbol{\beta}. \qquad (10)\end{aligned}$$

Its minimizer is

$$\widehat{\boldsymbol{\beta}}(\eta_b, \eta_\beta, \mathbf{U}) = [\mathbf{X}^T\mathbf{Q}_b^2\mathbf{X} + \exp(\eta_\beta)\mathbf{R}_\beta]^{-1}\mathbf{X}^T\mathbf{Q}_b^2\mathbf{y}.$$

Letting order nN super-matrices

$$\mathbf{P}_\beta = \mathbf{Q}_b\mathbf{X}[\mathbf{X}^T\mathbf{Q}_b^2\mathbf{X} + \exp(\eta_\beta)\mathbf{R}_\beta]^{-1}\mathbf{X}^T\mathbf{Q}_b \qquad \text{and}$$

$$\mathbf{Q}_\beta = \mathbf{I}_{nN} - \mathbf{P}_\beta,$$

we can express the minimizer of (10) as

$$\min_{\boldsymbol{\beta}}\{H(\boldsymbol{\beta}|\eta_b, \eta_\beta, \mathbf{U})\} = \|\mathbf{Q}_\beta\mathbf{Q}_b\mathbf{y}\|^2 + \exp(\eta_\beta)\widehat{\boldsymbol{\beta}}^T\mathbf{R}_\beta\widehat{\boldsymbol{\beta}}.$$

If we define the order nN operator matrix $\mathbf{P}$ as

$$\mathbf{P}(\eta_b, \eta_\beta) = \mathbf{P}_\beta(\eta_\beta) - \mathbf{P}_\beta(\eta_\beta)\mathbf{P}_b(\eta_b) + \mathbf{P}_b(\eta_b) \qquad \text{and}$$

$$\mathbf{Q} = \mathbf{I}_{nN} - \mathbf{P},$$

then the fit to $\mathbf{y}$ is $\widehat{\mathbf{y}} = \mathbf{Py}$. Then the third-level optimality criterion (5) remains essentially unchanged as

$$\mathrm{GCV}_b(\boldsymbol{\gamma}|\eta_\beta) = nN\frac{\|\mathbf{Q}(\boldsymbol{\gamma}|\eta_\beta)\mathbf{y}\|^2}{\mathrm{trace}[\mathbf{Q}(\boldsymbol{\gamma}|\eta_\beta)]^2}, \quad \text{where}$$

$$\boldsymbol{\gamma} = [\eta_b, \mathrm{vec}(\mathbf{U})^T]^T.$$

If required, the fourth-level optimality criterion that is minimized with respect to the fixed-effect smoothing parameter η_β is

$$\mathrm{GCV}_\beta(\eta_\beta) = nN\frac{\|\mathbf{Q}(\eta_\beta, \widehat{\boldsymbol{\gamma}}(\eta_\beta))\mathbf{y}\|^2}{\mathrm{trace}[\mathbf{Q}(\eta_\beta, \widehat{\boldsymbol{\gamma}}(\eta_\beta))]^2}. \qquad (11)$$

3.3 Parameter Cascading and Marginalization

Many statistical models have some parameters not of direct interest, which are called *nuisance parameters*, for example, the random-effect parameters $\mathbf{b}$ in the LME model. Some parameters, holding the primary interest, are called *structural parameters*, for example the fixed-effect parameters $\boldsymbol{\beta}$ in the LME model. The usual approach to the estimation of structural parameters $\boldsymbol{\beta}$ in the presence of nuisance parameters $\mathbf{b}$ is to eliminate them by marginalization with respect to a measure $\pi(\mathbf{b})$. That is, given a joint likelihood $L_J(\boldsymbol{\beta}, \mathbf{b}|\mathbf{y})$, we optimize the marginal likelihood

$$L_M(\boldsymbol{\beta}) = \int_{\mathcal{C}} L_J(\boldsymbol{\beta}, \mathbf{b}|\mathbf{y})\pi(\mathbf{b})\, d\mathbf{b}. \qquad (12)$$

There is a close formal connection between the marginalization strategy and parameter cascading that is captured by the following theorem. This theorem holds for any statistical model with the nuisance parameters $\mathbf{b}$ and structural parameters $\boldsymbol{\beta}$, not limited to just the LME models.

