IMS Lecture Notes–Monograph Series
Optimality: The Third Erich L. Lehmann Symposium
Vol. 57 (2009) 216–236

DOI: 10.1214/09-LNMS5713

Mixed Models, Posterior Means and Penalized Least-Squares

Yolanda Muñoz Maldonado[1]

Michigan Technological University

Abstract: This paper reviews the connections between estimators that derive from three different modeling methodologies: Mixed-effects models, Bayesian models and Penalized Least-squares. Extension of classical results on the equivalence for smoothing spline estimators and best linear unbiased prediction and/or posterior analysis of certain Gaussian signal-plus-noise models is examined in a more general setting. These connections allow for the application of an efficient, linear time algorithm, to estimate parameters, compute random effects predictions and evaluate likelihoods in a large class of model scenarios. We also show that the methods of generalized cross-validation, restricted maximum likelihood and unbiased risk prediction can be used to estimate the variance components or adaptively select the smoothing parameters in any of the three settings.

Contents

1. Introduction

Mixed-effects model methodology, penalized least-squares and Bayesian random-effects models are widely used statistical tools. However, due to the dissimilar nature of the settings in which they are typically formulated, connections between these three techniques as well as the fundamental reasons for the connections, have often been overlooked. In this paper, we review some of the well known results that connect smoothing spline estimators, Gaussian signal-plus-noise models and best linear unbiased prediction of mixed-effects models and show that they are but one aspect of a general framework that allows for "cross-platform" development in mixed-effects models, using frequentist or Bayesian approaches, and/or penalized least-squares (PLS) criteria.

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI USA, e-mail: ymunoz@mtu.edu

AMS 2000 subject classifications: Primary 62J05; secondary 65C20.

Keywords and phrases: smoothing splines, p-splines, ridge regression, varying coefficient models, Bayesian prediction, Kalman fiter, adaptive selection.

The relationship between particular cases of frequentist and Bayesian mixed-effects models and PLS has been exposed before. For example, Lindley and Smith [29] proposed the use of prior information on the parameters of a fixed effects linear model under the assumption of the parameters having exchangeable distributions. In the early development of the Bayesian theory for smoothing splines, Wahba [43] noticed the intimate connection between estimators resulting from spline smoothing and a Gaussian model with diffuse initial conditions. Robinson [35] remarked on applications of Best Linear Unbiased Predictors (BLUP's) for estimation of variance parameters, randomized block designs and their link to empirical Bayes methods and Kriging. Speed [40] pointed out, in a comment to Robinson's article, that smoothing spline estimators were in fact BLUP's of a certain mixed effects model. In the PLS framework, it is well known that smoothing splines estimators are a special case of penalized splines estimators (P-splines) [37]. Wahba [47] and Cressie [7] discussed the links between splines and kriging estimates and Nychka used the representation of smoothing splines as a type of ridge estimator to further relate smoothing spline estimation and kriging [32].

More recently, researchers have been using the connection between smoothing spline estimators and particular mixed-effects models to compute smoothing spline estimators [see 4, 19, 48]. Ruppert et al. [36] mentioned the correspondence between penalized spline smoothers and prediction in the mixed-effects model and remarked on the advantages of using existing mixed-effects model techniques and software in a semi-parametric regression setting. Eubank et al. [11] took advantage of the relationship between smoothing splines and the Gaussian model of [43] to provide a general development that includes the efficient computation of estimators in a varying coefficient model context.

Using connections that have been established for various special cases, we synthesize them and present a formal result that details precisely when penalized least-squares estimation, BLUP for a mixed-effects model and posterior mean analysis of a mixed-effects model with diffuse priors on some of the random effects (hereafter referred to as simply the Bayesian model) produce identical estimators. We then describe how this can be exploited in many cases of interest to provide a computationally efficient algorithm for evaluation of estimators and likelihoods, computation of predictions, and construction of Bayesian prediction intervals. The implemented algorithm reduces the computational effort of calculating the aforementioned quantities by two orders of magnitude over what would normally be the case for a direct mixed-effects model approach. We also establish a result showing that the methods of Generalized Cross-validation (GCV), Restricted Maximum Likelihood (REML) or the equivalent technique of Generalized Maximum Likelihood (GML) and Unbiased Risk Prediction (UBR) can be used in any of the three settings to adaptively estimate the smoothing parameters or variance components.

The following three examples will be used throughout the paper to illustrate the utility of our approach.

Example 1 (Varying Coefficient Models). Varying coefficient models generalize ordinary linear regression models by allowing for regression coefficients that change dynamically as a function of independent variables. The simplest example of this are the so-called time varying coefficient models where there is only one effect modifying covariate. In that setting, we have response variables y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, that depend on some predictor variables $x_{1ij}, \ldots, x_{Kij}$ through a

relationship of the form

$$y_{ij} = \sum_{k=1}^{K} \beta_k(t_{ij}) x_{kij} + e_{ij}, \tag{1}$$

where the $\beta_k(\cdot)$'s are unknown coefficient functions of a covariate t and the e_{ij} represent random error terms. Models like (1) were first introduced by [21] who proposed obtaining estimators through minimization of the PLS criterion

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ y_{ij} - \sum_{k=1}^{K} f_k(t_{ij}) x_{kij}(t_{ij}) \right\}^2 + \sum_{k=1}^{K} \lambda_k \int_0^1 [f_k^{(r)}(t)]^2 dt \tag{2}$$

over functions $f_1, \ldots, f_K$ having r square integrable derivatives, and $g^{(s)}(t)$ being the s^{th}derivative of the function g. The parameters $\lambda_k \geq 0$ control the smoothness of the coefficient functions and the minimizers can be shown to be natural splines of degree $2r-1$ with knots at the unique elements of the set $\{t_{ij}\}$.

Example 2 (Ridge Regression). Consider the linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of responses, $\mathbf{X}$ is a known $n \times p$ matrix of predictor variables of rank p, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients and $\boldsymbol{e}$ is a normally distributed vector of errors with $\mathrm{E}(\boldsymbol{e}) = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}) = \sigma_e^2 \mathbf{I}$, with "T" denoting the transpose of a matrix and $\mathbf{I}$ an identity matrix of suitable dimension. The generalized ridge regression estimator of $\boldsymbol{\beta}$ is then given by $\hat{\boldsymbol{\beta}} = [\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}]^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$. This estimator can be obtained by minimizing the PLS criterion

$$(\boldsymbol{y} - \mathbf{X}\boldsymbol{a})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\boldsymbol{a}) + \boldsymbol{a}^{\mathrm{T}}\mathbf{K}\boldsymbol{a}, \tag{4}$$

over $\{\boldsymbol{a} : \boldsymbol{a} \in I\!R^p\}$, with $\mathbf{K}$ a diagonal matrix having elements $\lambda_i \geq 0$, for $i = 1, \ldots, p$. A special instance of (4) is given by ordinary ridge regression in which case the predictor variables are usually standardized and $\mathbf{K}$ has the form $\lambda\mathbf{I}$, for $\lambda > 0$. Other variations of generalized ridge regression are the P-splines estimators of [9] and of [36]. We will now describe the later approach in more detail.

Suppose that we have a collection of points on the plane, (t_i, y_i), $i = 1, \ldots, n$, and want to fit them using scatter-plot smoothing methodology. P-splines provide one popular approach for accomplishing this that arise from using a spline basis to construct the $\mathbf{X}$ matrix in (3). That is, for some integer $m \geq 0$ and a fixed set of knots $\xi_1 < \xi_2 < \cdots < \xi_p$, we take $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m+p}]$ with $\boldsymbol{x}_1$ a n-vector of all ones, $\boldsymbol{x}_j = [t_1^{j-1}, \ldots, t_n^{j-1}]^{\mathrm{T}}$, $j = 2, \ldots, m$ and $\boldsymbol{x}_{m+j} = [(t_1 - \xi_j)_+^{m-1}, \ldots, (t_n - \xi_j)_+^{m-1}]^{\mathrm{T}}$, for $j = 1, \ldots, p$ with $(x)_+^r$ being x^r for $x \geq 0$ and zero otherwise. A P-spline smoother is then found by minimizing (4), with the matrix $\mathbf{K}$ having the form

$$\mathbf{K} = \begin{bmatrix} \mathbf{0}_{m\times m} & \mathbf{0}_{m\times p} \\ \mathbf{0}_{p\times m} & \lambda\mathbf{I} \end{bmatrix}, \tag{5}$$

with $\mathbf{0}_{r\times s}$ being an r by s matrix of all zeros.

Example 3 (Randomized Block Design). Linear mixed-effects models have been applied for analysis of data arising from situations involving repeated measures and experimental designs with factors that can be seen as a combination of

fixed and random effects. Some types of randomized block designs fall in the last category, for example, when the experimental units are randomly selected and each has repeated measurements. For this particular type of design, the experimental units are assumed to be the factor (or blocking criterion), that makes them relatively homogeneous with respect to a measured response. One way of modeling this type of problems is

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{b} + \boldsymbol{e}, \tag{6}$$

where $\mathbf{X}$ is the design matrix for the fixed-effects, $\boldsymbol{\theta}$ is the parameter vector for the fixed-effects and $\boldsymbol{b}$ is a random vector of blocking factors. This is not the only model that can be used with this type of design, but it will serve the purpose of this paper.

