Electronic Journal of Statistics
Vol. 3 (2009) 797–823
ISSN: 1935-7524
DOI: 10.1214/09-EJS428

Explicit connections between longitudinal data analysis and kernel machines

N.D. Pearce

School of Mathematics and Statistics,
University of New South Wales, Sydney, 2052, Australia
e-mail: nathan@maths.unsw.edu.au

M.P. Wand

School of Mathematics and Applied Statistics,
University of Wollongong, Wollongong, 2522, Australia
e-mail: mwand@uow.edu.au

Abstract: Two areas of research – *longitudinal data analysis* and *kernel machines* – have large, but mostly distinct, literatures. This article shows explicitly that both fields have much in common with each other. In particular, many popular longitudinal data fitting procedures are special types of kernel machines. These connections have the potential to provide fruitful cross-fertilization between longitudinal data analytic and kernel machine methodology.

Keywords and phrases: Best linear unbiased prediction, classification, generalized linear mixed models, machine learning, linear mixed models, reproducing kernel Hilbert spaces, penalized likelihood, support vector machines.

Received June 2009.

1. Introduction

Longitudinal data analysis is concerned with regression modelling and inference for data consisting of repeated measures on *units*, such as humans in a medical study. Since the seminal work of Harville [12] and Laird and Ware [17], linear mixed models have been the mainstay of longitudinal data analyses. The predominant distinguishing feature of linear mixed models, when compared with linear models, is the dichotomization of effects into fixed and random types. The fitting of fixed and random effects differ in that the latter is subject to a degree of shrinkage, dependent on the values of covariance parameters in the model. The concept of best linear unbiased prediction appealingly accommodates the handling of both types of effects (e.g. [26]). Expositions on longitudinal data analysis, including the role of linear mixed models, can be found in Diggle *et al.*

[*]We are grateful to the associate editor and a referee for their comments on an earlier version of this article. This research was partially supported by Australian Research Council Discovery Project DP0877055.

[6], Fitzmaurice, Laird and Ware [9], Fitzmaurice *et al.* [8], McCulloch, Searle and Neuhaus [19] and Verbeke and Molenberghs [35].

Kernel machines is a younger field that has most of its literature outside of Statistics. Essentially, kernel machines are flexible non-linear regression-type methods that use regularization to avoid overfitting. The name 'kernel' comes from the fact that the theory and methods of kernel machines is underpinned by *reproducing kernel Hilbert space (RKHS)* theory (e.g.[1, 16]), although other frameworks such as Gaussian process theory (e.g. [25]) can also be used. Whilst kernel machines can handle general response variables, the majority of their literature is geared towards classification in which the response is categorical. In particular, *support vector machines* (e.g. [4, 20]) are a sub-class of kernel machines and, in the 1990s, emerged as a powerful and elegant family of classifiers. The monograph of Schölkopf and Smola [29] and the web-site maintained by these authors, `www.kernel-machines.org`, are two portals to the large and expanding literature on kernel machines.

The main goal of this article is to expose the commonalities shared by longitudinal data analysis and kernel machines. In particular we show, explicitly, that many popular longitudinal fitting procedures are special types of kernel machines. There are at least two potential payoffs from such links: (a) the enrichment of longitudinal models to cope with non-linear predictor effects, and (b) adaptation of kernel machine classifiers to account for within-subject correlation when applied to longitudinal data. Sections 4.1.1–4.1.3 give some details on (a). Section 5.2 contains an illustration of (b).

Some recent related work is Gianola, Fernando and Stella [10] and Liu, Lin and Ghosh [18], each of whom combine linear mixed models with kernel machines to analyze very high-dimensional genetic data-sets. However, neither of these papers deal with regular longitudinal data analysis models. James and Hastie [15] and Müller [21] are examples of articles concerned with classification when the data are longitudinal.

The connections between longitudinal data analysis and kernel machines are stronger in the case of continuous responses. A concise overview of continuous response longitudinal data analysis is given in Section 2. A summary of kernel machines and their RKHS substructure is given in Section 3. Section 4 forms the main body of the paper and gives an explicit case-by-case description of kernel machine representations of popular longitudinal data analytic models, as well as explaining some non-linear (kernel-based) extensions. Generalized response models and kernel machines are treated in Section 5. Concluding discussion is given in Section 6.

2. Continuous response longitudinal data analysis

In this section, and for following two sections, we suppose that the response variables are continuous, and without strong departures from normality. In this case, the main vehicle for longitudinal data analysis is the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \qquad \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{bmatrix} \right) \tag{1}$$

where, for a general random vector $\boldsymbol{v}$, the notation $\boldsymbol{v} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is shorthand for the mean vector $E(\boldsymbol{v})$ equal to $\boldsymbol{\mu}$ and the covariance matrix $\text{Cov}(\boldsymbol{v})$ equal to $\boldsymbol{\Sigma}$. A common special case of (1) is

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{bmatrix} \right). \tag{2}$$

The use of (1) for longitudinal data analysis dates back to Laird and Ware [17]. Good summaries of estimation and prediction within this linear mixed model structure may be found in, for example, Robinson [26], Verbeke and Molenberghs [35], McCulloch *et al.* [19] and Ruppert, Wand and Carroll (Chapter 4) [28]. We will just present the main results here.

For given covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ the theory of best linear unbiased prediction (BLUP) can be used to guide choice of $\boldsymbol{\beta}$ and $\boldsymbol{u}$, and results in the criterion:

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u}. \tag{3}$$

This is minimized by

$$\begin{aligned} \boldsymbol{\beta}_{\text{BLUP}} &= (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}, \\ \boldsymbol{u}_{\text{BLUP}} &= \boldsymbol{G} \boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\text{BLUP}}) \end{aligned} \tag{4}$$

where $\boldsymbol{V} = \text{Cov}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R}$. Expressions (4) are known as the BLUPs of $\boldsymbol{\beta}$ and $\boldsymbol{u}$.

In practice, longitudinal data are fitted via the steps:

1. Estimation of $\boldsymbol{G}$ and $\boldsymbol{R}$. Usually, these matrices are restricted to a parametrized class of covariances matrices. Most commonly, estimation is achieved via maximum likelihood, or restricted maximum likelihood (REML), under the normality assumption (2).
2. Substitution of the estimated covariance matrices into (4). The resulting estimators, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$, are commonly known as estimated BLUPs, or EBLUPs for short.

The EBLUP phrase can be transferred to any linear function of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$. Thus, $\boldsymbol{A}\widehat{\boldsymbol{\beta}} + \boldsymbol{B}\widehat{\boldsymbol{u}}$ is the EBLUP of $\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{u}$ for any pair of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ for which the multiplications and summation are defined.

These two steps show a division into two types of estimation targets that arise in longitudinal data analysis: the *covariance parameters* in the $\boldsymbol{G}$ and $\boldsymbol{R}$ matrices, and the *effects* $\boldsymbol{\beta}$ and $\boldsymbol{u}$. The strong connections between longitudinal data analysis and kernel machines occur at the EBLUP step for estimation of the fixed and random effects. For this reason, we will not dwell on the estimation of the covariance parameters. These will be taken as given when we revisit longitudinal data analysis in Sections 4 and 5.

3. Kernel machines

In the past decade or so *kernel machines* have emerged as an important methodology for classification and regression problems. A special sub-class of kernel machines is *support vector machines* where classifiers are constructed with respect to constraints on the margin of the classification boundary, and may be represented as a regularized optimization problem with respect to the *hinge loss* function (e.g. [4]). However, through the allowance of general loss functions, kernel machines are a much broader class of methods. Kernel machines with squared error loss (e.g. [33]) include popular statistical methods such as kriging (e.g. [5, 32]), smoothing splines (e.g. [11, 36]) and additive models (e.g. [13]). Zhu and Hastie [40] explored the use of binomial log-likelihood loss in the kernel machine framework and coined the term *kernel logistic regression*. Kernel logistic regression and support vector machines are both special cases of *large margin* kernel machines (e.g. [39]). Another prominent class of kernel machines is that corresponding to robust loss functions (e.g. [30]). Pearce and Wand [24] delineated connections between kernel machines and penalized spline methodology.

General kernel machines can be formulated in a number of ways. Among the most common are: optimization and projection within reproducing Hilbert spaces (e.g. [16]), maximum a posterior estimation in Gaussian processes (e.g. [25]) and Tikhonov regularization of ill-posed problems (e.g. [34]). Because of its prominence in the Statistics literature (e.g. [36]) we will use the first of these formulations in the remainder of the paper. Recent summaries of RKHS theory include Wahba [37] and Evgeniou, Pontil and Poggio [7]. An early reference is Aronszajn [1]. Relevant background material on Hilbert space projection theory may be found in, for example, Simmons [31] and Rudin [27].

In Section 3 of Pearce and Wand [24] we summarized the theory of RKHS for the purpose of explaining connections between kernel machines and penalized splines. That same section also provides useful background material for the current paper. For convenience, we reproduce the core aspects of it here. A RKHS on $\mathbb{R}^d$ is a Hilbert space of real-valued functions that is generated by a bivariate symmetric, positive definite function $K(\boldsymbol{s},\boldsymbol{t})$, $\boldsymbol{s},\boldsymbol{t}\in\mathbb{R}^d$, called the *kernel*. The steps for RKHS construction from K are:

1. Determine the eigen-decomposition of K: $K(\boldsymbol{s},\boldsymbol{t}) = \sum_{j=0}^{\infty}\lambda_j\phi_j(\boldsymbol{s})\phi_j(\boldsymbol{t})$. This series is assumed to be well-defined (e.g. uniformly convergent).
2. Define the space of real-valued functions on $\mathbb{R}^d$:
$$\mathcal{H}_K = \left\{ f : f = \sum_{j=0}^{\infty} a_j\phi_j, \quad \text{such that} \quad \sum_{j=0}^{\infty} a_j^2/\lambda_j < \infty \right\}.$$
3. Endow $\mathcal{H}_K$ with the inner product
$$\left\langle \sum_{j=0}^{\infty} a_j\phi_j, \sum_{j=0}^{\infty} a_j'\phi_j \right\rangle_{\mathcal{H}_K} = \sum_{j=0}^{\infty} a_j a_j'/\lambda_j.$$

4. Complete $\mathcal{H}_K$ by including limits.

It follows from Step 3 that the norm of $f = \sum_{j=0}^{\infty} a_j \phi_j$ in $\mathcal{H}_K$ is

$$\|f\|^2_{\mathcal{H}_K} = \sum_{j=0}^{\infty} a_j^2/\lambda_j.$$

The adjective 'reproducing' arises from the result

$$\langle K(\boldsymbol{s},\cdot), K(\boldsymbol{t},\cdot)\rangle_{\mathcal{H}_K} = K(\boldsymbol{s},\boldsymbol{t}). \tag{5}$$

This has important implications, and gives rise to the *'kernel trick'* that we discuss shortly.