Theorem 3.1. Assume that the joint likelihood L_J is continuous with respect to $\mathbf{b} \in \mathcal{C}$ for all $\boldsymbol{\beta} \in \mathcal{T}$, $\mathcal{T}$ is compact and $\mathcal{C}$ is closed. Then there exists at least one function $\widehat{\mathbf{b}}(\boldsymbol{\beta})$ and a fitting criterion $H(\boldsymbol{\beta}, \widehat{\mathbf{b}}(\boldsymbol{\beta})|\mathbf{y})$ such that

$$\arg\max_{\boldsymbol{\beta}} H(\boldsymbol{\beta}, \widehat{\mathbf{b}}(\boldsymbol{\beta})|\mathbf{y}) = \arg\max_{\boldsymbol{\beta}} L_M(\boldsymbol{\beta}|\mathbf{y}).$$

Proof. By the integral version of the multivariate mean value theorem, there exists for any $\boldsymbol{\beta} \in \mathcal{T}$ at least one value $\mathbf{b}^*(\beta)$ such that

$$L_M(\boldsymbol{\beta}) = \int_{\mathcal{C}} L_J(\boldsymbol{\beta}, \mathbf{b}|\mathbf{y})\pi(\mathbf{b})\, d\mathbf{b} = L_J(\boldsymbol{\beta}, \mathbf{b}^*|\mathbf{y})\int_{\mathcal{C}} \pi(\mathbf{b})\, d\mathbf{b}.$$

Although multiple solutions to this equation may exist for some values $\boldsymbol{\beta}$, we can always construct a function such that $\widehat{\mathbf{b}}(\boldsymbol{\beta}) = \mathbf{b}^*$ over $\mathcal{T}$. The specification of such a function does not have any impact on the identification $H(\boldsymbol{\beta}, \widehat{\mathbf{b}}(\boldsymbol{\beta})|\mathbf{y}) = L_J(\boldsymbol{\beta}, \mathbf{b}^*|\mathbf{y})$, and the theorem follows.

Moreover, H also arises from an application of a regularization process. Because L_M as defined in (12) is constructed from L_J by a linear operation, it can be reexpressed as a solution of a functional quadratic optimization problem, namely,

$$L_M = \arg\max_f \int_{\mathcal{C}} [L_J(\cdot, \mathbf{b}|\mathbf{y}) - f]^2 \exp(\ln\pi(\mathbf{b}) + C)\, d\mathbf{b},$$

where f is a real-valued function on $\mathcal{T}$ and C is an arbitrary constant. That is, L_M is a smoothed likelihood where localized variation in $\mathbf{b}$ has been removed.

In this sense, then, the principal contrast between parameter cascading and marginalization is in terms of the computational overhead implied by integration in (12), and parameter cascading can be viewed as a strategy for bypassing computationally intensive integration strategies such as Markov chain Monte Carlo.

4. LME SMOOTHING FOR VANCOUVER TEMPERATURES

Figure 2 displays the daily temperatures in Vancouver on every fifth day for 34 years over the period 1961–1994 (N = 34 and n = 73). For our parameter cascading analyses, both the fixed effect $\mu(t)$ and the random effects $r_i(t)$ were expanded using 73 Fourier basis functions. These were partitioned into $\boldsymbol{\phi}(t) = (\boldsymbol{\theta}(t)^T, \boldsymbol{\psi}(t)^T)^T$, where the kernel basis functions in $\boldsymbol{\theta}(t)$ are $\theta_1(t) = 1, \theta_2(t) = \sin(\omega t), \theta_3(t) = \cos(\omega t)$, and the complement basis functions in $\boldsymbol{\psi}$ are $\psi_{2k-3}(t) = \sin(\omega k t)$ and $\psi_{2k-2}(t) = \cos(\omega k t), k = 2, \ldots, 36$, and $\omega = 2\pi/365$. Since the fixed and random effects are periodic functions, the roughness penalty for both fixed and random effects is defined by the harmonic acceleration operator $Lx(t) = \omega^2\, dx(t)/dt + d^3x(t)/dt^3$.

