From the SelectedWorks of Philip T. Reiss

January 2010

Fast Function-on-Scalar Regression with Penalized Basis Expansions

Contact Author

Start Your Own SelectedWorks

Notify Me of New Work

Available at: http://works.bepress.com/phil_reiss/16

The International Journal of Biostatistics

Volume 6, *Issue* 1 2010 *Article* 28

Fast Function-on-Scalar Regression with Penalized Basis Expansions

Philip T. Reiss* Lei Huang† Maarten Mennes‡

*New York University and Nathan S. Kline Institute for Psychiatric Research, phil.reiss@nyumc.org

†New York University, huangracer@gmail.com

‡New York University, mennes.maarten@gmail.com

Copyright ©2010 The Berkeley Electronic Press. All rights reserved.

Fast Function-on-Scalar Regression with Penalized Basis Expansions*

Philip T. Reiss, Lei Huang, and Maarten Mennes

Abstract

Regression models for functional responses and scalar predictors are often fitted by means of basis functions, with quadratic roughness penalties applied to avoid overfitting. The fitting approach described by Ramsay and Silverman in the 1990s amounts to a penalized ordinary least squares (P-OLS) estimator of the coefficient functions. We recast this estimator as a generalized ridge regression estimator, and present a penalized generalized least squares (P-GLS) alternative. We describe algorithms by which both estimators can be implemented, with automatic selection of optimal smoothing parameters, in a more computationally efficient manner than has heretofore been available. We discuss pointwise confidence intervals for the coefficient functions, simultaneous inference by permutation tests, and model selection, including a novel notion of pointwise model selection. P-OLS and P-GLS are compared in a simulation study. Our methods are illustrated with an analysis of age effects in a functional magnetic resonance imaging data set, as well as a reanalysis of a now-classic Canadian weather data set. An R package implementing the methods is publicly available.

KEYWORDS: cross-validation, functional data analysis, functional connectivity, functional linear model, smoothing parameters, varying-coefficient model

*The authors are grateful to the referees for very helpful feedback, to Mike Milham, Eva Petkova, Thad Tarpey and Simon Wood for highly informative discussions, and to Giles Hooker for advice on using the R package fda. The first author's research is supported in part by National Science Foundation grant DMS-0907017.

1 Introduction

A broad array of regression models have been proposed for settings in which the observations take the form of entire curves or functions. It has become conventional to distinguish among three basic situations (Ramsay and Silverman, 2005; Chiou et al., 2004):

1. both the responses and the predictors are functions—what might be termed function-on-function regression;

2. the responses are scalars and the predictors are functions (scalar-on-function regression);

3. the responses are functions and the predictors are scalars (function-on-scalar regression).

The terms "functional regression" and "functional linear model" have been applied to each of these scenarios, so it is important to be clear as to which of the three is under discussion. This paper describes methodological and computational advances for the third scenario, function-on-scalar regression.

Our starting point is the formulation of function-on-scalar regression in Section 13.4 of Ramsay and Silverman (2005; hereafter RS), which we begin by briefly recapitulating. The basic model is

$$y(t) = Z\beta(t) + \varepsilon(t). \tag{1}$$

Here the argument t ranges over some finite interval $T \subset \mathscr{R}$; $y(t)$ is an N-dimensional "vector of functional responses," i.e., a vector-valued function with values in $\mathscr{R}^N$; Z is an $N \times q$ design matrix; $\beta(t) = [\beta_1(t), \ldots, \beta_q(t)]^T$ is the vector of functional effects that we seek to estimate; and $\varepsilon(t)$ is a vector of error functions $\varepsilon_1(t), \ldots, \varepsilon_N(t)$, assumed to be drawn from a stochastic process with expectation zero at each t. The actual response data may come in one of two forms. We may be given a raw response matrix

$$Y = \left[y_i(t_j)\right]_{1 \le i \le N, 1 \le j \le n} \tag{2}$$

derived by sampling the N response curves at points $t_1, \ldots, t_n$. Alternatively, if the outcomes lie in the span of a set of basis functions $\theta_1, \ldots, \theta_K$, they can be specified by an $N \times K$ matrix C of basis coefficients such that $y(t) = C\theta(t)$, where $\theta(t) = [\theta_1(t), \ldots, \theta_K(t)]^T$.