Let $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \le i \le n$ be a data set, $\mathcal{L}(\cdot,\cdot)$ be a *loss* function and $\lambda > 0$ be a *regularization* parameter. The fit $\widehat{f}$ within $\mathcal{H}_K$, with respect to $\mathcal{L}$ and λ, is the solution to

$$\min_{f \in \mathcal{H}_K} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|^2_{\mathcal{H}_K} \right\}. \tag{6}$$

The solution to (6) can be shown to be of the form $\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{c}_i K(\boldsymbol{x}_i, \boldsymbol{x})$ for some $\widehat{c}_i$, $1 \le i \le n$, depending on $\mathcal{L}$, λ and the data. This result is known as the *representer theorem* of reproducing kernel Hilbert spaces. The 'kernel trick' is that we do not need to calculate the eigenfunctions, $\phi_0, \phi_1, \ldots$, in order to find the $\widehat{c}_i$. Because of (5) we only need evaluations of the inner products of the form $K(\boldsymbol{s},\boldsymbol{t})$. Popular kernels, particularly in machine learning contexts, include the pth degree polynomial kernel, $K(\boldsymbol{s},\boldsymbol{t}) = (1 + \boldsymbol{s}^T\boldsymbol{t})^p$, and the radial basis kernel, $K(\boldsymbol{s},\boldsymbol{t}) = \exp\left(-\gamma \parallel \boldsymbol{s} - \boldsymbol{t} \|^2\right)$, for some $\gamma > 0$.

The $\lambda\|f\|^2_{\mathcal{H}_K}$ term in (6) imposes a penalty on f. However, it is often desirable that certain functions in $\mathcal{H}_K$ are unpenalized. The simplest example of this is as follows. Let $\mathcal{H}_0$ be a subspace of $\mathcal{H}_K$ for which penalization is not desired. Mathematically, this means that fits over $\mathcal{H}_0$ are found by simply minimising $\sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i))$. Let $\mathcal{H}_1 = \mathcal{H}_0^{\perp}$ be the orthogonal complement of $\mathcal{H}_0$ and P_1 denote the linear operator corresponding to projection onto $\mathcal{H}_1$. Then

$$\mathcal{H}_K = \mathcal{H}_0 \oplus \mathcal{H}_1 \equiv \{\boldsymbol{v}_0 + \boldsymbol{v}_1 : \boldsymbol{v}_0 \in \mathcal{H}_0 \text{ and } \boldsymbol{v}_1 \in \mathcal{H}_1\},$$

with $\mathcal{H}_0$ being the *null space* of P_1, i.e.,

$$\mathcal{H}_0 = \{\boldsymbol{v} \in \mathcal{H}_K : P_1\boldsymbol{v} = \boldsymbol{0}\}.$$

With respect to the null space $\mathcal{H}_0$, smoothing parameter λ, and loss function $\mathcal{L}$, we define fits $\widehat{f}$ according to

$$\min_{f \in \mathcal{H}_K} \left\{ \sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \|P_1 f\|^2_{\mathcal{H}_K} \right\}. \tag{7}$$

It can also be shown ([1]) that $\mathcal{H}_0$ and $\mathcal{H}_1$ are reproducing kernel Hilbert spaces in their own right, with kernels K_0 and K_1 satisfying $K_0 + K_1 = K$.

4. Explicit connections for continuous responses

In this section we show, explicitly, how longitudinal data analyses are connected intimately with kernel machine methodology. Indeed, all longitudinal data analyses that use EBLUPs are actually just fitting a special type of kernel machine. To make these connections clear, and accessible to readers with a longitudinal data analysis background, we first treat some special cases of (1). We build up to fuller generality in the later subsections.

4.1. Random intercept model

The simple linear random intercept model is

$$y_{ij} = \beta_0 + \beta_1\, x_{ij} + U_i + \varepsilon_{ij}, \quad 1 \le j \le n_i, \quad 1 \le i \le m \tag{8}$$

where (x_{ij}, y_{ij}) is the jth predictor/response pair for subject i, and the $\varepsilon_{ij} \sim (0, \sigma_\varepsilon^2)$ are independent within-subject errors. The regression coefficients β_0 and β_1 are fixed effects, while the subject-specific intercepts $U_i \sim (0, \sigma_u^2)$ are independent random effects.

Given estimates $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\varepsilon^2$ of the variance components, the fitted line for subject i is

$$\widehat{\beta}_0 + \widehat{\beta}_1\, x + \widehat{U}_i, \quad 1 \le i \le m, \tag{9}$$

where $\widehat{\beta}_0$, $\widehat{\beta}_1$ and the $\widehat{U}_i$ are EBLUPs, as given by (4), with

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n_1} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{2n_2} \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_{m1} \\ \vdots & \vdots \\ 1 & x_{mn_m} \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{10}$$

Figure 1 shows the EBLUPs for data on longitudinally recorded weights of 48 pigs (source: [6]), with σ_u^2 and σ_ε^2 estimated via REML.

We now explain how (9) and the fitted lines in Figure 1 can be obtained as a solution to a RKHS optimization problem – thereby making them a special case of kernel machines. In the following discussion, we assume that the estimates of σ_u^2 and σ_ε^2 have been obtained (either via REML, or some other means) and are equal to $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\varepsilon^2$, respectively.

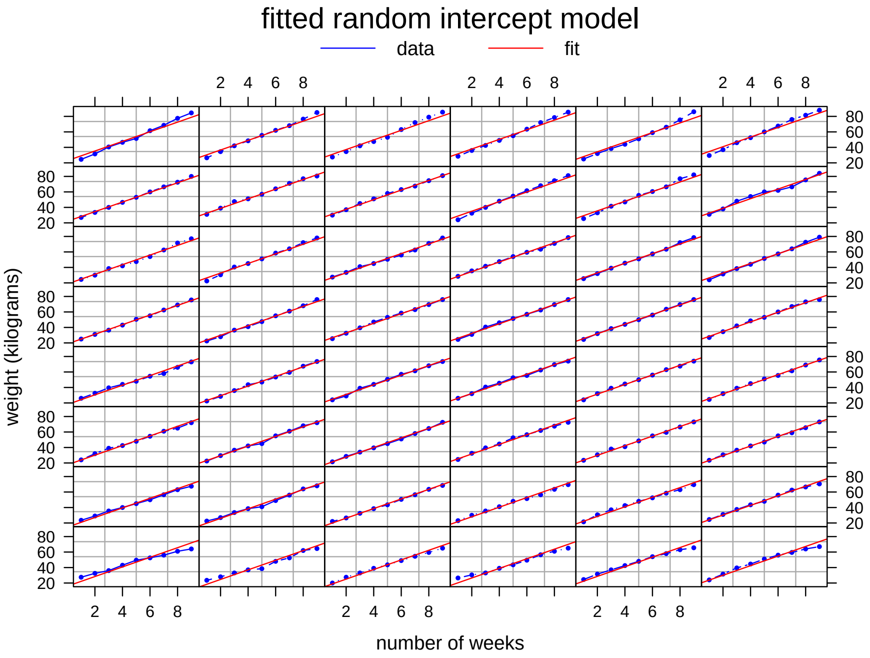

Fig 1. *The EBLUP-fitted lines to the pig-weights data for the simple linear random intercept model. The panels are ordered according to the size of the 48 pigs.*

Let $n = \sum_{i=1}^{m} n_i$ and re-subscript the (x_{ij}, y_{ij}) and ε_{ij} sequentially; i.e. according to the map:

$$\begin{array}{ccccccccccc} (1,1), & \ldots, & (1,n_1), & (2,1), & \ldots, & (2,n_2), & \ldots, & (m,1), & & \ldots, & (m,n_m) \\ \downarrow & \cdots & \downarrow & \downarrow & \cdots & \downarrow & \cdots & \downarrow & & \cdots & \downarrow \\ 1, & \ldots, & n_1, & n_1+1, \ldots, & n_1+n_2, & \ldots, & \sum_{j=1}^{m-1} n_j + 1, & \ldots, & & & n. \end{array} \tag{11}$$

This leads to the representation

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{m} U_j Z_{ij} + \varepsilon_i, \quad 1 \le i \le n,$$

where Z_{ij} is (i,j) entry of $\boldsymbol{Z}$ as given in (10) and is an indicator of (x_i, y_i) being measurements for subject j $(1 \le i \le n,\ 1 \le j \le m)$. Next, form the RKHS of real-valued functions on $\mathbb{R}^{m+1}$:

$$\mathcal{H}_K = \left\{ f : f(x, z_1, \ldots, z_m) = \beta_0 + \beta_1 x + \sum_{j=1}^{m} U_j z_j, \right\} \tag{12}$$

with kernel

$$K(\boldsymbol{s},\boldsymbol{t}) = K((s_1,\ldots,s_{m+1}),(t_1,\ldots,t_{m+1})) = 1 + s_1t_1 + \sum_{j=1}^{m} s_{1+j}t_{1+j}. \qquad (13)$$

Note that, while $\mathcal{H}_K$ is defined on the whole of $\mathbb{R}^{m+1}$, its members of interest in longitudinal data analysis are actually on:

$$\mathbb{R} \times (1,0,0,\ldots,0) \times (0,1,0,\ldots,0) \times (0,0,0,\ldots,1) \subset \mathbb{R}^{m+1}.$$

Let

$$\mathcal{H}_\beta = \{f : f(x,z_1,\ldots,z_m) = \beta_0 + \beta_1 x\} \qquad (14)$$

be a subspace of $\mathcal{H}_K$.

Theorem 1. *Let $(x_i,y_i,Z_{i1},\ldots,Z_{im})$, $1 \le i \le n$, be a sequentially subscripted longitudinal data set. Consider the RKHS $\mathcal{H}_K$ defined by (12) and (13) and subspace $\mathcal{H}_\beta$ given by (14). Let P_u be the projection operator onto $\mathcal{H}_u \equiv \mathcal{H}_\beta^\perp$. Then the solution to the RKHS optimization problem*

$$\min_{f\in\mathcal{H}_K}\left[\sum_{i=1}^{n}\{y_i - f(x_i,Z_{i1},\ldots,Z_{im})\}^2 + \lambda_u\|P_uf\|^2_{\mathcal{H}_K}\right] \qquad (15)$$

with $\lambda_u \equiv \widehat{\sigma}^2_\varepsilon/\widehat{\sigma}^2_u$ corresponds to the observed EBLUPs of (9). Explicitly, the solution to (15) is

$$\begin{aligned}
\widehat{f}(x,1,0,\ldots,0) &= \widehat{\beta}_0 + \widehat{\beta}_1\,x + \widehat{U}_1,\\
\widehat{f}(x,0,1,\ldots,0) &= \widehat{\beta}_0 + \widehat{\beta}_1\,x + \widehat{U}_2,\\
&\vdots\\
\textit{and}\quad \widehat{f}(x,0,0,\ldots,1) &= \widehat{\beta}_0 + \widehat{\beta}_1\,x + \widehat{U}_m,
\end{aligned}$$

where $x \in \mathbb{R}$, $\widehat{\beta}_0$, $\widehat{\beta}_1$ and the $\widehat{U}_i$ are given by (4) with $\boldsymbol{G} = \widehat{\sigma}^2_u\boldsymbol{I}$ and $\boldsymbol{R} = \widehat{\sigma}^2_\varepsilon\boldsymbol{I}$ and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are given by (10).