The R version of the LME-REML method was also used, and for this analysis the fixed effect $\mu(t)$ was expanded using 73 Fourier basis functions, but the random effects $r_i(t)$ were defined in two ways in order to vary the complexity of the model. Option (1) expanded $r_i(t)$ using the three kernel basis functions in $\boldsymbol{\theta}(t)$ and with Σ unconstrained except for being positive definite. Option (2) expanded $r_i(t)$ with 73 Fourier basis functions, but with the covariance matrix being block-diagonal: the order 3 covariance of the kernel function coefficients was unconstrained-positive-defiinite, and the covariance matrix for the coefficients of the remaining 70 basis functions in $\boldsymbol{\psi}(t)$ was a positive multiple of the identity matrix.

The parameter cascading method estimated $\hat{\eta}_\beta = -1.85$, which corresponding to the smoothing parameter value $\exp(\hat{\eta}_\beta) = 0.16$ for the functional fixed effect $\mu(t)$. The estimated fixed effect $\mu(t)$ is shown in Figure 3 as a solid line, and is quite smooth. The standard error estimate was $\hat{\sigma}_e = 2.6$ degrees Celsius, a value quite consistent with known day-to-day temperature variation.

The smoothing parameter estimate for the random effects $r_i(t)$ was $\exp(\hat{\eta}_b) = 0.01$, implying that a substantial amount of random-effect variation was captured by the high-frequency basis functions in $\boldsymbol{\psi}(t)$. Figure 4 displays in both the kernel basis

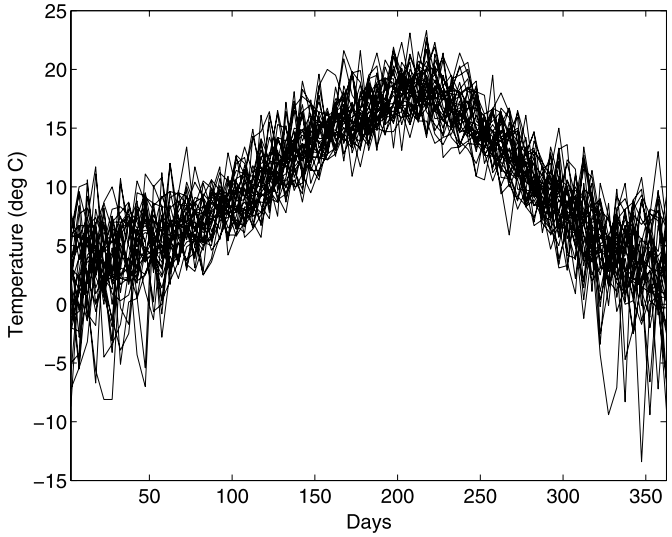

Figure 2. The temperature in degrees Celsius at Vancouver recorded every fifth day for the years 1961 to 1994.

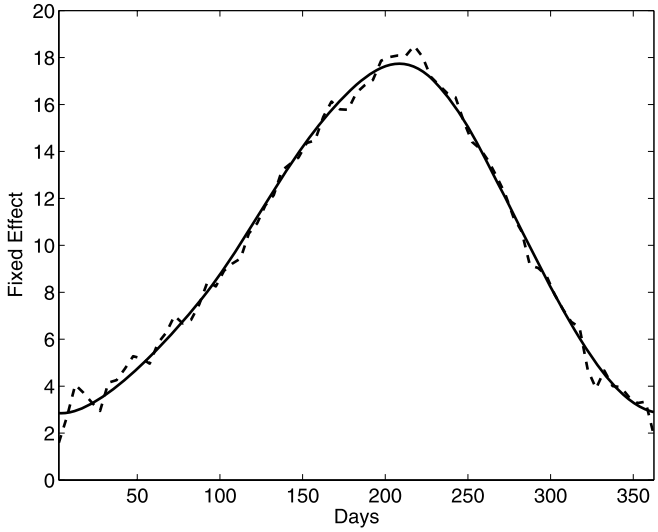

Figure 3. The solid line is the parameter cascading estimate for the fixed effect $\mu(t)$. The dashed line is the LME-REML fixed-effect estimate that is common to Models (1) and (2).