Whether the responses are given in raw form or as basis coefficients, we posit that the coefficient functions $\beta_1, \ldots, \beta_q$ lie in the span of $\theta_1, \ldots, \theta_K$. The problem is

thereby reduced to estimating $B = (b_1 \dots b_q)^T$, the $q \times K$ matrix of basis coefficients determining the coefficient functions via

$$\beta_k(t) = b_k^T \theta(t) \text{ for } k = 1, \dots, q. \tag{3}$$

A key feature of this basis function approach is the use of quadratic roughness penalties to avoid overfitting in the estimation of B.

In RS's formulation, the response and coefficient functions may be expanded in two different bases. For our purposes, as explained in Appendix A, it suffices to assume the same basis is used for both.

The above framework is probably the most common "entry point" for biostatisticians and other data analysts interested in function-on-scalar regression, due to the relative simplicity and accessibility of the basis function-roughness penalty approach, the status of RS as a foundational text of functional data analysis, and the wide dissemination of associated software for R and Matlab (Ramsay et al., 2009). Still, this basic model does not seem to have been incorporated into many data analysts' toolkits, in spite of its wide and growing range of potential applications. We attribute this to several factors:

1. Although the roughness penalty approach is familiar to readers of RS, the function-on-scalar regression problem entails somewhat involved matrix algebra, so that the route to obtaining the optimal B may seem more opaque than for other problems solved by roughness penalization.

2. The critical choice of the optimal degree of smoothing requires cross-validation, which is very time-consuming (Ramsay et al., 2009, p. 154). Given the need for rapid implementation of modern data analyses, this may render function-on-scalar regression infeasible in many applications.

3. The basis coefficient matrix C of the responses, as opposed to the raw responses Y, is emphasized in RS's treatment and is assumed in the associated software (Ramsay et al., 2009), whereas other authors have generally worked with the raw data.[1] This and perhaps other differences in formulation have hindered cross-fertilization between RS's work and related models, some of which are not explicitly "functional."[2]

[1]This difference is at least partly driven by disparate applications. RS are primarily interested in densely sampled data for which reduction to basis coefficients may be a practical necessity. Related work often deals with longitudinal data that is sparsely and/or irregularly sampled, possibly with significant measurement error; here raw responses are more appropriate. The methods of this paper are developed with the former class of applications in mind.

[2]In particular, RS (p. 259) note that functional-response models are related to varying-coefficient models, and suggest that methods developed for the latter may be useful for the former. Indeed, model (1) can be viewed as a repeated-measures varying-coefficient model.

DOI: 10.2202/1557-4679.1246

Our reconsideration of model (1) addresses each of these issues:

1. By a subtle modification of RS's development, we reduce the objective function to the familiar generalized ridge regression form appearing in the roughness penalty approach to nonparametric regression.

2. This reformulation motivates a computational shortcut that enables cross-validation to be performed much faster than with existing software (Section 8.1 discusses a classic data set for which speed is improved by two orders of magnitude).

3. We show how our framework casts the RS estimate as a penalized ordinary least squares (P-OLS) solution. This motivates a penalized generalized least squares (P-GLS) model in basis coefficient space, extending RS's model and bridging the gap between it and recent work in the varying-coefficient model literature.

Our work was motivated by a functional magnetic resonance imaging (fMRI) study investigating *functional connectivity*, or temporal correlation among signals in different brain locations. By a procedure outlined in Section 8.2, one can express connectivity between locations as a function of the distance between them. Figure 1 displays functions of this kind for each of 59 participants. There is neuroscientific evidence that the relationship between distance and connectivity may vary with age, and the figure suggests that there may indeed be systematic differences between the older and younger participants, in particular at short distances. Our application of function-on-scalar regression to model the effect of age on the connectivity functions will be presented in Section 8.2.

The contributions of this paper can be summarized as follows. First, we describe computationally efficient cross-validation for P-OLS, implemented in the new `refund` package for R (R Development Core Team, 2010). Second, we extend RS's model to P-GLS, with fast automatic selection of multiple smoothing parameters. Third, we present a generalized ridge regression framework for function-on-scalar regression that encompasses both P-OLS and P-GLS. Fourth, we introduce a novel notion of pointwise model selection, which may be more useful than traditional "overall" model selection in many applications. Fifth, we describe two distinct interval estimation approaches appropriate for P-OLS and for P-GLS respectively. Sixth, we compare P-OLS and several variants of P-GLS via simulations and with reference to the neuroimaging data set mentioned above.