The remainder of the paper is organized as follows. In Section 2 we present a result that connects estimators/predictions that are obtained from mixed-effects models, penalized least-squares estimation and Bayesian formulations. We also address the issue of estimation of the variance components and smoothing parameters that arise from their respective contexts. In this latter respect, we establish that GCV, REML/GML and UBR can all be used to obtain the above mentioned estimators. Section 3 illustrates the implementation of our main result using the three examples mentioned in this section. Section 4 concludes with some comments about the use of the theorems in Section 2 and the employment of the Kalman filter algorithm.

2. Equivalence Theorem

To begin, we will give a detailed description of the three modeling scenarios that are the focus of this section.

- **Mixed-effects model:** Consider first the linear mixed-effects model

$$\boldsymbol{y} = \mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b} + \boldsymbol{e}, \tag{7}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of responses and $\mathbf{T}$ and $\mathbf{U}$ are design matrices for the fixed and random effects of dimensions $n \times m$ and $n \times q$, respectively. Here, we take $\boldsymbol{\theta}$ to be a $m \times 1$ vector of fixed effects and $\boldsymbol{b}$ to be a $q \times 1$ normally distributed random vector with zero mean and variance-covariance matrix $\mathrm{Var}(\boldsymbol{b}) = \sigma_b^2\mathbf{R}$. The random effects $\boldsymbol{b}$ are assumed to be independent of the $n \times 1$ vector of random errors, $\boldsymbol{e}$, which in turn, is assumed to be normally distributed with zero mean and variance-covariance matrix $\sigma_e^2\mathbf{I}$. For this model, as well as for the Bayesian model below, the parameters σ_e^2 and σ_b^2 are the so called variance components. It is often convenient to reparameterized the variance components as $\lambda = \sigma_b^2/\sigma_e^2$ so that $\mathrm{Var}(\boldsymbol{y}) = \sigma_e^2(\lambda\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \mathbf{I})$. The value of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ can be predicted using its BLUP.

- **Bayesian Model:** Similar to the previous case, in this setting we assume that

$$\boldsymbol{y} = \mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b} + \boldsymbol{e}, \tag{8}$$

with $\mathbf{T}$ and $\mathbf{U}$ fixed matrices. However, we now also take $\boldsymbol{\theta}$ to be random and model it as being independent of $\boldsymbol{b}$ and $\boldsymbol{e}$, with a zero mean, normal prior distribution having variance-covariance matrix $\mathrm{Var}(\boldsymbol{\theta}) = \nu\mathbf{I}$. The vector of random effects, $\boldsymbol{b}$, is also assumed to be normally distributed with zero mean and $\mathrm{Var}(\boldsymbol{b}) = \sigma_b^2\mathbf{R}$. Prediction of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ can be accomplished via the use of its posterior mean. In

the absence of an informative prior for $\boldsymbol{\theta}$ a diffuse formulation can be employed wherein ν is allowed to diverge. *Note: notice that this is not truly a Bayesian model since there are no priors on the variance components. It is named Bayesian model for the sake of identification.*

- **Penalized Least-Squares:** In this case we have $\boldsymbol{y} = \mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b} + \boldsymbol{e}$ with $\boldsymbol{\theta}$ and $\boldsymbol{b}$ being non random and $\boldsymbol{e}$ is a vector of zero mean, normally distributed random errors with variance-covariance matrix $\mathrm{Var}(\boldsymbol{e}) = \sigma_e^2\mathbf{I}$. The parameters are to be estimated by minimizing the PLS criterion

$$\mathrm{PLS}(\boldsymbol{a}, \boldsymbol{c}) = (\boldsymbol{y} - \mathbf{T}\boldsymbol{a} - \mathbf{U}\boldsymbol{c})^T(\boldsymbol{y} - \mathbf{T}\boldsymbol{a} - \mathbf{U}\boldsymbol{c}) + \lambda \boldsymbol{c}^T\mathbf{R}^{-1}\boldsymbol{c}, \tag{9}$$

with respect to $\boldsymbol{a}$ and $\boldsymbol{c}$. Here, $\mathbf{R}^{-1}$ is a penalty matrix and λ is the parameter that controls how heavily we penalize the coefficients $\boldsymbol{c}$.

Having these three scenarios in mind, we now state the following theorem.

Theorem 2.1. *The Best Linear Unbiased Predictor (BLUP) of* $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ *in (7) is given explicitly by*

$$\hat{\boldsymbol{y}} = \mathbf{A}_\lambda \boldsymbol{y}, \tag{10}$$

where

$$\mathbf{A}_\lambda = \{\mathbf{I} - \mathbf{Q}^{-1}[\mathbf{I} - \mathbf{T}(\mathbf{T}^T\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{Q}^{-1}]\}, \tag{11}$$

and

$$\mathbf{Q} = (\lambda \mathbf{U}\mathbf{R}\mathbf{U}^T + \mathbf{I}). \tag{12}$$

This result is numerically the same as the limiting value (as $\nu \to \infty$*) of* $E[\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}]$ *in (8) and the minimizer of (9).*

Proof. To simplify the proof let us assume that the design matrices $\mathbf{U}$ and $\mathbf{T}$, as well as $\mathbf{R}$, are all full rank matrices (we will later relax this assumption).

Under model (7), the first two moments of $\boldsymbol{y}$ are given by

$$\mathrm{E}(\boldsymbol{y}) = \mathbf{T}\boldsymbol{\theta} \ \text{ and } \ \mathrm{Var}(\boldsymbol{y}) = \sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \sigma_e^2\mathbf{I}.$$

Using the distribution of $\boldsymbol{y}$ given $\boldsymbol{b}$ and the distribution of $\boldsymbol{b}$, we can then find the joint density of $\boldsymbol{y}$ and $\boldsymbol{b}$ and obtain the normal equations of [23]:

$$\begin{aligned} \mathbf{T}^{\mathrm{T}}\mathbf{T}\boldsymbol{\theta} + \mathbf{T}^{\mathrm{T}}\mathbf{U}\boldsymbol{b} &= \mathbf{T}^{\mathrm{T}}\boldsymbol{y}, \\ \mathbf{U}^{\mathrm{T}}\mathbf{T}\boldsymbol{\theta} + (\mathbf{U}^{\mathrm{T}}\mathbf{U} + \mathbf{R}_\lambda^{-1})\boldsymbol{b} &= \mathbf{U}^{\mathrm{T}}\boldsymbol{y} \end{aligned}$$

for $\mathbf{R}_\lambda = \lambda\mathbf{R}$.

After some algebra and using the Sherman-Morrison-Woodbury formula in [24] we have

$$\mathbf{Q}^{-1} = \mathbf{I} - \mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{U} + \mathbf{R}_\lambda)^{-1}\mathbf{U}^{\mathrm{T}}, \tag{13}$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\boldsymbol{y}$$

and

$$\hat{\boldsymbol{b}} = (\mathbf{U}^{\mathrm{T}}\mathbf{U} + \mathbf{R}_\lambda^{-1})^{-1}\mathbf{U}^{\mathrm{T}}[\mathbf{I} - \mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}]\boldsymbol{y}.$$

In this way, the predicted values of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ are given by

$$\hat{\boldsymbol{y}} = \{\mathbf{I} - \mathbf{Q}^{-1}[\mathbf{I} - \mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}]\}\boldsymbol{y}. \tag{14}$$

To show that minimization of the PLS criterion produces the same numerical answer as the BLUP of (7), we differentiate PLS$(\boldsymbol{a}, \boldsymbol{c})$ with respect to $\boldsymbol{a}$ and $\boldsymbol{c}$ to obtain normal equations which together with (13) give us the same answer as in (14).

It remains to show that under the Bayesian model with diffuse prior, $\lim_{\eta\to\infty}$ E$(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y})$ also agrees with (14). In this case, the joint distribution of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ and $\boldsymbol{y}$ is found to be normal with zero mean vector and variance-covariance matrix given by

$$\begin{pmatrix} \nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} & \nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} \\ (\nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}})^{\mathrm{T}} & \nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \sigma_e^2\mathbf{I} \end{pmatrix}.$$

Standard multivariate analysis results then produce

$$\begin{aligned} \mathrm{E}(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}) &= \mathrm{Cov}(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}, \boldsymbol{y})[\mathrm{Var}(\boldsymbol{y})]^{-1}\boldsymbol{y} \\ &= (\nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}}) \\ &\quad \times(\nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \sigma_e^2\mathbf{I})^{-1}\boldsymbol{y}. \end{aligned}$$

Letting λ be as in (12), $\eta = \nu/\sigma_e^2$ and recalling equation (12) we obtain

$$\mathrm{E}(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}) = (\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{U}\mathbf{R}_\lambda\mathbf{U}^{\mathrm{T}})(\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{Q})^{-1}\boldsymbol{y}. \tag{15}$$

Applying the Sherman-Morrison-Woodbury formula [24] on $(\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{Q})^{-1}$ and using a little algebra we get

$$(\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{Q})^{-1} = \mathbf{Q}^{-1} - Q^{-1}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}[\eta^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1} + \mathbf{I}]^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}.$$

For η sufficiently large, the eigenvalues of $(\eta^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}$ are all less than one. So, applying a power series expansion on $(\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{Q})^{-1}$ [16], substituting this expansion in (15), and with the aid of some straight forward calculus we have that $\lim_{\eta\to\infty}$E$(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y})$ is exactly the same expression as in (14).