Proof. Note that the inner product for $\mathcal{H}_K$ (induced by K) is given by

$$\left\langle \beta_0 + \beta_1 x + \sum_{j=1}^{m} U_j z_j, \beta_0' + \beta_1' x + \sum_{j=1}^{m} U_j' z_j \right\rangle_{\mathcal{H}_K} = \beta_0\beta_0' + \beta_1\beta_1' + \sum_{j=1}^{m} U_jU_j'.$$

Next, if $f(x,z_1,\ldots,z_m) = \beta_0 + \beta_1 x + \sum_{j=1}^{m} U_j z_j$ is a typical element of $\mathcal{H}_K$, then

$$y_i - f(x_i,Z_{i1},\ldots,Z_{im}) = (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})_i,$$

where $\boldsymbol{y}$ is the $n \times 1$ vector containing the y_is, $\boldsymbol{\beta} = (\beta_0,\beta_1)^T$ and

$$\boldsymbol{u} = (U_1,\ldots,U_m)^T.$$

It follows immediately that

$$\sum_{i=1}^{n}\{y_i - f(x_i, Z_{i1}, \ldots, Z_{im})\}^2 = \|\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}\|^2$$

where $\|\cdot\|$ denotes Euclidean norm. The orthogonal complement of $\mathcal{H}_\beta$ is

$$\mathcal{H}_u = \mathcal{H}_\beta^\perp = \left\{ f : f(x, z_1, \ldots, z_m) = \sum_{j=1}^{m} U_j z_j, \right\}.$$

Then $P_u f = \sum_{j=1}^{m} U_j z_j$ and

$$\|P_u f\|^2_{\mathcal{H}_K} = \left\langle \sum_{j=1}^{m} U_j z_j, \sum_{j=1}^{m} U_j z_j \right\rangle_{\mathcal{H}_K} = \sum_{j=1}^{m} U_j^2 = \|\boldsymbol{u}\|^2.$$

The RKHS optimization problem (15) is therefore equivalent to

$$\min_{\boldsymbol{\beta},\boldsymbol{u}}\{(1/\widehat{\sigma}_\varepsilon^2)\|\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}\|^2 + (1/\widehat{\sigma}_u^2)\|\boldsymbol{u}\|^2\}$$

which corresponds to EBLUP for the random intercept model. □

4.1.1. Kernel-based extension to general mean curves

Note that the kernel for the simple linear random intercept model can be written as

$$K((s_1, \ldots, s_{m+1}), (t_1, \ldots, t_{m+1})) = K_\beta(\boldsymbol{s}, \boldsymbol{t}) + K_u(\boldsymbol{s}, \boldsymbol{t})$$

where $K_u(\boldsymbol{s}, \boldsymbol{t}) \equiv \sum_{j=1}^{m} s_{1+j} t_{1+j}$ corresponds to the random intercept structure and $K_\beta(\boldsymbol{s}, \boldsymbol{t}) \equiv 1 + s_1 t_1$ corresponds to the population mean structure. More general population mean structures can be obtained by replacement of $K_\beta(\boldsymbol{s}, \boldsymbol{t})$ by $K_0(\boldsymbol{s}, \boldsymbol{t}) + K_c(\boldsymbol{s}, \boldsymbol{t})$ where, for all $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^{m+1}$,

$$K_0(\boldsymbol{s}, \boldsymbol{t}) = K_{0,1}(s_1, t_1) \quad \text{and} \quad K_c(\boldsymbol{s}, \boldsymbol{t}) = K_{c,1}(s_1, t_1)$$

for kernels $K_{0,1}(s_1, t_1)$ and $K_{c,1}(s_1, t_1)$ defined on $\mathbb{R}^2$. The kernel $K_{0,1}$ corresponds to unpenalized functions and, typically, $K_{0,1}(s_1, t_1) = 1$. We can take $K_{c,1}$ to be any positive definite function on $\mathbb{R}^2$ such that its eigen-decomposition does not include functions in the RKHS generated by $K_{0,1}$. Usually we would want $K_{c,1}$ to have a rich eigen-decomposition so that non-linear mean structure can be well-handled. Examples include:

$$K_{c,1}(s_1, t_1) = \exp\{-\omega^2(s_1 - t_1)^2\}$$

and

$$K_{c,1}(s_1, t_1) = (1 + \omega|s_1 - t_1|)\exp(-\omega|s_1 - t_1|)$$

where $\omega > 0$ is a scale factor. Each of these kernels have infinite-length eigen-decompositions and result in an infinite dimensional RKHS. The 'kernel trick' implies that fitting and representation only depends on evaluations of $K_{c,1}$.

Let $\mathcal{H}_c$ be the RKHS generated by K_0 and K_c, respectively. Then

$$\mathcal{H}_K = \mathcal{H}_0 \oplus \mathcal{H}_c \oplus \mathcal{H}_u \tag{16}$$

is the RKHS generated by $K_0 + K_c + K_u$. Let $P_c : \mathcal{H}_K \to \mathcal{H}_c$ be the linear operator corresponding to projection onto $\mathcal{H}_c$ and let P_u be defined similarly for $\mathcal{H}_u$. Then a mean curve, with random intercept shifts, can be fitted via the RKHS minimization problem

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{n} \{y_i - f(x_i, Z_{i1}, \ldots, Z_{im})\}^2 + \lambda_c \|P_c f\|^2_{\mathcal{H}_K} + \lambda_u \|P_u f\|^2_{\mathcal{H}_K} \right]. \tag{17}$$

By the RKHS representer theorem, the solution is

$$f(x) = \beta_0 + \sum_{i=1}^{n} c_i K_c(x_i, x) + \sum_{j=1}^{m} U_i Z_{ij}$$

for coefficients $c_1, \ldots, c_n$. Via arguments similar to those used in the proof of Theorem 1, (17) reduces to the matrix algebraic problem,

$$\min_{\beta_0, \boldsymbol{c}, \boldsymbol{u}} \left(\|\boldsymbol{y} - \mathbf{1}\beta_0 - \boldsymbol{K}\boldsymbol{c} - \boldsymbol{Z}\boldsymbol{u}\|^2 + \lambda_c \boldsymbol{c}^T \boldsymbol{K}\boldsymbol{c} + \lambda_u \|\boldsymbol{u}\|^2 \right).$$

where $\boldsymbol{c} \equiv (c_1, \ldots, c_n)^T$ and $\boldsymbol{K} = [K_c(x_i, x_{i'})]_{1 \le i, i' \le n}$. Explicit solutions are a special case of (20) below.

4.1.2. Extension to additional linear predictors

Our final extension of the random intercept model involves the possible inclusion of additional predictors, assumed to have a linear effect on the mean of the response variable. Corresponding to each y_i, $1 \le i \le n$, let $\boldsymbol{x}_i^{\ell}$ an $p \times 1$ vector of such predictors. Then we should replace (16) by

$$\mathcal{H}_K = \mathcal{H}_\beta \oplus \mathcal{H}_c \oplus \mathcal{H}_u$$

where each of these RKHSs are now on $\mathbb{R}^{m+p+1}$ and

$$\mathcal{H}_\beta = \{ f : f(\boldsymbol{x}^{\ell}, x, z_1, \ldots, z_m) = [1 \ (\boldsymbol{x}^{\ell})^T]\, \boldsymbol{\beta} \}$$

corresponds to the fixed effects. The RKHS minimization problem is now of the form

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{n} \{y_i - f(\boldsymbol{x}_i^{\ell}, x_i, Z_{i1}, \ldots, Z_{im})\}^2 + \lambda_c \|P_c f\|^2_{\mathcal{H}_K} + \lambda_u \|P_u f\|^2_{\mathcal{H}_K} \right], \tag{18}$$

which reduces to

$$\min_{\boldsymbol{\beta},\boldsymbol{c},\boldsymbol{u}} \left(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{K}\boldsymbol{c}-\boldsymbol{Z}\boldsymbol{u}\|^2+\lambda_c\boldsymbol{c}^T\boldsymbol{K}\boldsymbol{c}+\lambda_u\|\boldsymbol{u}\|^2\right) \tag{19}$$

where $\boldsymbol{X} = [1\ (\boldsymbol{x}_i^\ell)^T]_{1\le i\le n}$. Vector differential calculus, combined with some algebra, lead to the solutions:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}, \quad \widehat{\boldsymbol{c}} = \lambda_c^{-1}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ \text{and } \widehat{\boldsymbol{u}} &= \lambda_u^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \quad \text{where } \boldsymbol{V}\equiv\lambda_c^{-1}\boldsymbol{K}+\lambda_u^{-1}\boldsymbol{Z}\boldsymbol{Z}^T+\boldsymbol{I}. \end{aligned} \tag{20}$$

We now provide illustration of fits for this most general random intercept model for longitudinal data on spinal bone mineral density for 230 girls and young women (source: [2]). The subjects are categorized as belonging to one of the four ethnicity groups: Asian, Black, Hispanic and White. With double subscript notation, the model is

$$y_{ij} = [1\ (\boldsymbol{x}^\ell)^T]\,\boldsymbol{\beta}+U_i+c(x_{ij})+\varepsilon_{ij} \tag{21}$$

where the y_{ij} are spinal bone mineral measurements (g/cm^2), the $\boldsymbol{x}_i^\ell$ contain indicators for ethnicity and the x_{ij} are age measurements. The function $c(\cdot)$ indicates a curve corresponding to the kernel K_c. We used the Gaussian kernel $K_c(s,t)=\exp\{-0.05(s-t)^2\}$. The values of λ_c and λ_u were chosen using REML.

Figure 2 shows the fitted mean curves for each ethnicity group. The mean age effect is clearly non-linear and is estimated well by the Gaussian kernel.

Figure 3 shows the fitted curves for those subjects that had four spinal bone mineral density measurements. Note that the fitted random effects $\widehat{U}_i$ are required for this display. The fits are very good for about half of the subjects. Some lack-of-fit is apparent for the other half.

4.1.3. Extension to multivariate kernels

We briefly mention one last extension: the replacement of $c(x_i)$ by $c(\boldsymbol{x}_i)$ where the $\boldsymbol{x}_i\in\mathbb{R}^d$. This can be achieved by making K_c a d-variate kernel as opposed to the univariate kernels treated so far in this section. The relevant RKHS is now on $\mathbb{R}^{m+p+d}$, but the optimization problems (18) and (19) are only different in that $x_i\in\mathbb{R}$ is now $\boldsymbol{x}_i\in\mathbb{R}^d$ and $\boldsymbol{K}=[K_c(\boldsymbol{x}_i,\boldsymbol{x}_{i'})]_{1\le i,i'\le n}$ where the kernel K_c is on $\mathbb{R}^d\times\mathbb{R}^d$. Models of a similar type were recently considered by Liu, Lin and Ghosh [18].