random effects expanded with $\boldsymbol{\theta}(t)$, and the complement basis functions with $\boldsymbol{\psi}(t)$. We see that there are important effects with periods of half a year possessing considerable phase variation, and that a few calendar years have unusually large variation in the complement space.

The LME-REML method does not permit independent control of the smoothness of the fixed effect, and the estimate for both of its random-effect models turned out to be that shown as the rather rough dashed line in Figure 3. The standard error estimates for the two models were 2.6 and 0.93 degrees for Models (1) and (2), respectively. The LME-REML random-effect estimates for the low-dimensional random effect Model (1) resembled the PC kernel effects shown in Figure 4, but since this model had no capacity to represent higher-frequency effects, it tended to over-smooth or under-fit the data. The LME-REML random-effect estimates for Model (2) came close to interpolating the differences between the data and the fixed effect, and therefore strongly under-smoothed or over-fit the data, which is reflected by the unrealistically small value of $\hat{\sigma}_e$.

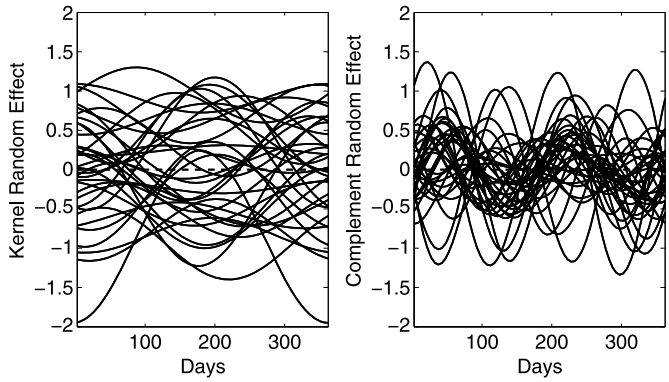

Figure 4. The left panel displays the random effects expanded with the kernel basis functions in $\boldsymbol{\theta}(t)$. The right panel displays the random effects expanded with the complement basis functions in $\boldsymbol{\psi}(t)$.

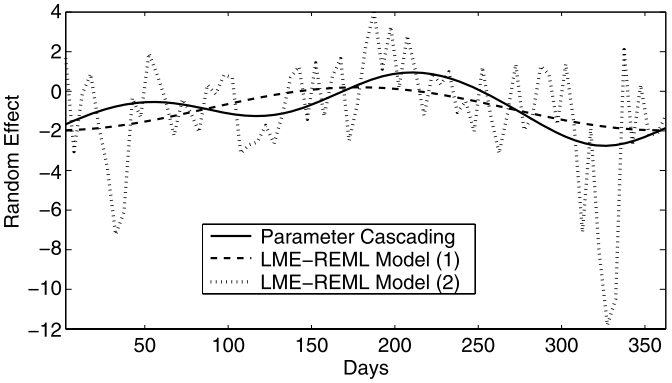

Figure 5. The estimated random effects $r_i(t)$ for year 1985.

Figure 5 shows the estimated random effects $r_i(t)$ for the year 1985. Parameter cascading yields an estimate that reflects variation in both the low-frequency and higher-frequency domains, as one might expect. The fitted curve using LME-REML for Model (1) is the closest to the estimated fixed effect in most of the time interval because it underestimates the random effect, an seems too smooth. The fitted curve using LME-REML for Model (2) has large variation because of under-smoothing. Nevertheless, all three random effects suggest that Vancouver's December in 1985 was unusually cold.