Now, let us go back to our assumption of $\mathbf{U}$, $\mathbf{T}$ and $\mathbf{R}$ being full rank matrices. This may not be always the case. For example, if we approach estimation from the PLS criterion perspective, there are cases (such as spline smoothing), where $\mathbf{R}$ has less than full rank. To deal with this instance, suppose that the matrix $\mathbf{U}\mathbf{R}\mathbf{U}^T$ is not invertible. In this situation, the matrix $\mathbf{Q} = (\lambda\mathbf{U}\mathbf{R}\mathbf{U}^T + \mathbf{I})$ will still be invertible and our only concern is that the matrix $\mathbf{T}$ is less than full rank. In that case, we can employ conditional inverses (e.g., [18], pp. 31) and the theorem will still hold. □

A result such as Theorem 2.1 is important because, as pointed out by [4, 19, 48] and [36], one can take advantage of existing methodology and software to facilitate and enhance our analyses. The difference here is that Theorem 2.1 is not restricted to the smoothing spline case of [43]; the BLUP result by [40] and referenced by [4, 19, 48]; or to the Bayesian mixed model of [29]. Instead we see that, quite generally, methodology from any particular one of the three frameworks can be potentially applied to obtain useful developments for the other two.

In each of the scenarios described by Theorem 2.1, it will generally be necessary to estimate the parameter λ. The following result is a generalization of Theorem 5.6 in [12] that allows us to apply three standard methods to the problem of adaptively

selecting this parameter. The methods considered here are GCV, UBR and GML which respectively produce estimators of λ via minimization of

$$\text{GCV}(\lambda) = \frac{n^{-1}\text{RSS}(\lambda)}{[n^{-1}\text{tr}(\mathbf{I} - \mathbf{A}_\lambda)]^2}, \tag{16}$$

$$\text{UBR}(\lambda) = n^{-1}\text{RSS}(\lambda) + 2n^{-1}\sigma_e^2\text{tr}(\mathbf{A}_\lambda), \tag{17}$$

and

$$\text{GML}(\lambda) = \frac{\boldsymbol{y}^T(\mathbf{I} - \mathbf{A}_\lambda)\boldsymbol{y}}{|\mathbf{I} - \mathbf{A}_\lambda|_+^{1/(n-m)}}. \tag{18}$$

Here, tr denotes the trace of a matrix, $\text{RSS}(\lambda) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}})$and $|\mathbf{I} - \mathbf{A}_\lambda|_+$ is the product of the nonzero eigenvalues of $\mathbf{I} - \mathbf{A}_\lambda$.

We note in passing that GML is equivalent to the method of REML that is a popular approach to variance component estimation. (See, e.g. [40].) In terms of the relationship between criteria (16)-(18) we can establish the following result.

Theorem 2.2. E[GCV(λ)], E[UBR(λ)] *and* E[REML/GML(λ)] *are all minimized at* $\lambda = \sigma_b^2/\sigma_e^2$.

Proof. To establish Theorem 2.2 first note that the arguments in [12, pp. 244–247] can be easily modified to account for either the GCV or UBR part of the theorem. The main difference is that here we are not working with diffuse priors. Thus, we will concentrate on sketching the part of the proof that pertains to equation (18).

Let $\lambda_o = \sigma_b^2/\sigma_e^2$ and write $\mathbf{I} - \mathbf{A}_\lambda = \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{Q}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$, for a $\mathbf{B}$ such that $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}$, $\mathbf{B}\mathbf{B}^{\mathbf{T}} = \mathbf{I} - \mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{T} = 0$. Then,

$$\begin{aligned}\mathbf{B}^{\mathrm{T}}\mathbf{Q}\mathbf{B} &= \mathbf{B}^{\mathrm{T}}(n\lambda\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \mathbf{I})\mathbf{B} \\ &= n\lambda\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}}\mathbf{B} + \mathbf{I}.\end{aligned}$$

Define the matrix of eigenvalues for $\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}}\mathbf{B}$ with corresponding matrix of eigenvectors $\mathbf{V}$ as $\mathbf{\Lambda} = \text{diag}\{d_1, \ldots, d_{n-m}\}$. Then, we can write

$$\mathbf{B}^{\mathrm{T}}\mathbf{Q}\mathbf{B} = \mathbf{V}(\lambda\mathbf{\Lambda} + \mathbf{I})\mathbf{V}^{\mathrm{T}}.$$

Now, taking expectation with respect to $\boldsymbol{e}$ and $\boldsymbol{b}$ we can show that

$$\begin{aligned}\text{E[REML/GML}(\lambda)] &= \frac{\sigma_e^2\text{tr}[(\mathbf{I} - \mathbf{A}_\lambda)] + \lambda_o\text{tr}[(\mathbf{I} - \mathbf{A}_\lambda)(\mathbf{Q} - \mathbf{I})]}{[\prod_{i=1}^{n-m}(\lambda d_i + 1)^{-1/(n-m)}]} \\ &= \frac{\sigma_e^2}{\prod_{i=1}^{n-m}(\lambda d_i + 1)^{-1/(n-m)}} \sum_{i=1}^{n-m} \frac{(\lambda_o d_i + 1)}{(\lambda d_i + 1)}.\end{aligned}$$

Now, take the difference of the logarithms of the expectations E[REML/GML(λ)] and E[REML/GML(λ_o)]. A sufficient condition for minimization of the REML/ GML criterion at λ_o is then seen to be

$$\log\left[\frac{1}{n-m}\sum_{i=1}^{n-m}\frac{(\lambda_o d_i + 1)}{(\lambda d_i + 1)}\right] - \frac{1}{(n-m)}\sum_{i=1}^{n-m}\log\left[\frac{(\lambda_o d_i + 1)}{\lambda d_i + 1}\right] \geq 0.$$

However, this is an immediate consequence of Jensen's inequality. □

Criteria (16)–(18) have long been used for the selection of smoothing or penalty parameters. Golub et al. [17] proposed (16) as a method to choose the ridge regression parameter in a standard regression model like (3) and Craven and Wahba [6] introduced GCV as a method for choosing the smoothing parameter in nonparametric regression. Wahba [46], Kohn et al. [27] and Stein [41] compared the performance of GCV and REML/GML for the smoothing spline case.

Unlike the methods of REML/GML in the PLS framework, GCV and UBR have not been applied in the context of mixed-effects models. Theorem 2.2 suggests that GCV may be another suitable method for estimation of variance components in this context. The fact that the GCV estimator of the variance components shares the REML/GML estimator attribute of minimizing the expectation of the risk, seems to indicate that both estimators will have similar properties and behavior under the mixed-effects model (as it has been shown for the PLS and the Bayesian models [see 27, 46]). However, this needs to be confirmed by studying the distributional and consistency properties of the GCV estimator of σ_e^2 and σ_b^2 under the mixed-effects model and this is a topic for future research.

3. Examples

In this section we focus on the examples introduced in section 1 and exemplify the advantages of using existing methodology for one particular framework (the Bayesian model) to the other two. In particular, we will use a Kalman filter algorithm to compute estimators and predictions that arise in the three scenarios considered in Theorem 2.1. Perhaps the most common application of the Kalman filter has been in a Bayesian context (see [3, 28]). Specifically, Kohn and Ansley [25], using Wahba's Gaussian model (a particular case of our Bayesian model), reformulated the model into a state-space representation and thereby obtained an efficient $O(n)$ algorithm for computing smoothing spline estimators. Theorem 2.1 allows us to extend this approach to non spline smoothing situations and obtain an efficient, Kalman filter based, computational algorithm provided that the random components in Theorem 2.2 admit a state-space representation. This algorithm also permits the evaluation of likelihood functions, making it possible to obtain REML/GML estimators for variance components or smoothing parameters.

Description of the Kalman filter is beyond the scope of this paper. Instead, we will focus on establishing a state-space representation for the three examples and refer the reader to [11, 13] and [14] for a more complete development. To accomplish this, it suffices to give only a brief discussion concerning the form of a state-space model.

Any response y_i can be represented using a state-space model if the observation at time i can be expressed as a function of the observation at time $i-1$. More formally, a state-space model is composed of a set of response equations

$$y_i = \mathbf{h}^{\mathrm{T}}(t_i)\boldsymbol{x}(t_i) + e_i, \tag{19}$$

and a system of state equations

$$\boldsymbol{x}(t_{i+1}) = \mathbf{F}(t_i)\boldsymbol{x}(t_i) + \boldsymbol{u}(t_i). \tag{20}$$

with $t_i \in [0,1]$ and $0 = t_0 \leq t_1 < \cdots < t_n$. The y_i are observed quantities and the e_i, $\boldsymbol{u}(t_i)$, $\boldsymbol{x}(t_i)$, are all unobservable with $\boldsymbol{u}(t_0), \ldots, \boldsymbol{u}(t_{n-1})$, $e_1, \ldots, e_n$ and the initial state, $\boldsymbol{x}(t_0)$, all being zero mean, uncorrelated normal random variables.