4.2. Random intercept and slope model

Close inspection of Figure 1 shows that the parallel lines restriction imposed by the random intercept model is questionable. A more realistic model is one that allows each pig to have his/her own slope. This is achieved through the random intercept and slope model

$$y_{ij}=\beta_0+U_i+(\beta_1+V_i)x_{ij}+\varepsilon_{ij}, \quad 1\le j\le n_i, \quad 1\le i\le m \tag{22}$$

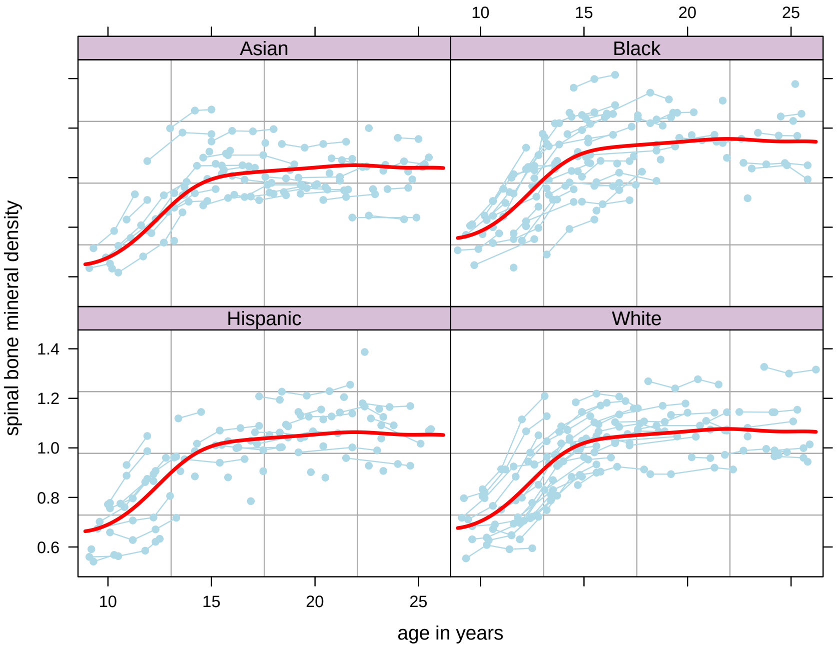

Fig 2. *Kernel-based fit to the spinal bone mineral density data.*

where, as with (8), $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$ are independent, while

$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix} \right), \quad \text{independently,} \tag{23}$$

allow for subject specific deviations in both intercept and slope from the mean line $\beta_0 + \beta_1 x$. Figure 4 shows an EBLUP fit of this model to the pig weights data, with the covariance matrix parameters estimated via REML.

The extension to random slopes involves replacement of $\boldsymbol{Z}$ and $\boldsymbol{G}$ by

$$\boldsymbol{Z} = \begin{bmatrix} 1 & x_{11} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{1m} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & x_{n1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & x_{nm} \end{bmatrix} \quad \text{and} \quad \boldsymbol{G} \equiv \underset{1 \le i \le m}{\text{blockdiag}} \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix} \tag{24}$$

in the BLUP equations (4).

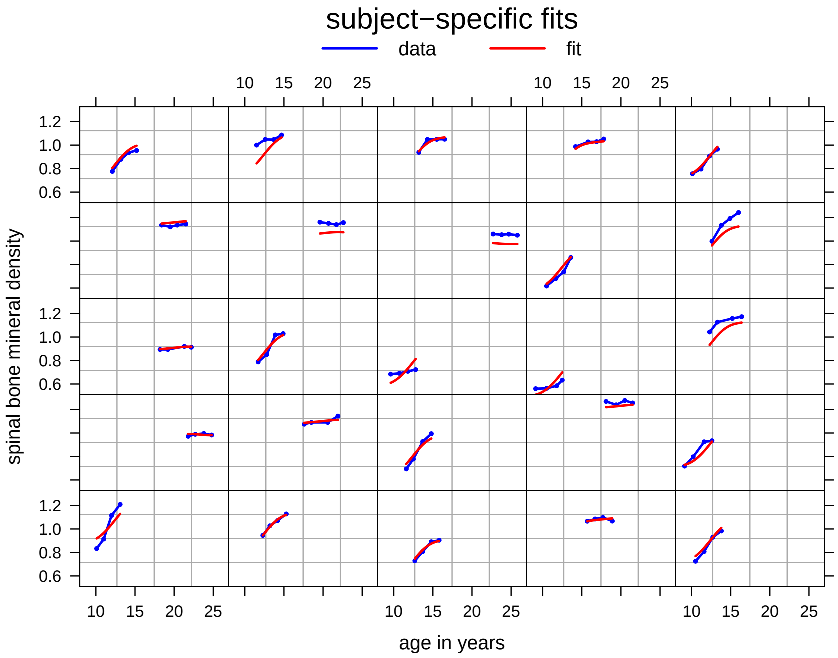

Fig 3. *Fitted subject-specific curves corresponding to model (21) for 25 randomly chosen subjects among those with 4 measurements each.*

We will now re-write (22) in a canonical form that is amenable to RKHS representation. It involves the singular value decomposition (or spectral decomposition) of the random effects covariance matrix:

$$\begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix} = \begin{bmatrix} \alpha & \sqrt{1-\alpha^2} \\ \sqrt{1-\alpha^2} & -\alpha \end{bmatrix} \begin{bmatrix} d_u & 0 \\ 0 & d_v \end{bmatrix} \begin{bmatrix} \alpha & \sqrt{1-\alpha^2} \\ \sqrt{1-\alpha^2} & -\alpha \end{bmatrix}$$

where the eigenvalues d_u and d_v are given by

$$\begin{aligned} d_u &= d_u(\sigma_u,\sigma_v,\rho) \equiv (\sigma_u^2+\sigma_v^2)/2 + \sqrt{(\sigma_u^2-\sigma_v^2)^2/4 + (\sigma_u\sigma_v\rho)^2}, \\ \text{and}\quad d_v &= d_v(\sigma_u,\sigma_v,\rho) \equiv (\sigma_u^2+\sigma_v^2)/2 - \sqrt{(\sigma_u^2-\sigma_v^2)^2/4 + (\sigma_u\sigma_v\rho)^2}. \end{aligned}$$

The first normalized eigenvector component α takes the form

$$\alpha = \alpha(\sigma_u,\sigma_v,\rho) \equiv \begin{cases} \sigma_u\sigma_v\rho/\sqrt{(\sigma_u\sigma_v\rho)^2 + (\sigma_u^2 - d_u)^2}, & \text{if } \rho \neq 0 \text{ or } \sigma_u \neq \sigma_v, \\ 1, & \text{otherwise.} \end{cases}$$

The matrix

$$\boldsymbol{\mathcal{U}} \equiv \begin{bmatrix} \alpha & \sqrt{1-\alpha^2} \\ \sqrt{1-\alpha^2} & -\alpha \end{bmatrix}$$

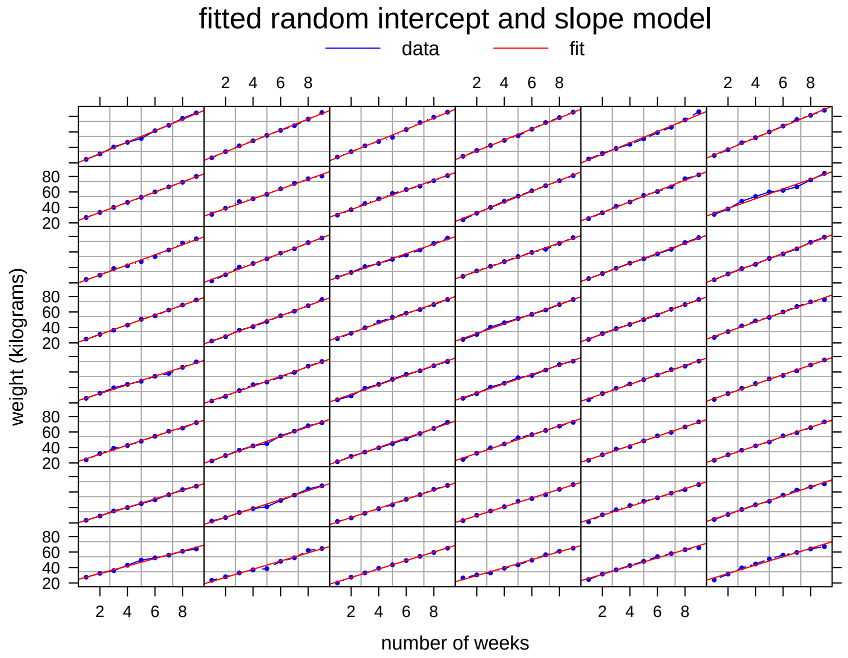

FIG 4. *The EBLUP-fitted lines to the pig-weights data for the simple linear random intercept and slope model. The panels are ordered according to the size of the 48 pigs.*

is orthogonal: $\boldsymbol{\mathcal{U}}\boldsymbol{\mathcal{U}}^T = \boldsymbol{\mathcal{U}}^T\boldsymbol{\mathcal{U}} = \boldsymbol{I}$. Even though $\boldsymbol{\mathcal{U}}$ is symmetric in this case, we will write $\boldsymbol{\mathcal{U}}^T$ to allow comparison with more general results. The random intercept and slope model (22) can be written as

$$\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 x_{ij} + [1\ x_{ij}]\boldsymbol{\mathcal{U}}\boldsymbol{\mathcal{U}}^T \begin{bmatrix} U_i \\ V_i \end{bmatrix} + \varepsilon_{ij} \\
&= \beta_0 + \beta_1 x_{ij} + \left(\boldsymbol{\mathcal{U}}^T \begin{bmatrix} 1 \\ x_{ij} \end{bmatrix}\right)^T \left(\boldsymbol{\mathcal{U}}^T \begin{bmatrix} U_i \\ V_i \end{bmatrix}\right) + \varepsilon_{ij} \\
&= \beta_0 + \beta_1 x_{ij} + [\widetilde{x}_{ij}^U\ \widetilde{x}_{ij}^V] \begin{bmatrix} \widetilde{U}_i \\ \widetilde{V}_i \end{bmatrix} + \varepsilon_{ij} \\
&= \beta_0 + \beta_1 x_{ij} + \widetilde{U}_i \widetilde{x}_{ij}^U + \widetilde{V}_i \widetilde{x}_{ij}^V + \varepsilon_{ij}
\end{aligned} \tag{25}$$

where, for $1 \le j \le n_i$, $1 \le i \le m$,

$$\widetilde{U}_i \equiv \alpha U_i + \sqrt{1-\alpha^2}\, V_i, \qquad \widetilde{V}_i \equiv \sqrt{1-\alpha^2}\, U_i - \alpha V_i,$$
$$\widetilde{x}_{ij}^u \equiv \alpha + \sqrt{1-\alpha^2}\, x_{ij} \quad \text{and} \quad \widetilde{x}_{ij}^v \equiv \sqrt{1-\alpha^2} - \alpha\, x_{ij}$$

are linear transformations of the random effects and predictors based on the $\boldsymbol{\mathcal{U}}$

matrix. Note that

$$\begin{bmatrix} \widetilde{U}_i \\ \widetilde{V}_i \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} d_u & 0 \\ 0 & d_v \end{bmatrix} \right), \quad \text{independently}, \quad 1 \le i \le m.$$

Suppose that the covariance parameters are replaced by estimates: $\widehat{\sigma}_u$, $\widehat{\sigma}_v$, $\widehat{\sigma}_\varepsilon$ $\widehat{\rho}$ and consider the EBLUPs

$$\widehat{\beta}_0 + \widehat{U}_i + (\beta_1 + \widehat{V}_i)x \tag{26}$$

based on (4). Let $\widehat{d}_u = d_u(\widehat{\sigma}_u, \widehat{\sigma}_v, \widehat{\rho})$, $\widehat{d}_v = d_v(\widehat{\sigma}_u, \widehat{\sigma}_v, \widehat{\rho})$ and $\widehat{\alpha} = \alpha(\widehat{\sigma}_u, \widehat{\sigma}_v, \widehat{\rho})$ be the estimates for the eigen-parameterization of the random effects covariance matrix.