5. SIMULATIONS FOR LME SMOOTHING

Simulation was used to compare parameter cascading with LME-REML in estimating the LME smoothing model (7). The true functional fixed effect $\mu(t)$ and random effects $r_i(t)$ are fixed to be the parameter cascading estimates from the real Vancouver temperature data analyzed in Section 4. We generated 100 simulated datasets by adding the measurement error $\boldsymbol{\epsilon}_i \sim \text{Normal}(\mathbf{0}_{n\times 1}, \sigma_e^2 \mathbf{I}_{n\times n})$. We set the number of curves $N = 34$, the number of equally spaced observations $n = 73$ in the time interval [1, 365], and $\sigma_e = 2.6$, which are all consistent with the Vancouver temperature example. The basis functions used for the parameter cascading method are the same as those used in the Vancouver temperature example. The LME-REML method was used to estimate Models (1) and (2) described in Section 4.

Table 3 shows the the bias, STD and RMSE, averaged over 73 time points, for the fixed effect and random-effect estimates. For the fixed-effect estimates, the parameter cascading method reduces the average bias, STD, RMSE by 75%, 94%, and 91%, respectively, when compared with REML. For the random-effect estimates, Model (1) has a smaller RMSE than Model (2), but has a slightly larger bias than Model (2). This makes sense because Model (2) has a larger number of basis functions used for expanding the random effects. On the other hand, parameter cascading estimation of random effects reduces the average bias, STD and RMSE by 72%, 69%, and 70%, respectively, when compared with the LME-REML estimates for Model (1), or by 50%, 96%, and 94%, respectively, when compared with the LME-REML estimates for Model (2).

Table 3. Simulation results for 100 simulated samples using parameter cascading and LME-REML for Models (1) and (2)

	Fixed effect			Random effects		
	$\lvert BIAS\rvert$	STD	RMSE	$\lvert BIAS\rvert$	STD	RMSE
PC	0.032	0.026	0.042	0.105	0.092	0.143
`lme()` for Model (1)	0.128	0.443	0.467	0.373	0.298	0.479
`lme()` for Model (2)	0.128	0.443	0.467	0.208	2.194	2.204

NOTE: The middle three columns are the absolute value of the bias, standard deviation (STD), and root mean squared error (RMSE), averaged over 73 time points, for the fixed-effect estimates. The right three columns are the absolute value of the bias, STD, RMSE, averaged over 73 time points and 34 random effects, for the random-effect estimates.

The parameter cascading method also gives good estimation for LME smoothing models in a simulation study with different scales of measurement errors. These results are discussed in Section 2 of the supplementary file.

6. CONCLUSIONS AND DISCUSSION

Our primary objective was to show that parameter cascading is an acceptable alternative to marginalization as a method for dealing with high-dimensional nuisance parameters. The desiderata that we had in mind were: (1) the quality of estimates of parameter values and their sampling variances; (2) the ease of mathematical development and programming; (3) computational overhead; and (4) the stability of the algorithms involved. We chose the linear mixed-effects model because marginalization-based methods are already widely used and well developed, and because the model itself is widely applied.

Our simulation results indicate that, in this context, parameter cascading is as good as the main current methods in terms of bias and sampling variance for estimation of fixed effects. Table 1 indicates that parameter cascading estimates, along with those produced by function `lme()`, had rather better standard errors than the closed form UNC-MLE and UNC-REML estimates available for balanced designs, and also avoided the frequent occurrence of negative eigenvalues in estimated covariance matrices. Sampling variance estimates produced by parameter cascading showed negligible bias in Table 2, and were as good as or better than those produced by the other three methods. In the random design simulation results, which are included in the supplementary file, we saw that parameter cascade estimates of fixed effects and variance components were of good to excellent quality over a wide range of error variances.

Computational overhead for parameter cascading can be a serious issue for the four-level cascades that we considered in the smoothing situation, since matrices of the order of nN must be multiplied, but when these computations are programmed in Matlab and other languages capable of using multiple processors, computation times for single analyses of the size of the Vancouver weather data ($nN = 2482$) are a matter of a few minutes. The three-level computations described in this paper required fractions of a second.