In general, the $\boldsymbol{x}(t_i)$ and $\boldsymbol{u}(t_i)$ may be vector valued with $\boldsymbol{u}(t_i)$ having variance-covariance matrix $\mathbf{R}_{u(t_i)}$. For our purposes we will treat the vectors $\mathbf{h}(t_i)$ and the transition matrix $\mathbf{F}(t_i)$ in (19)–(20) as being known.

We will proceed now to demonstrate the application of the equivalence theorem in the context of our three examples.

3.1. Varying Coefficient Models

To illustrate the varying coefficient case, we will examine the progesterone profiles data (Figure 1) of [4]. The data consists of metabolite progesterone profiles, measured daily in urine over the course of a menstrual cycle in a group of 51 women.

The women in the study were divided into two groups: 29 in the non-conceptive group and 22 in the conceptive group. Each woman contributed a different number of cycles, ranging from 1 to 5 cycles and some of the cycles have missing values.

The goal of the analysis is to detect differences between the conceptive and non-conceptive group profiles. To do this we will express the varying coefficient model (1) with the formulation in (9), apply Theorem 2.1 and find the equivalent formulation (8) in the Bayesian framework in order to use the efficient Kalman filter algorithm of [13].

For simplicity, assume that we have complete data and the same number of cycles per woman (later we will relax these assumptions). Let the log progesterone level of the c^{th} cycle for the w^{th} woman at time t_i be denoted by y_{wci} and model this

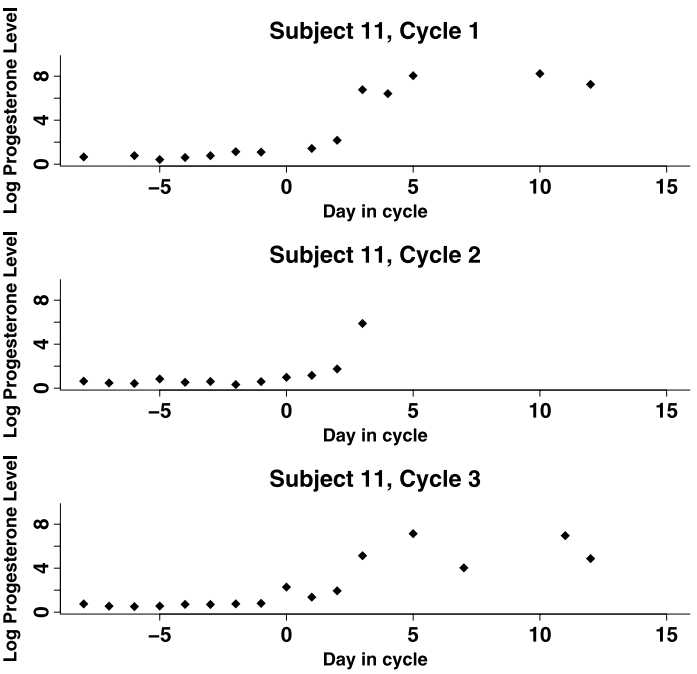

FIG 1. *Observed progesterone measurements for subject 11 in the non-conceptive group. The plots correspond to three of the four cycles for subject 11 and show the log progesterone concentration versus day in the cycle. All cycles have missing observations. Days corresponding to the menses were excluded.*

response as

$$y_{wci} = \beta_1(t_i)X_{1wci} + \beta_2(t_i)X_{2wci} + e_{wci},$$

where $i = 1, \ldots, 24$, and $t_1 = -8, t_2 = -7, \ldots, t_{24} = 15$ are the days in a menstrual cycle. The cycles c range from 1 to 5 and $w = 1, \ldots, 29$ correspond to women in the non-conceptive group and the rest belong to the conceptive group.

Assume that the $\beta_k(\cdot)$'s, $k = 1, 2$, are smooth functions of t. Usually, this translates into assuming that the functions belong to a Hilbert space of order m (see [22]). To find the estimated profiles we minimize a particular PLS criterion, where the penalty is applied to the integral of the square of the second derivative of the $\beta_k(\cdot)$'s. The minimizers of $\beta_1(\cdot)$, $\beta_2(\cdot)$ are natural splines of order m, with $m = 3$, that can be represented by a linear combinations of basis functions

$$\sum_{q=0}^{m-1} \theta_{kq} t_i^q + \sum_{r=1}^{24} b_{kr}\xi_r(t_i),$$

with knots at each of the design points t_i, and

$$\xi_r(t_i) = \int_0^{\min\{t_r, t_i\}} \frac{(t_i - u)^{m-1}(t_r - u)^{m-1}du}{[(m-1)!]^2}. \tag{21}$$

Equation (21) is one of the usual reproducing kernels of a Hilbert space of order m [22].

Let $\boldsymbol{y}_{wc} = [y_{wc}(t_1), \ldots, y_{wc}(t_{24})]^{\mathrm{T}}$ be the vector of responses for women w that contains all the daily observations in the c cycle, and $\boldsymbol{\xi}_i = [\xi_1(t_i), \xi_2(t_i), \ldots, \xi_{24}(t_i)]^{\mathrm{T}}$ and $\boldsymbol{t}_i = [t_i^0, t_i^1, \ldots, t_i^{m-1}]^{\mathrm{T}}$ be the vectors of basis functions evaluated at the times t_i's.

Denote the vector of coefficients for β_1 and β_2 as $\boldsymbol{\theta}_1 = [\theta_{10}, \theta_{11}, \ldots, \theta_{1,(m-1)}]^{\mathrm{T}}$, $\boldsymbol{\theta}_2 = [\theta_{20}, \theta_{21}, \ldots, \theta_{2,(m-1)}]^{\mathrm{T}}$, $\boldsymbol{b}_1 = [b_{10}, b_{11}, \ldots, b_{1,24}]^{\mathrm{T}}$, $\boldsymbol{b}_2 = [b_{21}, b_{22}, \ldots, b_{2,24}]^{\mathrm{T}}$, respectively. Construct $\boldsymbol{t}$, $\boldsymbol{\xi}$ and $\boldsymbol{X}$ such that

$$\boldsymbol{t} = \begin{bmatrix} \boldsymbol{t}_1^{\mathrm{T}} \\ \boldsymbol{t}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{t}_{24}^{\mathrm{T}} \end{bmatrix}, \ \boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\xi}_1^{\mathrm{T}} \\ \boldsymbol{\xi}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\xi}_{24}^{\mathrm{T}} \end{bmatrix} \text{ and } X = \begin{bmatrix} X_{1wc1} & X_{2wc1} \\ X_{1wc2} & X_{2wc2} \\ \vdots & \vdots \\ X_{1wc24} & X_{2wc24} \end{bmatrix}.$$

Let $\mathbf{T}_{wc} = \boldsymbol{t} \bigotimes \mathbf{X}$ and $\mathbf{U}_{wc} = \boldsymbol{\xi} \bigotimes \mathbf{X}$, where $\boldsymbol{A} \bigotimes \mathbf{B}$ denotes the Kronecker product of the matrices $\boldsymbol{A}$ and $\mathbf{B}$ and it is equal to $a_{ij}\mathbf{B}$.

For each woman's cycle we have the model $\mathbf{T}_{wc}\boldsymbol{\theta}^\star + \mathbf{U}_{wc}\boldsymbol{b}^\star + \boldsymbol{e}_{ew}$, where $\boldsymbol{\theta}^\star = [\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{b}^\star = [\boldsymbol{b}_1^{\mathrm{T}}, \boldsymbol{b}_2^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{e}_{wc}$ is the corresponding vector of errors. Denote by $\boldsymbol{y}$ and $\boldsymbol{e}$ the vectors resulting from stacking the vectors $\boldsymbol{y}_{wc}$ and $\boldsymbol{e}_{wc}$, (i.e., $\boldsymbol{y} = [\boldsymbol{y}_{1,1}^{\mathrm{T}}, \boldsymbol{y}_{1,2}^{\mathrm{T}}, \ldots, \boldsymbol{y}_{1,5}^{\mathrm{T}}, \boldsymbol{y}_{2,1}^{\mathrm{T}}, \ldots, \boldsymbol{y}_{51,5}^{\mathrm{T}}]^{\mathrm{T}}$), and let $\mathbf{T} = \mathrm{diag}\{\mathbf{T}_{wc}\}_{\substack{w=1,51 \\ c=1,5}}$ and $\mathbf{U} = \mathrm{diag}\{\mathbf{U}_{wc}\}_{\substack{w=1,51 \\ c=1,5}}$. Then, we can construct the model $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b} + \boldsymbol{e}$, where $\boldsymbol{\theta} = \mathbf{1} \bigotimes \boldsymbol{\theta}^\star$, $\boldsymbol{b} = \mathbf{1} \bigotimes \boldsymbol{b}^\star$, and minimize criteria (9), where $\mathbf{R}^{-1} = \mathbf{U}$.