We now describe RKHS representation of these EBLUPs. A first step is to switch from the double subscripting of longitudinal data analysis to single subscripting via (11). The single subscript version of the random intercept and slope model (22) is

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{m} (U_j + x_i V_j) Z_{ij} + \varepsilon_i, \quad 1 \le i \le n$$

where, as before, Z_{ij} is (i,j) entry of $\boldsymbol{Z}$ as given in (10). Form the RKHS of real-valued functions on $\mathbb{R}^{2m+1}$:

$$\mathcal{H}_K = \left\{ f : f(x, z_1^u, \ldots, z_m^u, z_1^v, \ldots, z_m^v) = \beta_0 + \beta_1 x + \sum_{j=1}^{m} (\widetilde{U}_j z_j^u + \widetilde{V}_j z_j^v) \right\} \tag{27}$$

with kernel

$$K(\boldsymbol{s}, \boldsymbol{t}) = K((s_1, \ldots, s_{2m+1}), (t_1, \ldots, t_{2m+1})) = 1 + \boldsymbol{s}^T \boldsymbol{t}. \tag{28}$$

Note that

$$K(\boldsymbol{s}, \boldsymbol{t}) = K_0(\boldsymbol{s}, \boldsymbol{t}) + K_u(\boldsymbol{s}, \boldsymbol{t}) + K_v(\boldsymbol{s}, \boldsymbol{t})$$

where

$$\begin{aligned} K_\beta(\boldsymbol{s}, \boldsymbol{t}) = 1 + s_1 t_1, \quad & K_u(\boldsymbol{s}, \boldsymbol{t}) = \textstyle\sum_{j=1}^{m} s_{1+j} t_{1+j} \\ \text{and} \quad & K_v(\boldsymbol{s}, \boldsymbol{t}) = \textstyle\sum_{j=1}^{m} s_{m+1+j} t_{m+1+j}. \end{aligned} \tag{29}$$

Let $\mathcal{H}_\beta$, $\mathcal{H}_u$ and $\mathcal{H}_v$ be the RKHSs generated by K_β, K_u and K_v, respectively, so that

$$\mathcal{H}_K = \mathcal{H}_\beta \oplus \mathcal{H}_u \oplus \mathcal{H}_v.$$

Theorem 2. *Let $(x_i, y_i, Z_{i1}, \ldots, Z_{im})$, $1 \le i \le n$, be a sequentially subscripted longitudinal data set as in Theorem 1. For $1 \le i \le n$ define $x_i^u = \widehat{\alpha} + \sqrt{1 - \widehat{\alpha}}\, x_i$ and $x_i^u = \sqrt{1 - \widehat{\alpha}}\, x_i - \widehat{\alpha}$ where $\widehat{\alpha} = \alpha(\widehat{\sigma}_u, \widehat{\sigma}_v, \widehat{\rho})$ is based on covariance parameter estimates appropriate for the random intercept and slope model (22) and (23). Consider the RKHS $\mathcal{H}_K$ defined by (27) and (28) and subspaces $\mathcal{H}_\beta$, $\mathcal{H}_u$ and*

$\mathcal{H}_v$ be generated by (29). Let P_u and P_v be, respectively, the projection operators onto $\mathcal{H}_u$ and $\mathcal{H}_v$. Then the solution to the RKHS optimization problem

$$\begin{aligned}\min_{f\in\mathcal{H}_K}\Big[\textstyle\sum_{i=1}^n\{y_i - f(x_i, Z_{i1}\widetilde{x}_i^u,\ldots,Z_{im}\widetilde{x}_i^u, Z_{i1}\widetilde{x}_i^v,\ldots,Z_{im}\widetilde{x}_i^v)\}^2 \\ +\lambda_u\|P_u f\|^2_{\mathcal{H}_K} + \lambda_v\|P_v f\|^2_{\mathcal{H}_K}\Big],\end{aligned} \tag{30}$$

with $\widehat{\lambda}_u = \widehat{\sigma}^2_\varepsilon/\widehat{d}_u$ and $\widehat{\lambda}_v = \sigma^2_\varepsilon/\widehat{d}_v$, corresponds to the EBLUPs (26). Explicitly, the solution to (30) is

$$\begin{aligned}
\widehat{f}(x,\widehat{\alpha}+\sqrt{1-\widehat{\alpha}^2}\,x,\mathbf{0}_{m-1},\sqrt{1-\widehat{\alpha}^2}-\widehat{\alpha}\,x,\mathbf{0}_{m-1}) &= \widehat{\beta}_0+\widehat{U}_1+(\widehat{\beta}_1+\widehat{V}_1)x,\\
\widehat{f}(x,0,\widehat{\alpha}+\sqrt{1-\widehat{\alpha}^2}\,x,\mathbf{0}_{m-1},\sqrt{1-\widehat{\alpha}^2}-\widehat{\alpha}\,x,\mathbf{0}_{m-2}) &= \widehat{\beta}_0+\widehat{U}_2+(\widehat{\beta}_1+\widehat{V}_2)x,\\
&\vdots\\
\widehat{f}(x,\mathbf{0}_{m-1},\widehat{\alpha}+\sqrt{1-\widehat{\alpha}^2}\,x,\mathbf{0}_{m-1},\sqrt{1-\widehat{\alpha}^2}-\widehat{\alpha}\,x) &= \widehat{\beta}_0+\widehat{U}_m+(\widehat{\beta}_1+\widehat{V}_m)x,
\end{aligned}$$

where $x\in\mathbb{R}$ and $\widehat{\beta}_0,\widehat{\beta}_1$ and the $\widehat{U}_i,\widehat{V}_i$ are given by (4) with $\boldsymbol{X}$ given by (10), $\boldsymbol{Z}$ given by (24), $\boldsymbol{R}=\widehat{\sigma}^2_\varepsilon\boldsymbol{I}$ and $\boldsymbol{G}$ given by (24) with $(\sigma_u,\sigma_v,\rho)=(\widehat{\sigma}_u,\widehat{\sigma}_v,\widehat{\rho})$, and $\mathbf{0}_r$ denotes the string 0,0,…,0 of length r.

Proof. The inner product for $\mathcal{H}_K$ (induced by K) is given by

$$\begin{aligned}\left\langle\beta_0+\beta_1x+\sum_{j=1}^m(\widetilde{U}_jz_j^u+\widetilde{V}_jz_j^v),\beta_0'+\beta_1'x+\sum_{j=1}^m(\widetilde{U}_j'z_j^u+\widetilde{V}_j'z_j^v)\right\rangle_{\mathcal{H}_K}\\
=\beta_0\beta_0'+\beta_1\beta_1'+\sum_{j=1}^m(\widetilde{U}_j\widetilde{U}_j'+\widetilde{V}_j\widetilde{V}_j').\end{aligned}$$

Let

$$f(x,z_1^u,\ldots,z_m^u,z_1^v,\ldots,z_m^v)=\beta_0+\beta_1x+\sum_{j=1}^m(\widetilde{U}_jz_j^u+\widetilde{V}_jz_j^v)$$

be a typical member of $\mathcal{H}_K$. Then, using a reversal of the argument presented at (25),

$$\begin{aligned}y_i - f(x_i, Z_{i1}\widetilde{x}_i^u,\ldots,Z_{im}\widetilde{x}_i^u, Z_{i1}\widetilde{x}_i^v,\ldots,Z_{im}\widetilde{x}_i^v)\\
&= y_i-(\boldsymbol{X\beta})_i-\sum_{j=1}^m Z_{ij}(\widetilde{U}_j\widetilde{x}_i^u+\widetilde{V}_j\widetilde{x}_i^v)\\
&= y_i-(\boldsymbol{X\beta})_i-\sum_{j=1}^m Z_{ij}(U_i+V_ix_i)\\
&= (\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})_i\end{aligned}$$

where $\boldsymbol{u} \equiv (U_1, V_1, U_2, V_2, \dots, U_m, V_m)^T$ and $\boldsymbol{Z}$, given by (24), are the random effects vector and corresponding design matrix for the random intercept and slope model. (In the definition of $\boldsymbol{Z}$ at (24), double subscripting is being used on the xs for clarity reasons.) Finally

$$
\begin{aligned}
(1/\widehat{d}_u)\|P_u f\|^2_{\mathcal{H}_K} + (1/d_v)\|P_v f\|^2_{\mathcal{H}_K} &= \sum_{j=1}^{m} (\widetilde{U}_j^2/\widehat{d}_u + \widetilde{V}_j^2/d_v) \\
&= \sum_{j=1}^{m} [\widetilde{U}_i \ \widetilde{V}_i] \begin{bmatrix} \widehat{d}_u & 0 \\ 0 & \widehat{d}_v \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{U}_i \\ \widetilde{V}_i \end{bmatrix} \\
&= \sum_{j=1}^{m} [\widetilde{U}_i \ \widetilde{V}_i] \widehat{\boldsymbol{\mathcal{U}}}^T \left(\widehat{\boldsymbol{\mathcal{U}}} \begin{bmatrix} \widehat{d}_u & 0 \\ 0 & \widehat{d}_v \end{bmatrix} \widehat{\boldsymbol{\mathcal{U}}}^T \right)^{-1} \widehat{\boldsymbol{\mathcal{U}}} \begin{bmatrix} \widetilde{U}_i \\ \widetilde{V}_i \end{bmatrix} \\
&= \sum_{j=1}^{m} [U_i \ V_i] \begin{bmatrix} \widehat{\sigma}_u^2 & \widehat{\rho}\widehat{\sigma}_u\widehat{\sigma}_v \\ \widehat{\rho}\widehat{\sigma}_u\widehat{\sigma}_v & \widehat{\sigma}_v^2 \end{bmatrix}^{-1} \begin{bmatrix} U_i \\ V_i \end{bmatrix} = \boldsymbol{u}^T \widehat{\boldsymbol{G}}^{-1} \boldsymbol{u}
\end{aligned}
$$

where

$$
\widehat{\boldsymbol{G}} \equiv \operatorname*{blockdiag}_{1 \le i \le m} \begin{bmatrix} \widehat{\sigma}_u^2 & \widehat{\rho}\widehat{\sigma}_u\widehat{\sigma}_v \\ \widehat{\rho}\widehat{\sigma}_u\widehat{\sigma}_v & \widehat{\sigma}_v^2 \end{bmatrix}
$$

is the estimated covariance matrix of $\boldsymbol{u}$. The criterion in the RKHS optimization problem (30) is therefore equivalent to $(1/\widehat{\sigma}_\varepsilon^2)\|\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}\|^2 + \boldsymbol{u}^T \widehat{\boldsymbol{G}}^{-1} \boldsymbol{u}$, the EBLUP criterion for the random intercept and slope model. □