Along the way we also applied LME to smoothing problems, and developed some new methodology that permits estimation of low-dimensional variance components in the context of both fixed and random effects of rather higher dimensionality than one encounters in usual variance components, random effects, and longitudinal data analyses. These results should prove useful for multicurve smoothing problems where the signal-to-noise ratio is so low that borrowing strength across records offers substantial improvements in both fixed effect and variance component estimation.

In practice, the noise variance $\mathbf{\Sigma}_i$ in (1) may have unknown elements, parameterized with a small number of coefficients, such as in an AR(1) model. These coefficients can be estimated along with the parameter vector $\boldsymbol{\gamma}$ in the top-level optimization. We will work on this problem in our further research, as well as applying parameter cascading to nonlinear mixed-effect models where marginalization results cannot not be expressed in closed form.

Multilevel parameter structures abound in contemporary data analysis, and are on the sharp increase as more and more complex data structures require our attention. The results in this paper encourage us to consider parameter cascading as an effective strategy in contexts much more nonlinear and complicated than this LME context.

SUPPLEMENTAL MATERIALS

Brief Description: Section 1 gives two simulation studies for estimating the LME variance component model. One simulation study is summarized in Section 2 for estimating the LME smoothing model. Section 3 shows one application of parameter cascading method in estimating a LME smoothing model. (Supplement.pdf)

[Received February 2009. Revised October 2009.]

REFERENCES

Bates, D. M., and Watts, D. B. (1988), *Nonlinear Regression Analysis and Its Applications*, New York: Wiley. [365,367]

Brumback, B. A., and Rice, J. A. (1998), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves," *Journal of the American Statistical Association*, 93, 961–976. [365,369]

Cao, J., and Ramsay, J. O. (2007), "Parameter Cascades and Profiling in Functional Data Analysis," *Computational Statistics*, 22 (3), 335–351. [365]

——— (2009), "Generalized Profiling Estimation for Global and Adaptive Penalized Spline Smoothing," *Computational Statistics and Data Analysis*, 53, 2550–2562. [365]

Demidenko, E. (2004), *Mixed Models: Theory and Applications*, New York: Wiley-Interscience. [368]

Demmler, A., and Reinsch, C. (1975), "Oscillation Matrices With Spline Smoothing," *Numerische Mathematik*, 24 (5), 375–382. [369]

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall. [367]

Gu, C. (2002), *Smoothing Spline ANOVA Models*, New York: Springer. [367, 369]

Guo, W. (2002), "Functional Mixed Effects Models," *Biometrics*, 58, 121–128. [365]

Henderson, C. R. (1973), "Sire Evaluation and Genetic Trends," in *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*, Champaign, IL: American Society of Animal Science–American Dairy Science Association–Poultry Science Association, pp. 10–41. [366,367]

Morris, J. S., and Carroll, R. J. (2006), "Wavelet-Based Functional Mixed Models," *Journal of the Royal Statistical Society, Ser. B*, 68, 179–199. [365]

Neyman, J., and Scott, E. L. (1948), "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32. [365]

Pinheiro, J. C., and Bates, D. M. (2000), *Mixed-Effects Models in S and S-Plus*, New York: Springer. [366]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer. [365]

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007), "Parameter Estimation for Differential Equations: A Generalized Smoothing Approach" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 69, 741–796. [365]

Rice, J., and Wu, C. O. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253–259. [365]

Robinson, G. K. (1991), "That BLUP Is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6, 15–32. [366]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge, U.K.: Cambridge University Press. [365]

Wahba, G. (1985), "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem," *The Annals of Statistics*, 13 (4), 1378–1402. [367]

——— (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics. [369]

Wand, M. P. (2003), "Smoothing and Mixed Models," *Computational Statistics*, 18, 223–249. [365]

Welham, S. J., Cullis, B. R., Kenward, M. G., and Thompson, R. (2006), "The Analysis of Longitudinal Data Using Mixed Model l-Splines," *Biometrics*, 62, 392–401. [365]