By Theorem 2.1, this is equivalent to find $\lim_{\nu \to \infty} \mathrm{E}[\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b} | \boldsymbol{y}]$, where $\boldsymbol{\theta}$, $\boldsymbol{b}$ and $\boldsymbol{e}$ are independent of each other and normally distributed with zero mean and variance-covariance matrices $\nu\mathbf{I}$, $\sigma_b^2\mathbf{U}^{-1}$, and $\sigma_e^2\mathbf{I}$, respectively. In this case, the smoothing parameter λ in the PLS model can be found using the variance components, σ_e^2 and σ_b^2 since $n\lambda = \sigma_b^2/\sigma_e^2$, where n is the total number of observations in the data.

TABLE 1
Run time comparisons between the Kalman filter algorithm of [11] implemented in SAS, SAS proc mixed and Brumback and Rice's [4] approach. Both Kalman filter and Brumback and Rice's approach include the time it took to calculate the smoothing parameter. The proc mixed time does not include this computation

Method	Real Time
Kalman Filter	**14.60 secs.**
PROC MIXED	4 hrs. 12 mins. 15 secs.
Brumback and Rice	1 hr. 50 mins.

The equivalent Bayesian representation of the varying coefficient model will allow us to make use of the Bayesian theory and apply it to our PLS setting. Specifically, we can follow [13] and transform the Bayesian model into a state-space model, as they indicate, and apply their efficient algorithm to compute the varying coefficients and respective confidence bands. Their approach also shows how to reformulate the matrices in the Bayesian model so the unbalanced design does not represent a problem in the computation of the estimators. For details on how to find the state-space model form, or on how to apply this efficient algorithm, we refer the readers to the appendix and to the above mentioned authors, respectively.

To see what are the advantages of using this equivalence representation of the PLS, let us first explore the extent that the Kalman filter can speed up computations. To investigate this issue we carried out a run time comparison between our Kalman filter approach, the "standard way" of estimation assuming a mixed-effects model approach (both in SAS), and, only as a reference, we provide the time used in the method developed by Brumback and Rice [4]. We need to point out that these are the reported times in their 1998 paper and that there has been great improvement in computational speed since the publication of this paper. Table 1 shows the required times for computing the estimated conceptive and non-conceptive functions (see Figure 2).

The first time in Table 1 corresponds to the time employed by the Kalman filter algorithm of [11] implemented in SAS and using a computer with a 3.2GHz processor and 1G RAM. This algorithm used 2004 observations (missing values were omitted) and calculated the estimated coefficient functions and corresponding 95% confidence intervals. The second time is the result of using a mixed-effects model representation and taking advantage of SAS proc mixed (the same equipment was used). The last time is the one reported by [4]. They implemented an eigenvalue-eigenvector decomposition on a mixed-effects model representation of the profiles, separately for each group, and combined the times for both groups and the estimation of the variance components. We calculated the smoothing parameters via REML/GML using the Kalman filter algorithm and it took approximately 10.5 seconds in SAS (this time is included in the computation of the Kalman filter in Table 1). These parameters were used in both the SAS and Kalman filter calculation of the varying coefficient functions (we didn't want to calculate the smoothing parameters with SAS Proc Mixed given that it already took 4 hrs. to calculate the profiles without estimating the variance components). The reason why SAS takes so long to estimate the functions is due to the complex covariance structure of the model and the number of observations. The convenient SAS built-in covariance structures were not an option (see comment by [4]), and the inversion of a general $n \times n$ matrix requires $O(n^3)$ operations versus the $O(n)$ used by the Kalman filter.

Another advantage of using the Bayesian interpretation in our PLS model is that the Kalman filter allows us to easily obtain confidence intervals as well as

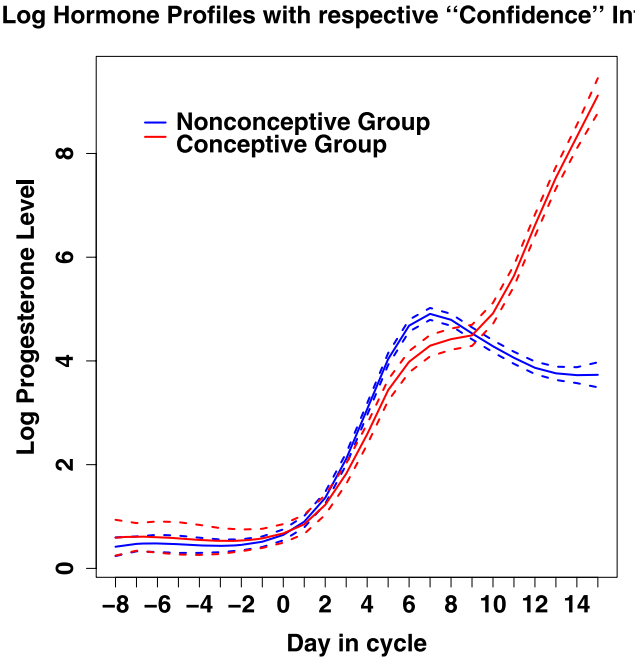

FIG 2. *Smooth estimates for non conceptive and conceptive mean groups with respective 95% pointwise confidence intervals. The corresponding smoothing parameters were computed using the GML method implemented through a Kalman filter algorithm.*

point estimators. In this respect, we use the relationship between PLS and the Bayesian model to provide Bayesian $100(1-\alpha)\%$ confidence (or prediction) intervals which parallel those developed by [44] and [31]. Specifically, we "estimate" the i^{th}component of $\beta_k(t_i)$ via the interval $\beta_k(t_i) \pm z_{1-\alpha/2}\sqrt{\hat{\sigma}_e^2 \times a_{ii}}$, where $\hat{\sigma}_e^2 = [(\boldsymbol{y} - \mathbf{A}_\lambda \boldsymbol{y})^T(\boldsymbol{y} - \mathbf{A}_\lambda \boldsymbol{y})]/(n-m)$, a_{ii} is the i^{th}diagonal element of the corresponding hat matrix $\mathbf{A}_\lambda$ for β_k and $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ standard normal percentile.

Wahba's "Bayesian Confidence Intervals" have been often used in the nonparametric community. Wahba [44] showed that the average of the coverage probability across points of these pointwise intervals is very close to nominal level for large n. She also commented that even if the confidence intervals are derived from a Bayesian perspective, they perform well in the frequentist realm. Nychka [32] offers an excellent discussion on why this is true.

In their paper, Brumback and Rice [4] utilized a hierarchical bootstrap method to assess the variability of the fitted functions instead of using the variance components estimators (it is well know that these estimators often underestimate the true parameters). For each bootstrapped sample 1.5 hours was required to obtain the estimated sample profiles (as reported by [4]). As a result, a partially parametric version of the method was implemented (see [4], for more details). They computed 35 bootstrap samples and this took approximately 45 mins. In contrast, the confidence intervals computed in this paper for the progesterone profiles were obtained with the same computational effort involved in the estimation of the profiles.

Our estimated function profiles seem to agree with the ones obtained by Brumback and Rice. In addition, the "confidence" intervals also allow us to see that, on average, the production of progesterone in the conceptive group drops significantly

from day 4 until around day 8 (when ovulation occurs) as compared to the hormone production of the non conceptive group. This result differs from the findings by [4]. Their bootstrap sample suggested that the decrease in progesterone for the conceptive group was not significant. The discrepancy between our findings and those of Brumback and Rice may be due to the small bootstrap sample they employed in their analysis or with our interpretation of the confidence intervals. Nychka [32] pointed out that these intervals may not be reliable at specific points, even more if those points are part of a sharp peak or deep valley in the function. However, he also mentioned that it provides "a reasonable measure of the spline estimate's accuracy provided that the point for evaluation is chosen independently of the shape" of the function. It is known that the women are more fertile around day 3 to day 9, making it an interval targeted for studying before the start of the analysis. Also, we do not consider that the bump that forms in the interval of interest is that sharp. Hence, we believe that the confidence intervals provide reasonable evidence that the profiles are different.

3.2. Ridge Regression and Penalized Spline Regression

To exemplify the use of Theorem 2.1 in the ridge regression setting, we have selected a data set that is freely distributed by *StatLib* at http://lib.stat.cmu.edu. The data set consists of 1150 heights measured at 1 micron intervals along the drum of a roller (i.e. parallel to the axis of the roller). The units of height are not given and the zero reference height is arbitrary.

To fit this data we used a model of the form (3) with $\mathbf{X} = [\mathbf{T}, \mathbf{U}]$ and corresponding vector of coefficients $\boldsymbol{\beta} = [\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{b}^{\mathrm{T}}]^{\mathrm{T}}$, where

$$\mathbf{T} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_{1150} \end{bmatrix} \quad \text{and} \quad \mathbf{U} = \begin{bmatrix} (t_1-\xi_1)_+ & \cdots & (t_1-\xi_k)_+ \\ (t_2-\xi_1)_+ & \cdots & (t_2-\xi_k)_+ \\ \vdots & \cdots & \vdots \\ (t_{1150}-\xi_1)_+ & \cdots & (t_{1150}-\xi_k)_+ \end{bmatrix}. \tag{22}$$

The generalized ridge regression estimator of $\boldsymbol{\beta}$ is then obtained by minimizing the PLS criterion (4), with $\mathbf{K}$ as in (5) and $m = 2$.