4.2.1. Kernel-based extension to general mean curves

As in Section 4.1.1 we can extend the random intercept and slope model to allow for non-linear mean structure by introducing a kernel. Here we will consider the extension of (22):

$$
y_{ij} = \beta_0 + U_i + c(x_{ij}) + V_i x_{ij} + \varepsilon_{ij}, \quad 1 \le j \le n_i, \quad 1 \le i \le m. \tag{31}
$$

In this model $\beta_0 + c(\cdot)$ is the smooth overall function. Subject specific deviations in both intercept and slope are allowed. The relevant RKHS is

$$
\mathcal{H}_K = \mathcal{H}_0 \oplus \mathcal{H}_c \oplus \mathcal{H}_u \oplus \mathcal{H}_v
$$

where $\mathcal{H}_0$ and $\mathcal{H}_c$ are as defined in Section 4.1.1, whilst $\mathcal{H}_u$ and $\mathcal{H}_v$ carry the definitions ascribed to them in the lead up to Theorem 2. The RKHS minimization problem is similar to that given by (30) with the addition of a penalty term

$$
\lambda_c \|P_c f\|^2_{\mathcal{H}_K}.
$$

The minimization problem reduces to

$$
\min_{\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{u}} \left\{ \|\boldsymbol{y} - \boldsymbol{1}\beta_0 - \boldsymbol{Kc} - \boldsymbol{Zu}\|^2 + \lambda_c \boldsymbol{c}^T \boldsymbol{Kc} + \boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u} \right\}
$$

where $\boldsymbol{u} \equiv (U_1, V_1, U_2, V_2, \ldots, U_m, V_m)^T$, $\boldsymbol{Z} \equiv \text{blockdiag}_{1\le i\le m}([1\ x_{ij}]_{1\le j\le n_i})$ and $\boldsymbol{G} = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}$ for some 2×2 symmetric positive definite matrix $\boldsymbol{\Sigma}$. Here $\otimes$ denotes Kronecker product. The solution is

$$\begin{aligned} \widehat{\beta}_0 &= (\boldsymbol{1}^T\boldsymbol{V}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^T\boldsymbol{V}^{-1}\boldsymbol{y}, \quad \widehat{\boldsymbol{c}} = \lambda_c^{-1}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{1}\widehat{\beta}_0) \\ \text{and} \quad \widehat{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{1}\widehat{\beta}_0) \quad \text{where} \quad \boldsymbol{V} \equiv \lambda_c^{-1}\boldsymbol{K} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{I}. \end{aligned}$$

4.3. Extension to general random effects structure

The general form of the $\boldsymbol{X\beta} + \boldsymbol{Zu}$, $\boldsymbol{u} \sim (\boldsymbol{0}, \boldsymbol{G})$, structure for parametric longitudinal data analysis has

$$\boldsymbol{X} = \begin{bmatrix} 1 & \boldsymbol{X}_1^{\text{F}} \\ \vdots & \vdots \\ 1 & \boldsymbol{X}_m^{\text{F}} \end{bmatrix}, \quad \boldsymbol{Z} = \underset{1\le i\le m}{\text{blockdiag}}(\boldsymbol{X}_i^{\text{R}}), \quad \text{and} \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_m \end{bmatrix}$$

with

$$\boldsymbol{G} = \text{Cov}(\boldsymbol{u}) = \underset{1\le i\le m}{\text{blockdiag}}(\boldsymbol{\Sigma}) = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}.$$

Here $\boldsymbol{X}_i^{\text{F}}$ is an $n_i \times p$ matrix corresponding to the ith subject's fixed effects contribution $(\boldsymbol{X}_i^{\text{F}}\boldsymbol{\beta})$, $\boldsymbol{X}_i^{\text{R}}$ is an $n_i \times q$ matrix and $\boldsymbol{u}_i$ is a $q \times 1$ random effects vector corresponding the ith subject's contribution $(\boldsymbol{Z}_i^{\text{R}}\boldsymbol{u}_i)$ and $\boldsymbol{\Sigma}$ is an unstructured $q \times q$ covariance matrix satisfying $\text{Cov}(\boldsymbol{u}_i) = \boldsymbol{\Sigma}$, $1 \le i \le m$.

Let $\boldsymbol{\Sigma} = \boldsymbol{\mathcal{U}}\text{diag}(\boldsymbol{d})\boldsymbol{\mathcal{U}}^T$ be the spectral decomposition of $\boldsymbol{\Sigma}$, where

$$\boldsymbol{d} = (d_1, \ldots, d_q)$$

are the eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mathcal{U}}$ is a $q \times q$ orthogonal matrix of normalized eigenvectors. Set

$$\widetilde{\boldsymbol{Z}} = \boldsymbol{Z}\,(\boldsymbol{I}_m \otimes \boldsymbol{\mathcal{U}}) \quad \text{and} \quad \widetilde{\boldsymbol{u}} = (\boldsymbol{I}_m \otimes \boldsymbol{\mathcal{U}}^T)\boldsymbol{u}$$

so that the model has canonical form

$$\boldsymbol{X\beta} + \widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{u}}, \quad \widetilde{\boldsymbol{u}} \sim (\boldsymbol{0}, \text{diag}(\boldsymbol{d})).$$

The BLUPs for $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{u}}$ minimize

$$\begin{aligned} &(1/\sigma_\varepsilon^2)\|\boldsymbol{y} - \boldsymbol{X\beta} - \widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{u}}\|^2 + \widetilde{\boldsymbol{u}}^T\text{diag}(1/\boldsymbol{d})\widetilde{\boldsymbol{u}} \\ &\qquad = (1/\sigma_\varepsilon^2)\|\boldsymbol{y} - \boldsymbol{X\beta} - \widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{u}}\|^2 + \textstyle\sum_{k=1}^q (1/d_k)\|\widetilde{\widetilde{\boldsymbol{u}}}_k\|^2 \end{aligned} \qquad (32)$$

where

$$\widetilde{\widetilde{\boldsymbol{u}}}_1 \equiv \begin{bmatrix} \widetilde{\boldsymbol{u}}_{11} \\ \vdots \\ \widetilde{\boldsymbol{u}}_{m1} \end{bmatrix}, \quad \widetilde{\widetilde{\boldsymbol{u}}}_2 \equiv \begin{bmatrix} \widetilde{\boldsymbol{u}}_{12} \\ \vdots \\ \widetilde{\boldsymbol{u}}_{m2} \end{bmatrix}, \ \ldots\ , \quad \widetilde{\widetilde{\boldsymbol{u}}}_q \equiv \begin{bmatrix} \widetilde{\boldsymbol{u}}_{1q} \\ \vdots \\ \widetilde{\boldsymbol{u}}_{mq} \end{bmatrix}$$

and $\widetilde{\boldsymbol{u}}_{ij}$ is the jth entry of $\widetilde{\boldsymbol{u}}_i$, $1 \le i \le m$, $1 \le j \le q$.

Theorems 1 and 2 can be generalized to the situation where BLUP corresponds to the solution of a RKHS optimization problem. The relevant RKHS, $\mathcal{H}_K$, consists of real-valued functions on $\mathbb{R}^{p+mq}$ with kernel

$$K(\boldsymbol{s},\boldsymbol{t}) = K((s_1,\ldots,s_{p+mq}),(t_1,\ldots,t_{p+mq})) = 1 + \boldsymbol{s}^T\boldsymbol{t}.$$

Sub-spaces of interest are those generated by

$$K_F(\boldsymbol{s},\boldsymbol{t}) \equiv 1 + \sum_{j=1}^{p} s_j t_j \text{ and } \widetilde{\widetilde{K}}_k(\boldsymbol{s},\boldsymbol{t}) = \sum_{j=1}^{m} s_{p+(k-1)m+j}\,t_{p+(k-1)m+j},\ 1 \le k \le q.$$

We denote these by $\mathcal{H}_F, \widetilde{\widetilde{\mathcal{H}}}_1,\ldots,\widetilde{\widetilde{\mathcal{H}}}_q$. Then we have

$$\mathcal{H}_K = \mathcal{H}_F \oplus \widetilde{\widetilde{\mathcal{H}}}_1 \oplus \cdots \oplus \widetilde{\widetilde{\mathcal{H}}}_q.$$

Next, let

$$\widetilde{\widetilde{\boldsymbol{x}}} \equiv \begin{bmatrix} 1 & \boldsymbol{X}_1^{\mathrm{R}} \\ \vdots & \vdots \\ 1 & \boldsymbol{X}_m^{\mathrm{R}} \end{bmatrix} \boldsymbol{\mathcal{U}},$$

$\boldsymbol{Z}_i = (Z_{i1},\ldots,Z_{im})$ be the ith row of $\boldsymbol{Z}$ and $\widetilde{\widetilde{\boldsymbol{x}}}_i$ be the ith row of $\widetilde{\widetilde{\boldsymbol{x}}}$. Then the BLUPs given by (32) correspond to the RKHS optimization problem

$$\min_{f\in\mathcal{H}_K}\left[\sum_{i=1}^{n}\{y_i - f(\widetilde{\widetilde{\boldsymbol{x}}}_i \otimes \boldsymbol{Z}_i)\}^2 + \lambda_1\|\widetilde{\widetilde{P}}_1 f\|^2_{\mathcal{H}_K} + \ldots + \lambda_q\|\widetilde{\widetilde{P}}_q f\|^2_{\mathcal{H}_K}\right]$$

where, for $1 \le k \le q$, $\lambda_k \equiv \sigma_\varepsilon^2/d_k$ and $\widetilde{\widetilde{P}}_k$ is the projection operator onto $\widetilde{\widetilde{\mathcal{H}}}_k$.