Following some brief remarks on related models in Section 2, we begin our main development in Section 3, in which the P-OLS estimator of RS is derived as a generalized ridge estimator. Our P-GLS extension of RS's model is described

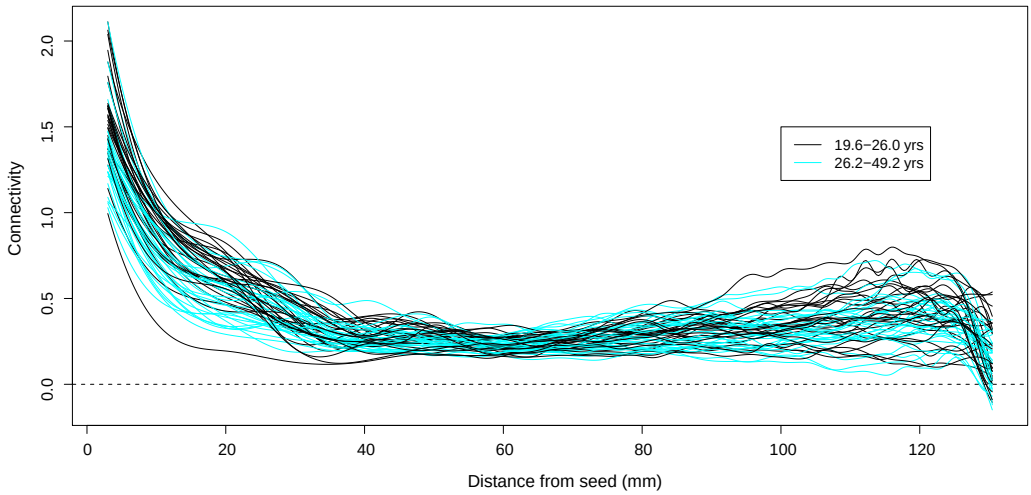

Figure 1: Functional connectivity between brain regions as a function of distance between the regions, for each of 59 individuals. The darker curves represent participants whose age is below the sample median.

in Section 4. Section 5 shows how our methods achieve computationally efficient automatic smoothing parameter selection. Section 6 discusses confidence intervals, hypothesis testing, and model selection. Comparative simulations and real data analyses appear in Sections 7 and 8, and Section 9 offers concluding remarks.

2 Some related methods

Before proceeding we shall offer some remarks on related models. In practical terms (at least when working with the raw responses), model (1) is equivalent to a varying-coefficient model (Hastie and Tibshirani, 1993) with repeated measurements. In conceptual terms, however, the functional data analysis (FDA) viewpoint casts the model somewhat differently than the traditional repeated measures or longitudinal data viewpoint. If our data come from N subjects, each observed at times $t_1, \ldots, t_n$, then the longitudinal data paradigm views the values $y_{i1}, \ldots, y_{in}$ for subject i as n distinct responses, whereas the FDA perspective posits a single functional datum $y_i(\cdot)$ that has been sampled at n points.

In the FDA framework, if we were to replace$_i z$, the vector of predictors for the ith functional response, with $z_i(t)$—for t representing time, this would be referred to as time-varying predictors—the problem would be changed from function-on-

DOI: 10.2202/1557-4679.1246

scalar to a form of function-on-function regression, specifically the "concurrent" model treated in Chapter 14 of RS. In this paper we restrict attention to function-on-scalar regression, i.e., non-time-varying predictors, for simplicity.

The most commonly used penalized basis functions for functional data are splines. The low-rank penalized spline bases favored by RS may be contrasted with two alternative spline approaches. On the one hand, roughness penalization allows for the use of a rich basis, as opposed to unpenalized spline approaches (Huang et al., 2004) that may require a careful choice of a limited number of knots. On the other hand, low-rank spline bases may offer substantial computational savings over smoothing splines with a knot at each observation point, even when the latter are efficiently implemented as in Eubank et al. (2004).

A key challenge in function-on-scalar regression is how to contend with dependence among the error terms for a given functional response. More explicitly, suppose we are given raw responses (2). Writing the associated stochastic terms as $\left[\varepsilon_i(t_j)\right]_{1\leq i\leq N, 1\leq j\leq n}$, we assume that $\varepsilon_{i_1}(t_{j_1})$, $\varepsilon_{i_2}(t_{j_2})$ are independent when $i_1 \neq i_2$, but need not be when $i_1 = i_2$. One way to address this within-function dependence is to try to remove it, by incorporating curve-specific effects in the model such that the remaining error can be viewed as independent and identically distributed. Individual curves may be treated as fixed effects (Brumback and Rice, 1998), but have more often been modeled as random effects (Guo, 2002; Crainiceanu and Ruppert, 2004; Bugli and Lambert, 2006). In this paper we are interested in fast computation with a possibly large number of functional responses, for which estimating individual curves may become infeasible. We therefore focus on the P-OLS and P-GLS methods, which retain within-function dependence but offer contrasting ways of dealing with it.