Applying the results of Theorem 2.1 we can write a parallel mixed-effects model representation for this ridge regression problem. This particular framework was considered by [36] who describe in Section 4.9 of their book how to represent p-splines as BLUP's and illustrated how to use available software packages, like SAS proc mixed or the S-PLUS function lme, to obtain a fitted curve for the data. In view of the equivalence theorem, an alternative approach would be to use the connection between PLS and the Bayesian model so the Kalman filter can be implemented for purposes of computing estimators and "confidence" intervals. Another comprehensive description of the use of P-splines in the semi-parametric regression setting using Bayesian techniques is given in [5]. In this paper, we will use the Bayesian connection.

Assume that the vectors $\boldsymbol{\theta}$, $\boldsymbol{b}$ and $\boldsymbol{e}$ are independently normally distributed with zero mean and respective variance-covariance matrices $\nu\mathbf{I}$, $\sigma_b^2\mathbf{I}$ and $\sigma_e^2\mathbf{I}$. Then, by the equivalence theorem, the minimizer of (4) is the same as the limit, when ν is allowed to go to infinity, of the posterior mean of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}$.

Again, this Bayesian model representation of the ridge regression example will permit the use of the Kalman filter algorithm for the computation of the estimated

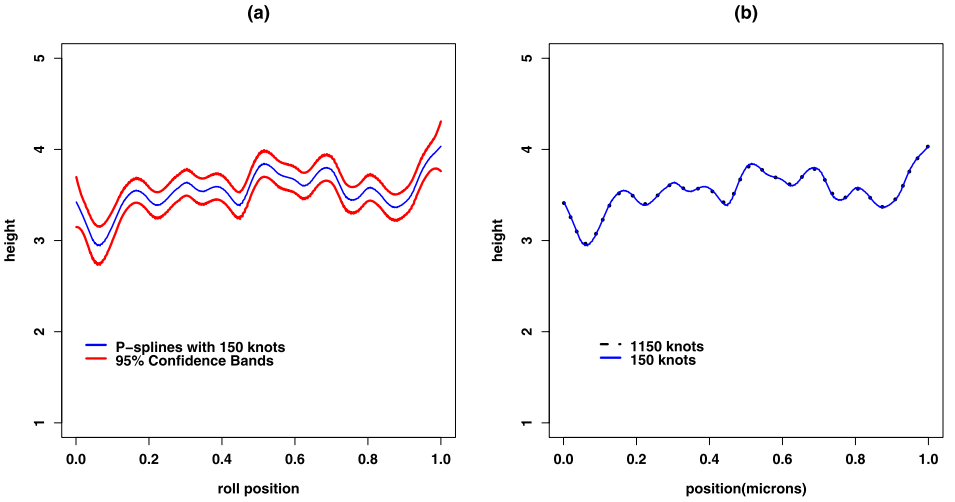

FIG 3. *Figure (a) shows the P-spline estimator with 150 knots for the roller height with its respective 95% confidence bands. The corresponding figure for the P-spline estimator with 1150 knots looks exactly the same and has been omitted. Figure (b) shows the comparison between the p-spline estimator with 150 knots and the P-spline with 1150 knots. There is no visual difference and both procedures yielded an estimated error variance of 0.36.*

function and its respective "confidence intervals". For the explicit form of the state-space model see the appendix.

For this particular example, we considered two different model versions, one using $k = 150$ knots and the other with $k = 1150$ knots. This because we wanted to contrast the computational performance of the P-splines versus the computational effort required using smoothing splines and the Kalman filter. We should remark here that, when $k = 1150$, basically we have a smoothing spline estimator which basis functions are the polynomials and the truncated power functions $(t_i - \xi_k)_+$, for $i = 1, \ldots, 1150$.

Figure 3 shows the smooth estimated curve for the roller height and corresponding 95% "confidence intervals". The smoothing parameters were, respectively, $\lambda_{150} = 0.043$ and $\lambda_{1150} = 0.095$. They were selected via GCV and as we can see the GCV methods adjusts the smoothing parameters according to the number of knots used.

One of the main arguments in favor of using P-splines in lieu of smoothing splines is that, by reducing the number of knots involved in the model, we increase the computational efficiency involved in calculating the spline estimator. This is true when using brute force methods, i.e., direct inversion of the matrix (12). However, when using the proposed Kalman filter algorithm, the computational advantage of the P-splines over the smoothing splines disappear as we can see in Table 2.

3.3. Mixed-Effects Model

In this last example, we want to illustrate the application of the equivalence theorem in the mixed-effects model setting. By finding the equivalent Bayesian model of a mixed-effects model representation we will demonstrate the use of Kalman filtering for estimating parameters in a setting that it has seldom being used and that it can benefit from the reduced computational burden of estimating parameters

TABLE 2
Run time comparisons between a p-splines estimator with 150 knots, a P-spline estimator with 1150 knots and a smoothing spline estimator (technically 1150 knots but using as basis functions the polynomials and equation (21)). Both estimators were computed using code in R and the time does not include computation of the smoothing parameter

Knots	Real Time
P-spline 150	48.34 secs.
P-spline 1150	54.55 secs.
Smoothing Spline	48.26 secs.

and variance components. It is true that, if we have a "reasonable" number of observations and a specific covariance structure, like the ones provided by existing software, it will be advisable to use these procedures in lieu of the Kalman filter. However, there are occasions where the number of observations is really large. Then we can take advantage of the computational efficiency of the Kalman filter.

Our example deals with a randomized block design, where the data consists of 37 patients, which represent the random blocks, and a set of consecutive Hamilton depression scores measured over the course of 6 weeks (see Figure 4). The data set is part of a study conducted by [34] and it is available at `http://tigger.uic.edu/~hedeker/`.

We model the data as

$$y_{ij} = \beta_0 + \beta_1 \text{week} + b_i + e_{ij},$$

where the y_{ij}'s are the depression scores, for $i = 1, \ldots, 37$, β_0 and β_1 are fixed parameters and week $= 0, 1, \ldots, 5$, is the week number where the score was measured. The random effects due to each patient are denoted by the b_i's and they are independent of the errors e_{ij}'s which are generated by an autoregressive process of order 1, i.e., $e_{ij} = \phi e_{i,j-1} + a_j(t_i)$, with ϕ a constant and $a_j(t_i)$ independent, identically distributed zero mean errors with variance σ_e^2.

Let $\boldsymbol{y}$ be the vector of depression scores such that $\boldsymbol{y} = [\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_{37}^{\mathrm{T}}]^{\mathrm{T}}$, for $\boldsymbol{y}_i = [y_{i0}, y_{i1}, \ldots, y_{i5}]^{\mathrm{T}}$. Denote by $\mathbf{1}_n$, the vector of all ones of dimension $n \times 1$ and

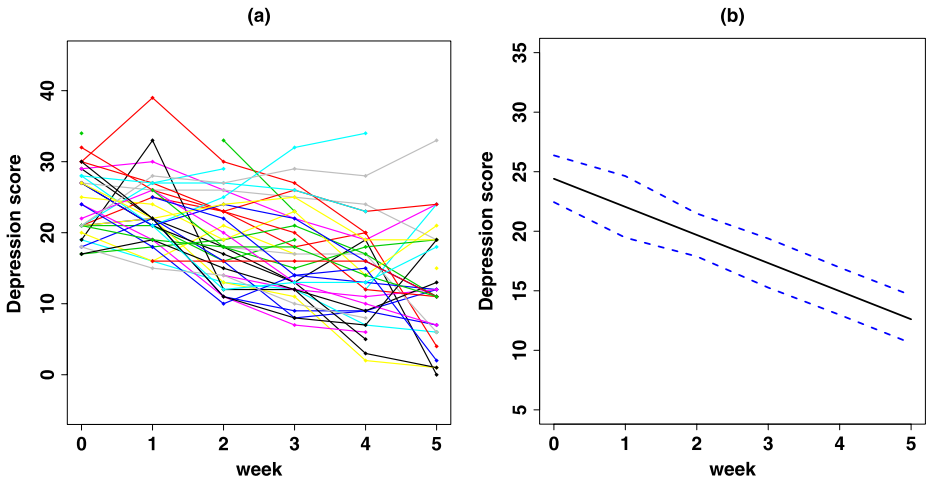

FIG 4. *(a) Hamilton depression scores for 37 patients measured over the period of 6 weeks. (b) Estimated regression line,* $y = 24.41 - 2.36$ week*, with respective 95% confidence bands.*