4.4. Correlated errors

Each of the longitudinal models considered so far have

$$\boldsymbol{R} = \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2\,\boldsymbol{I}.$$

However, in longitudinal data analysis it is common to allow more general structure in the $\boldsymbol{R}$ matrix. An example is the random intercept model with first-order autoregressive (AR(1)) errors:

$$y_{ij} = \beta_0 + U_i + \beta_1\,x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} = \rho\varepsilon_{ij} + \xi_{ij},$$

for $1 \le i \le m$, $1 \le j \le n_i$, where $|\rho| < 1$ and the $\xi_{ij} \sim (0,\sigma_\xi^2)$ are independent. The $\boldsymbol{R}$ matrix in this case is

$$\boldsymbol{R} = \sigma_\xi^2 \underset{1\le i\le m}{\mathrm{blockdiag}}\begin{bmatrix} 1 & \rho & \cdots & \rho^{n_i-1} \\ \rho & 1 & \cdots & \rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \cdots & 1 \end{bmatrix}.$$

Longitudinal data analysis models such as these do not fit as comfortably into the RKHS framework. However, if the first term of the BLUP criterion (4) is written as

$$(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}) = \|\boldsymbol{R}^{-1/2}\boldsymbol{y}-(\boldsymbol{R}^{-1/2}\boldsymbol{X})\boldsymbol{\beta}-(\boldsymbol{R}^{-1/2}\boldsymbol{Z})\boldsymbol{u}\|^2$$

then it is apparent that the RKHS theory with $\mathcal{L}(s,t)=(s-t)^2$ applies provided we work with the transformed data vector $\boldsymbol{y_R} \equiv \boldsymbol{R}^{-1/2}\boldsymbol{y}$ and corresponding transformation of the predictor structure.

4.5. Alternative loss functions

So far in this section we have only considered squared error loss $\mathcal{L}(s,t)=(s-t)^2$. However, the connections between longitudinal data analysis and kernel machines remain if other continuous response loss functions are used instead. For example, to counteract the influence of outliers in the response data it is common to work with a different loss such as absolute error $\mathcal{L}(s,t)=|s-t|$ or those arising in M-estimation (e.g. [14]):

$$\mathcal{L}(s,t;\delta)=\begin{cases}(s-t)^2, & |s-t|\le\delta\\ 2\delta|s-t|-\delta^2, & |s-t|>\delta\end{cases}$$

where $\delta>0$ is the 'bending' parameter. Another class robust of loss functions is that which arises from modelling the responses as having a *t*-distribution with ν degrees of freedom and scale parameter $\sigma>0$:

$$\mathcal{L}(s,t;\sigma,\nu)=\log\left\{1+\nu^{-1}\left(\frac{s-t}{\sigma}\right)^2\right\}.$$

Yet another alternative loss function is $\mathcal{L}(s,t;\varepsilon)=(|s-t|-\varepsilon)_+$, for a fixed $\varepsilon>0$. It is known as *ε-insensitive loss* and ignores errors of size less than ε. Finally, we mention the possibility of non-convex loss functions, such as those described in Shen, Tseng, Zhang and Wong [30].

5. Generalized response extension

Many longitudinal studies have a non-continuous response; such as count or binary variable. In such circumstances the linear mixed model (1) is not appropriate and alternative approaches are required. The most common involve generalized linear mixed models (GLMM) and generalized estimating equations (GEE). In this section we describe explicit connections between kernel machines and the popular penalized quasi-likelihood (PQL) methodology for fitting GLMMs to generalized response longitudinal data. The main message is that the RKHSs developed in Section 4 for longitudinal data analysis all apply to the generalized response situation. Only the loss functions require modification.

To keep the notation simple, we will work with GLMMs confined to the canonical one-parameter exponential family framework:

$$f(\boldsymbol{y}|\boldsymbol{u}) = \exp\{\boldsymbol{y}^T(\boldsymbol{X\beta}+\boldsymbol{Zu}) - \mathbf{1}^T b\,(\boldsymbol{X\beta}+\boldsymbol{Zu}) + \mathbf{1}^T c(\boldsymbol{y})\}, \quad \boldsymbol{u} \sim (\mathbf{0}, \boldsymbol{G}) \tag{33}$$

where $f(\boldsymbol{y}|\boldsymbol{u})$ denotes the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ and b and c depend upon the family member. The most common examples are Bernoulli, with $b(x) = \log(1+e^x)$, $c(x) = 0$, and Poisson with $b(x) = e^x$, $c(x) = -\log(x!)$. The matrices in the linear predictor $\boldsymbol{\eta} \equiv \boldsymbol{X\beta}+\boldsymbol{Zu}$, as well as $\boldsymbol{G}$, have definition and structure identical to those in the continuous response situation described in Sections 2 and 4. The simplest example is the generalized response random intercept model

$$\begin{aligned} f(y_{ij}|U_1,\ldots,U_m) = \exp\Big[\textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} \Big\{ & y_{ij}(\beta_0+\beta_1\,x_{ij}+U_i) \\ & -b\,(\beta_0+\beta_1\,x_{ij}+U_i)\Big\}\Big] \end{aligned} \tag{34}$$

with U_i are independent $(0,\sigma_u^2)$, $1 \le i \le m$, which corresponds to (33) with $\boldsymbol{X}$ and $\boldsymbol{Z}$ as in (10) and $\boldsymbol{G} = \sigma_u^2\boldsymbol{I}$.

A common approach to fitting GLMMs is maximum likelihood for $(\boldsymbol{\beta},\boldsymbol{G})$ and best prediction for $\boldsymbol{u}$ under the normality assumption $\boldsymbol{u} \sim N(\mathbf{0},\boldsymbol{G})$. However this requires numerical integration techniques and, especially if the integrals are multi-dimensional, approximations are used instead. The most common of these is PQL (e.g. [3]). However, we will not treat quasi-likelihoods here, so the label *penalized likelihood (PL)* is appropriate. For (33) with $\boldsymbol{u} \sim N(\mathbf{0},\boldsymbol{G})$ and $\boldsymbol{G}$ known this involves maximization of the penalized likelihood

$$\exp\{\boldsymbol{y}^T(\boldsymbol{X\beta}+\boldsymbol{Zu}) - \mathbf{1}^T b\,(\boldsymbol{X\beta}+\boldsymbol{Zu}) - \tfrac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\} \tag{35}$$

to obtain the estimates $\widehat{\boldsymbol{\beta}}_{\text{PL}}$ and $\widehat{\boldsymbol{u}}_{\text{PL}}$.

We now show that the penalized likelihood (35) can be treated as an RKHS optimization problem. Hence, obtaining $\widehat{\boldsymbol{\beta}}_{\text{PL}}$ and $\widehat{\boldsymbol{u}}_{\text{PL}}$ for a given $\boldsymbol{G}$ involves a particular kernel machine. Again, with simplicity in mind, we give the full explanation for the random intercept model (34). The general case follows via the linear algebraic arguments and structures given in Sections 4.2 and 4.3.

Re-subscript the (x_{ij}, y_{ij}) sequentially (as in Section 4) and, as before, let Z_{ij} be the (i,j) entry of the matrix $\boldsymbol{Z}$ defined at (10). Then (34) is

$$f(y_i|U_1,\ldots,U_m) = \exp\Big[\sum_{i=1}^n \Big\{y_i\big(\beta_0+\beta_1\,x_i+\sum_{j=1}^m Z_{ij}U_j\big) - b\big(\beta_0+\beta_1\,x_i+\sum_{j=1}^m Z_{ij}U_j\big)\Big\}\Big].$$

Let $\mathcal{H}_K$, K and $\mathcal{H}_\beta$ be defined by (12), (13) and (14) respectively. Then penalized likelihood estimation of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ is equivalent to the RKHS optimization problem

$$\min_{f\in\mathcal{H}_K}\Big[\sum_{i=1}^n \mathcal{L}(y_i, f(x_i, Z_{i1},\ldots,Z_{im})) + \lambda\|P_u f\|^2_{\mathcal{H}_K}\Big] \tag{36}$$

where P_u is the projection operator onto $\mathcal{H}_u = \mathcal{H}_\beta^\perp$, $\lambda = 1/\sigma_u^2$ and the loss function is given by $\mathcal{L}(s,t) = -2\{st - b(t)\}$. For example,

$$\mathcal{L}(s,t) = \begin{cases} -2\{st + \log(1+e^t)\}, & \text{in the Bernoulli case,} \\ -2(st + e^t), & \text{in the Poisson case.} \end{cases}$$

If $\widehat{f}$ is the solution to (36) then

$$\begin{aligned} \widehat{f}(x,1,0,\ldots,0) &= \widehat{\beta}_0 + \widehat{\beta}_1\,x + \widehat{U}_1, \\ \widehat{f}(x,0,1,\ldots,0) &= \widehat{\beta}_0 + \widehat{\beta}_1\,x + \widehat{U}_2, \\ &\vdots \\ \text{and} \quad \widehat{f}(x,0,0,\ldots,1) &= \widehat{\beta}_0 + \widehat{\beta}_1\,x + \widehat{U}_m, \end{aligned}$$

where $(\widehat{\beta}_0, \widehat{\beta}_1) = \widehat{\boldsymbol{\beta}}_{\text{PL}}$ and $(\widehat{U}_1,\ldots,\widehat{U}_m) = \widehat{\boldsymbol{u}}_{\text{PL}}$.

5.1. Kernel extension

Let $\mathcal{H}_K = \mathcal{H}_\beta \oplus \mathcal{H}_c \oplus \mathcal{H}_u$ be the RKHS defined in Section 4.1.1. Recall that $\mathcal{H}_\beta$ handles linear predictors ('fixed effects'), $\mathcal{H}_c$ handles non-linear effects of a predictor x and $\mathcal{H}_u$ handles random intercepts. The same structure can be used for the general response situation. For example, if y_i is binary then the appropriate minimization problem is

$$\min_{f\in\mathcal{H}_K}\left(\sum_{i=1}^n[-2\{y_i - f(x_i,Z_{i1},\ldots,Z_{im}) + \log(1+e^{f(x_i,Z_{i1},\ldots,Z_{im})})\}] \right. \\ \left. + \lambda_c\|P_c f\|^2_{\mathcal{H}_K} + \lambda_u\|P_u f\|^2_{\mathcal{H}_K}\right).$$

By arguments similar to those given in Section 2, this reduces to the matrix algebraic problem

$$\max_{\boldsymbol{\beta},\boldsymbol{c},\boldsymbol{u}}\{\boldsymbol{y}^T(\boldsymbol{X\beta}+\boldsymbol{Kc}+\boldsymbol{Zu}) - \boldsymbol{1}^T\log(\boldsymbol{1}+e^{\boldsymbol{X\beta}+\boldsymbol{Kc}+\boldsymbol{Zu}}) - \tfrac{1}{2}\lambda_c\boldsymbol{c}^T\boldsymbol{Kc} - \tfrac{1}{2}\lambda_u\|\boldsymbol{u}\|^2\}$$

which may be solved via Newton-Raphson iteration.

5.2. Alternative loss functions

Other loss functions, appropriate for the type of generalized response at hand, may be considered. For example, in the binary response case the Bernoulli log-likelihood loss could be replaced by *hinge loss*:

$$\mathcal{L}(s,t) = \{1-(2s-1)t\}_+,$$

which corresponds to support vector machine classification (e.g. [4]). Alternative *large margin* loss functions could also be considered, such as

$$\mathcal{L}(s,t)=[\{1-(2s-1)t\}_+]^q,\quad q>1\quad\text{and}\quad \mathcal{L}(s,t)=\{\rho-(2s-1)t\}_+,\quad \rho>0.$$

The definition of the class of large margin loss functions, and a fuller list of examples, is given in Section 2 of Wang and Shen [39].

We performed a small comparison of Bernoulli log-likelihood loss and hinge loss for a binary response longitudinal data set. The data involves longitudinal measurements on Indonesian children (source: [6]). The response variable is an indicator of presence of respiratory infection. The comparison between the two loss functions is cleaner if the response is coded as ±1, so we let

$$y_{ij}=\begin{cases}1 & \text{if child } i \text{ has respiratory infection at time of measurement } j,\\ -1 & \text{otherwise}\end{cases}$$

and x_{ij} denote the age, in years, corresponding to y_{ij}. Then, with (x_i,y_i), $1\leq i\leq n$, denoting the sequentially subscripted (x_{ij},y_{ij}) we considered two versions of

$$\min_{f\in\mathcal{H}_K}\left\{\sum_{i=1}^n \mathcal{L}(y_i,f(x_i,Z_{i1},\ldots,Z_{im}))+\lambda_c\|P_cf\|^2_{\mathcal{H}_K}+\lambda_u\|P_uf\|^2_{\mathcal{H}_K}\right\}\tag{37}$$

where $\mathcal{H}_K=\mathcal{H}_\beta\oplus\mathcal{H}_c\oplus\mathcal{H}_u$ is the RKHS defined in Section 4.1.1. The first version has

$$\mathcal{L}(s,t)=\log(1+e^{-st}),$$

corresponding to Bernoulli log-likelihood, while the second is hinge loss:

$$\mathcal{L}(s,t)=(1-st)_+.\tag{38}$$

For $\mathcal{H}_c$ we used a low-rank kernel

$$K_c(s,t)=\sum_{j=1}^k z_j(s)z_j(t)$$

based on a set of k canonical O'Sullivan spline basis functions $\{z_j(\cdot):1\leq j\leq k\}$ and corresponding to equation (6) of Wand and Ormerod [38]. In the case of hinge loss, such a kernel gives rise to a low-rank quadratic programming problem. This enabled us to use the methodology described in Ormerod, Wand and Koch [23] and the corresponding `R` package `LowRankQP` ([22]).

In the original data, there are many more values with $y_{ij}=-1$ than with $y_{ij}=1$ and a weighted version of hinge loss is appropriate. So that we could use ordinary hinge loss (38) for illustration we worked with a sub-sample of the $y_{ij}=-1$ data so the sample sizes for each type are equal (107 each). The regularization parameter for the random subject effect was obtained via penalized quasi-likelihood to be $\lambda_u=3.6$. We then varied the regularization parameter for the K_c component over the values $\lambda_c \in \{10, 1, 0.1, 0.01\}$. The fitted functions:

$$\widehat{f}_c(x; \lambda_c) = \sum_{i=1}^{n} \widehat{c}_i K_c(x, x_i)$$

are shown in Figure 5. Note that the $\widehat{c}_i$ depend on λ_c.

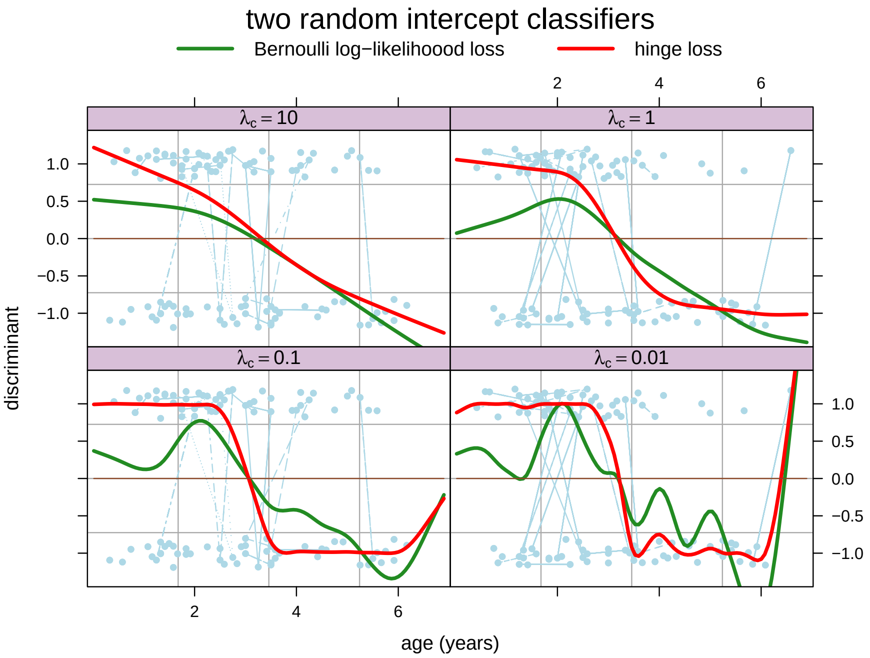

FIG 5. *Results of fitting (37) for both Bernoulli log-likelihood loss and hinge loss, with 4 different values of the regularization parameter λ_c corresponding to the spline kernel machine age effect. If viewed as a classification problem then the curves correspond to discriminants. The longitudinal data are jittered to enhance visualization.*

Even though the Indonesian children respiratory study was concerned with determination of risk factors, rather than classification, it is useful to suppose that the latter is the case when viewing Figure 5. Under this scenario, the $\widehat{f}_c(\cdot; \lambda_c)$ are *discriminants* for classification of having respiratory infection or not based on age. Hinge loss is seen to be less wiggly and more decisive on either side of the classification boundaries.

Both discriminants possess the rare property of taking account of the longitudinal nature of the data, through the presence of the $\lambda_u \|P_u f\|^2_{\mathcal{H}_K}$ term in the fitting process. While the classification rules themselves are minimally affected by this additional regularization, precision estimates are likely to change. Since precision estimates require additional probabilistic structure we do not do such comparison here.

6. Discussion

In this article we have shown that two ostensibly different areas of research – longitudinal data analysis and kernel machines – are, in fact, very similar in their underlying mathematics. It is anticipated that the explicit connections that have been established here will facilitate a more fluid exchange of ideas between the two fields. For longitudinal data analysis, there is the possibility of using kernel machines to better deal with non-linearity and to develop improved classification procedures. From the kernel machine perspective, we envisage kernel methodology that is tailored to longitudinal data models and accounts for complications such as within-unit correlation.

References

[1] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, **68**, 337–404. MR0051437

[2] Bachrach, L.K., Hastie, T., Wang, M.-C., Narasimhan, B. and Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth. A longitudinal study. *Journal of Clinical Endocrinology and Metabolism*, **84**, 4702–12.

[3] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

[4] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge: Cambridge University Press.

[5] Cressie, N. (1993). *Statistics for Spatial Data.* New York: John Wiley & Sons. MR1239641

[6] Diggle, P.J., Heagerty, P., Liang, K.-L. and Zeger, S. (2002). *Analysis of Longitudinal Data (Second Edition)*. Oxford: Oxford University Press. MR2049007

[7] Evgeniou, T., Pontil, M. and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, **13**, 1–50. MR1759187

[8] Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Eds.) (2008). *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods.* Boca Raton, Florida: Chapman & Hall/CRC.

[9] Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis.* New York: John Wiley & Sons. MR2063401

[10] Gianola, D., Fernando, R.L. and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, **173**, 1761–1776.

[11] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman & Hall. MR1270012

[12] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–338. MR0451550

[13] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall. MR1082147

[14] Huber P. (1981). *Robust Statistics.* Chichester: John Wiley & Sons. MR0606374

[15] James, G.M. and Hastie, T.J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B*, **63**, 533–550. MR1858401

[16] Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Its Application*, **33**, 82–95. MR0290013

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

[18] Liu, D., Lin, X. and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, **63**, 1079–1088. MR2414585

[19] McCulloch, C.E., Searle, S.R. and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models, Second Edition*. New York: John Wiley & Sons. MR2431553

[20] Moguerza, J.M. and Muñoz, A. (2006). Support vector machines with applications (with discussion). *Statistical Science*, **21**, 322–362. MR2339130

[21] Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, **32**, 223–240. MR2188671

[22] Ormerod, J.T. and Wand, M.P. (2006). LowRankQP 1.0. R package. `http://cran.r-project.org`

[23] Ormerod, J.T., Wand, M.P. and Koch, I. (2008). Penalised spline support vector classifiers: computational issues. *Computational Statistics*, **23**, 623–641.

[24] Pearce, N.D. and Wand, M.P. (2006). Penalized splines and reproducing kernel methods. *The American Statistician*, **60**, 233–240. MR2246756

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

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

[27] Rudin, W. (1991). *Functional Analysis, Second Edition*. New York: McGraw Hill. MR1157815

[28] Ruppert, D., Wand, M. P. and Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press. MR1998720

[29] Schölkopf, B. and Smola, A.J. (2002). *Learning with Kernels.* Cambridge USA: MIT Press.

[30] Shen, X., Tseng, G.C., Zhang, X. and Wong, W. (2003). On ψ-learning. *Journal of the American Statistical Association*, **98**, 724–734. MR2011686

[31] Simmons, G.F. (1983). *Introduction to Topology and Modern Analysis.* Melbourne USA: Krieger. MR0695310

[32] Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* New York: Springer. MR1697409

[33] Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B. and Vandewall, J. (2002) *Least squares support vector machines.* Singapore: World Scientific Publishing Company.

[34] Tarantola, A. (2005). *Inverse Problem Theory.* Philadelphia: Society of Industrial and Applied Mathematics. MR2130010

[35] Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag. MR1880596

[36] Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: Society of Industrial and Applied Mathematics. MR1045442

[37] Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In *Advances in Kernel Methods : Support Vector Learning* (eds. B. Scholkopf, C. Burges and A. Smola) pp. 69–88. Cambridge USA: MIT Press.

[38] Wand, M.P. and Ormerod, J.T. (2008). On O'Sullivan penalised splines and semiparametric regression. *Australian and New Zealand Journal of Statistics*, **50**, 179–198. MR2431193

[39] Wang, J. and Shen, X. (2007). Large margin semi-supervised learning. *Journal of Machine Learning Research*, **8**, 1867–1891. MR2353822

[40] Zhu, J. and Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, **14**, 185–205. MR2137897