$\mathbf{week} = [0, 1, \ldots, 5]^{\mathrm{T}}$. In matrix form our model becomes

$$\boldsymbol{y} = \mathbf{T}\boldsymbol{\theta} + \boldsymbol{b} + \boldsymbol{e},$$

with $\boldsymbol{\theta} = [\beta_0, \beta_1]^{\mathrm{T}}$, $\mathbf{T} = [\mathbf{1}_{37} \bigotimes \mathbf{1}_5, \mathbf{1}_{37} \bigotimes \mathbf{week}]$, and $\boldsymbol{b} = \mathbf{1}_5 \bigotimes [b_1, b_2, \ldots, b_{37}]^{\mathrm{T}}$ and $\boldsymbol{e} = [\boldsymbol{e}_1^{\mathrm{T}}, \boldsymbol{e}_2^{\mathrm{T}}, \ldots, \boldsymbol{e}_{37}^{\mathrm{T}}]^{\mathrm{T}}$, for $\boldsymbol{e}_i = [e_{i0}, e_{i1}, \ldots, e_{i5}]^{\mathrm{T}}$. Here, we model the $\boldsymbol{b}$ as normally distributed with zero mean and variance-covariance matrix $\mathbf{R} = \{\xi_r(t_1)\}_{\substack{r=1,5 \\ i=1,5}}$ and

$$\mathbf{W} = \begin{bmatrix} \frac{\sigma_e^2}{1-\phi^2} & \frac{\phi}{1-\phi^2} & \frac{\phi^2}{1-\phi^2} & \cdots & \frac{\phi^n}{1-\phi^2} \\ \frac{\phi}{1-\phi^2} & \frac{\sigma_e^2}{1-\phi^2} & \frac{\phi}{1-\phi^2} & \cdots & \frac{\phi^{n-1}}{1-\phi^2} \\ \vdots & \cdots & \ddots & \cdots & \vdots \\ \frac{\phi^n}{1-\phi^2} & \frac{\phi^{n-1}}{1-\phi^2} & \cdots & \frac{\phi}{1-\phi^2} & \frac{\sigma_e^2}{1-\phi^2} \end{bmatrix},$$

where W is the variance-covariance matrix of the AR(1) errors.

To find the corresponding Bayesian model, let $\boldsymbol{b}$ and $\boldsymbol{e}$ keep their distributions and assume that $\boldsymbol{\theta}$ is normally distributed with zero mean and variance-covariance matrix $\nu\mathbf{I}$. Once in the Bayesian form, we check that our observations Y_{ij} can be represented using the state-space equations (19)-(20). The equivalence theorem hold regardless of the state-space structure but, if we have that structure, then we can apply the efficient Kalman filter algorithm of [13] and estimate all our parameters with linear computational efficiency.

Figure 4 shows the estimated regression line for the Hamilton depression scores over the 6 week period. The variance components for this example are estimated via REML/GML and are $\hat{\phi} = 0.97$, $\hat{\sigma}_e^2 = 1.214$ and $\hat{\sigma}_b^2 = 0.00132$. The corresponding estimated values for the regression coefficients are $\hat{\theta}_0 = 24.41$ and $\hat{\theta}_1 = -2.36$.

4. Summary

In this paper, we have reviewed known results concerning the numerical equivalence of (1) a smoothing spline estimator and a particular mixed-effects model and (2) a smoothing spline estimator and the posterior mean of Wahba's Gaussian model and focus on the more general framework of frequentist and Bayesian mixed-effects models and penalized least-squares estimation as seen in Theorem 2.1. This result broadens the number of methodological resources available for computing BLUPs, posterior means, likelihoods and minimizers of penalized least squares criteria and facilitates the use of existing methodological tools, as exemplified by Theorem 2.2 and our examples.

The link between the Bayesian mixed-effects model and the two other model settings allowed us to obtain Bayesian "confidence" intervals for the profile groups (instead of the computationally demanding bootstrap method of Brumback and Rice) and facilitated the analysis of the profile differences during the fertile days. Example 2 showed us that the Kalman filter implementation is not restricted to Wahba's Bayesian model. More generally, the idea carries over to settings involving p-splines, Kernel estimators, differences, etc. Lastly, this link allows for the implementation of a computationally efficient Kalman filter algorithm in many cases of interest. Kalman filter algorithms have been used to compute smoothing splines type estimators [19, 25, 26, 48]. But, they have been sparsely used in mixed-effects model settings. To this author knowledge, only [38] and, more recently, [30] have applied the Kalman filter to mixed-effects models. In the mixed-effects framework, the

techniques employed for the analysis of large data sets require the use of computer intensive methods like the EM or MCMC algorithms [1, 39], conjugate gradient iterative methods [42], or the use of high performance computing environments. Some of the methods mentioned in these references assume that observations are generated by Brownian motion or ARMA processes and, whenever we have this type of processes, we have a state-space structure that can be exploited, as demonstrated in our examples, to reduce the computational burden. Observations generated by longitudinal analysis (as in example 3), repeated measurements or any process that depends on an ordering variable can also frequently be assumed to have a state-space representation and can, as a result, benefit from the computational efficiency of the Kalman filter.

Appendix: State-Space Forms

In this section we will explicitly describe the state-space forms used for the application of the Kalman filter in each of our examples. Since the form of the errors $\boldsymbol{u}(t_i)$ in equation (20) is assumed to be the same for the varying-coefficient case and the mixed-effects model, we will show the derivation for the varying coefficient case and detail the small changes needed for the mixed-effects case. We will leave for the last the ridge regression example.

To employ the Kalman filter for computation of the varying coefficient example we need to show that the varying coefficients have a state-space representation. That is, we need to be able to write equation (1) using equations (19)–(20). Since the $\beta_k(\cdot)$ are assumed to be smooth functions of t, we model them as

$$\beta_k(t_i) = \sum_{q=0}^{m-1} \theta_{kq} t^q + \sigma_b^2 Z_k(t_i), \tag{A.1}$$

for $k = 1, 2$ and $m = 2$, where (without loss of generality) we can take t in $[0, 1]$ and

$$Z_k(t) = \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} dW_k(u),$$

with $W_k(\cdot)$ standard Wiener processes. To simplify matters, first assume that $\beta_k(t_i) = \sigma_b^2 Z_k(t_i)$. Then, $\beta_k(t_{i+1})$ can be written as σ_b^2 times

$$\int_0^{t_i} \frac{(t_{i+1}-s)_+^{m-1}}{(m-1)!} dW_k(s) + \int_{t_i}^{t_{i+1}} \frac{(t_{i+1}-s)_+^{m-1}}{(m-1)!} dW_k(s).$$

Taking

$$u_k(t_i) = \int_{t_i}^{t_{i+1}} \frac{(t_{i+1}-u)_+^{m-1}}{(m-1)!} dW_k(u),$$

for $t_i < t_j$, the covariance between $u_k(t_i)$ and $u_k(t_j)$ is found to be equal to

$$\int_0^{t_i} \frac{(t_i-u)^{m-1}(t_j-u)^{m-1}}{[(m-1)!]^2} du.$$

For the remaining integral, add and subtract t_i inside $(t_{i+1}-u)^{m-1}$ and apply the Binomial theorem. Upon doing this, a state-space representation results with $\mathbf{F}(t_i)$

equal to

$$\text{(A.2)} \qquad \mathbf{F}(t_i) = \begin{bmatrix} 1 & (t_{i+1}-t_i) & \frac{(t_{i+1}-t_i)^2}{2!} & \dots & \frac{(t_{i+1}-t_i)^{m-1}}{(m-1)!} \\ 0 & 1 & (t_{i+1}-t_i) & \dots & \frac{(t_{i+1}-t_i)^{m-2}}{(m-2)!} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & . & . & . & 1 \end{bmatrix},$$

$\mathbf{Z}_k(t_i) = [Z_k(t_i), Z_k^{(1)}(t_i), \dots, Z_k^{(m-1)}(t_i)]^T$, $\boldsymbol{u}_k(t_i) = [u_k(t_i), u_k^{(1)}(t_i), \dots, u_k^{(m-1)}(t_i)]^T$ and $\boldsymbol{Z}_k(t_{i+1}) = \mathbf{F}(t_i)\boldsymbol{Z}_k(t_i) + \boldsymbol{u}_k(t_i)$.

Now, rearranging the observations with respect to the time t_i define

$$\boldsymbol{y}_{iw}^{\mathrm{T}} = [y_{iw1}, \dots, y_{iwc_w}]^{\mathrm{T}}$$

with $\boldsymbol{y}_{iw}^{\mathrm{T}}$ the responses for woman w at time t_i observed at cycles $1, \dots, c_w$, with corresponding vector of random errors $\boldsymbol{e}_{iw}$. Let $\boldsymbol{x}(t_i) = [\boldsymbol{Z}_1(t_i), \boldsymbol{Z}_2(t_i)]^T$, $\boldsymbol{u}(t_i) = [\boldsymbol{u}_1(t_i), \boldsymbol{u}_2(t_i)]^T$ and $\boldsymbol{X}_{kwi} = [X_{kw1i}, \dots, X_{kwc_wi}]^T$. Then, taking

$$\boldsymbol{h}(t_i) = [\boldsymbol{X}_{1wi}^T, \mathbf{0}, \dots, \mathbf{0}, \boldsymbol{X}_{2wi}^T, \mathbf{0}, \dots, \mathbf{0}]^T,$$

we arrive at the state-space model

$$\boldsymbol{y}_{iw} = \boldsymbol{h}(t_i)\boldsymbol{x}(t_i) + \boldsymbol{e}_{iw},$$

$$\boldsymbol{x}(t_{i+1}) = \mathbf{F}^{\star}(t_i)\boldsymbol{x}(t_i) + \boldsymbol{u}(t_i),$$

where $\mathbf{F}^{\star}(t_i)$ is the block diagonal matrix of size $2m \times 2m$ with diagonal blocks $\mathbf{F}(t_i)$, $i = 1, \dots, n$.

Application of the standard Kalman filter to the vector of observations $\boldsymbol{y}_{iw}$ will yield coefficient functions estimates that disregard the polynomial term in (A.1). To account for that, we must employ the diffuse Kalman filter as in [13]. This entails a slight modification of our approach wherein the Kalman filter is applied to the vector of observations $\boldsymbol{y}_{iw}$ and each of the vectors $\mathbf{1}_{\boldsymbol{n}}$, $\boldsymbol{t}$, $\boldsymbol{t}^2, \dots, \boldsymbol{t}^{(m-1)}$, where $\boldsymbol{t}^r = [t_1^r, t_2^r, \dots, t_n^r]^{\mathrm{T}}$ (see [11], for a detailed derivation).

For our mixed-effects example we need to show that $\boldsymbol{e}$ can be represented in a state-space form and stack the respective state vectors, errors and matrices. We will proceed as follows: since the errors $e(t_i)$ are generated by an AR(1) process, they can be written as $e(t_{i+1}) = \phi e(t_i) + a_j(t_i)$, with ϕ a non random coefficient. This entails that the transition matrix $\mathbf{F}^{\star}(t_i) = diag\{\mathbf{F}(t_i), \phi\}$, with $\mathbf{F}(t_i)$ as in (A.2) and $\boldsymbol{h}(t_i) = [1, 0, 1]$. Take the state vector, $\boldsymbol{x}(t_i)$, to be equal to $[\boldsymbol{Z}_k(t_i)^{\mathrm{T}}, e(t_i)]^{\mathrm{T}}$, $\boldsymbol{u}(t_i) = [\boldsymbol{u}_k(t_i)^{\mathrm{T}}, a_j(t_i)]^{\mathrm{T}}$ with $m = 2$, where $\boldsymbol{Z}_k(t_i)$ and $\boldsymbol{u}_k(t_i)$ are as in the varying coefficient case. Specific details about the form of the state vector and the vector $\boldsymbol{u}(t_i)$ of the state equation (20), as well as a more general form for an ARMA model, can be found in [14].

Lastly, the state-space representation for the ridge regression example is found by taking the state vector to be $\boldsymbol{x}(t_i) = [x(t_i), x^{(1)}(t_i), \dots, x^{(m-1)}(t_i)]^T$, with $x(t_i) = \sum_{k=1}^{j} \beta_k (t_i - \xi_k)^{m-1}$ for $t_i \in [\xi_j, \xi_{j+1})$ (using the definition of the truncated power function), and $x^{(r)}(t_i)$ the r^{th} derivative of $x(t_i)$, $r = 1, \dots, (m-1)$. Then,

$$\boldsymbol{x}(t_{i+1}) = \mathbf{F}(t_i)\boldsymbol{x}(t_i) + \boldsymbol{u}(t_i),$$

with $\mathbf{F}(t_i)$ as in (A.2) and $\boldsymbol{u}(t_i) = [u(t_i), u^{(1)}(t_i), \dots, u^{(m-1)}(t_i)]^{\mathrm{T}}$, where

$$u(t_i) = \begin{cases} 0 & \text{if} \quad t_{i+1} \in [\xi_j, \xi_{j+1}), \\ \beta_{j+1}(t_{i+1} - \xi_{j+1})^{m-1} & \text{if} \quad t_{i+1} \in [\xi_{j+1}, \xi_{j+2}). \end{cases}$$

To complete the state-space formulation, take the vector $\mathbf{h}(t_i)$ to have dimension $m \times 1$ with one in the first position and the rest of its elements equal to zero.

References

[1] Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Statist. Meth. Med. Res.* **55** 117–128.

[2] Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Prentice Hall, Englewood Cliffs, NJ.

[3] Ansley, C. F. and Kohn, R. (1985). Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. *Ann. Statist.* **13** 1286–1316.

[4] Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc.* **93** 961–993.

[5] Crainiceanu, C. M., Ruppert, D. and Wand, M. P. (2005). Bayesian analysis for penalized spline regression using WinBugs. *Journal of Statistical Software* **14** 1–24.

[6] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

[7] Cressie, N. (1990). Reply: Letters to the editor. *Amer. Statist.* **44** 256–258.

[8] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–22.

[9] Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–102.

[10] Eppright, E., Fox, H., Fryer, B., Lamkin, G., Vivian, V. and Fuller, E. (1972). Nutrition of infants and preschool children in the North Central Region of the United States of America. *World Review of Nutrition and Dietetics* **14** 269–332.

[11] Eubank, R. L., Huang, C., Muñoz Maldonado, Y., Wang, N., Wang, S. and Buchanan, R. J. (2004). Smoothing spline estimation in varying coefficient models. *J. Roy. Statist. Soc. Ser. B* **66** 653–667.

[12] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*, 1st ed. Marcel Dekker, New York.

[13] Eubank, R. L., Huang, C. and Wang, S. (2003). Adaptive order selection for spline smoothing. *J. Comput. Graph. Statist.* **12** 2546–2559.

[14] Eubank, R. L. (2006). *A Kalman Filter Primer*. Chapman & Hall/CRC, Boca Raton, FL.

[15] de Finetti, B. (1964). Foresight: Its logical laws, its subjective sources. In *Studies in Subjective Probability*. Wiley, New York.

[16] Gantmakher, F. R. (1959). *The Theory of Matrices*. Chelsea Pub. Co., New York.

[17] Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized-cross validation as a method for choosing a good ridge parameter. *Technometrics* **58** 215–223.

[18] Graybill, F. A. (1976). *The Theory and Application of the Linear Model*. Duxbury, North Scituate, MA.

[19] Guo, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128.

[20] Guo, W. (2003). Functional data analysis in longitudinal settings using smoothing splines. *Statist. Meth. Med. Res.* **13** 1–24.

[21] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **4** 757–796.
[22] Heckman, N. (1997). The theory and application of penalized least squares methods or reproducing kernel Hilbert spaces made easy. Technical Report 216, The Univ. British Columbia.
[23] Henderson, C. R., Kempthorne, O., Searle, S. R. and Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15** 192–218.
[24] Householder, A. (1964). *The Theory aof Matrices in Numerical Analysis*. Dover, New York.
[25] Kohn, R. and Ansley, C. F. (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J. Sci. Statist. Comput.* **8** 33–48.
[26] Kohn, R. and Ansley, C. F. (1989). A fast algorithm for signal extraction, influence and cross-validation in state-space models. *Biometrika* **76** 65–79.
[27] Kohn, R., Ansley, C. F. and Tharm, D. (1993). Performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Amer. Statist. Assoc.* **86** 1042–1050.
[28] Koopman, S. J. and Durbin, J. (1998). Fast filtering and smoothing for multivariate state space models. *J. Times Ser. Anal.* **21** 281–296.
[29] Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41.
[30] Piehpo, H. P. and Ogutu, J. O. (2007). Simple state-space models in a mixed-model framework. *Amer. Statist.* **61** 224–232.
[31] Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.* **83** 1134–1143.
[32] Nychka, D. (2000). Spatial process estimates as smoothers. In *Smoothing and Regression. Approaches, Computation and Application*. Wiley, New York.
[33] Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when cell sizes are unequal. *Biometrika* **58** 545–554.
[34] Riesby, N., Gram, L.F., Bech, P., Nagy, A., Petersen, G. O., Ortmann, J., Ibsen, I., Dencker, S. J., Jacobsen, O., Krautwald, O., Søndergaard, I. and Chirstiansen, J. (1977). Imipramine: Clinical effects and pharmacokinetic variability. *Psychopharmacology* **54** 263–272.
[35] Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* **6** 15–32.
[36] Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, New York.
[37] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11** 735–757.
[38] Sallas, W. M. and Harville, D. A. (1981). Best linear recursive estimation for mixed linear models. *J. Amer. Statist. Assoc.* **76** 860–869.
[39] Schafer, J. L. and Recai, M. Y. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *J. Comput. Graph. Statist.* **11** 437–457.
[40] Speed, T. (1991). That BLUP is a good Thing: The estimation of random effects: Comment. *Statist. Sci.* **6** 42–44.
[41] Stein, M. L. (2000). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Statist.* **18** 1139–1157.
[42] Strandén, I. and Lidauer, M. (1999). Solving large mixed linear models

using preconditioned conjugate gradient iteration. *Journal of Diary Science* **82** 2779–2787.

[43] Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

[44] Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

[45] Wahba, G. (1990). Spline models for observational data. *SIAM* **59**. Philadelphia, Pennsylvania.

[46] Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13** 1378–1402.

[47] Wahba, G. (1990). Comment on Cressie: Letters to the editor. *Amer. Statist.* **44** 255–256.

[48] Wang, Y. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93** 341–348.

[49] Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *J. Roy. Statist. Soc. Ser. B* **60** 159–174.