It should also be noted that function-on-scalar regression models can be fitted by approaches other than spline-type basis functions, including kernel and local polynomial smoothers (e.g., Fan and Zhang, 2000; Chiou et al., 2003, 2004) and wavelets (e.g., Morris and Carroll, 2006; Abramovich and Angelini, 2006; Antoniadis and Sapatinas, 2007; Ogden and Greene, 2010).

3 The Ramsay-Silverman penalized ordinary least squares estimator

This section revisits RS's function-on-scalar regression estimator. Our derivation differs from that of RS (Section 13.4), and in particular shows how the solution can be viewed as a generalized ridge regression estimator. This is not merely an exercise in matrix algebra: rather, the ridge regression viewpoint has two key advantages. First, it motivates the enormous computational improvements mentioned above. Second, it casts RS's estimator as a P-OLS estimator, and points the way toward a

P-GLS alternative. We begin with the more conventional raw response form, and then proceed to the basis coefficient response form; we also derive the latter as a limiting case of the former.

3.1 Responses in raw form: the simplest case

The generalized ridge regression form of the (raw-response) estimator is most transparent in the degenerate case in which $N = q = 1$. Here the raw response matrix Y and parameter matrix B reduce to row vectors $y^T \in \mathscr{R}^n$ and $b^T \in \mathscr{R}^K$. In view of (3), model (1) reduces to the simple nonparametric regression model $y = \Theta b + \varepsilon$, where Θ is the basis function evaluation matrix $[\theta_j(t_i)]_{1\le i\le n, 1\le j\le K}$, and $\varepsilon = [\varepsilon(t_1), \ldots, \varepsilon(t_n)]^T$. (Note that we are temporarily using y and ε to denote vectors in $\mathscr{R}^n$, as opposed to the $\mathscr{R}^N$-valued functions $y(\cdot)$ and $\varepsilon(\cdot)$ defined above.) Taking $X = \Theta$ allows us to write this model in the generic form

$$y = Xb + \varepsilon. \tag{4}$$

Assuming a rich basis, the least-squares solution $\hat{b}$ to (4) will yield a function estimate $\hat{\beta}(t) = \hat{b}^T \theta(t)$ that is excessively wiggly. Instead we minimize the penalized sum of squared errors (SSE) criterion

$$\|y - Xb\|^2 + b^T P b, \tag{5}$$

where P is a positive-semidefinite $K \times K$ matrix such that $b^T P b$ provides a measure of the wiggliness of $b^T \theta(t)$. This so-called roughness penalty is often given by $P = \lambda \int [L(b^T \theta)(t)]^2 dt$ where λ is a nonnegative tuning parameter and L is a linear differential operator such as the second derivative operator, for which the above integral equals $\int \beta''(t)^2 dt$. Criterion (5) is minimized by

$$\hat{b} = (X^T X + P)^{-1} X^T y. \tag{6}$$

We show next that, for general N and q, the RS estimator still has the generalized ridge regression form (6).

3.2 Responses in raw form: the general case

In general the raw responses are modeled as $Y = ZB\Theta^T + E$,[3] where E is the error matrix $[\varepsilon_i(t_j)]_{1\le i\le N, 1\le j\le n}$. RS (p. 239) propose to estimate B by the minimizer of

[3] Applications in which it is appropriate to use different bases for the q coefficient functions require a more complex formulation; see Section 14.4 of RS.

DOI: 10.2202/1557-4679.1246

the penalized SSE

$$\sum_{i=1}^{N}\sum_{j=1}^{n}[y_i(t_j)-(ZB\Theta^T)_{ij}]^2+\sum_{k=1}^{q}\lambda_k\int[L(b_k^T\theta)(t)]^2dt. \tag{7}$$

This is the double SSE over the n points at which each of the N functions is sampled, plus roughness penalties on each of the coefficient functions $\beta_k(\cdot)=b_k^T\theta(\cdot)$, $k=1,\ldots,q$.

To minimize criterion (7) we require the vec operator and Kronecker products. Recall that $\mathrm{vec}(M)$ is the vector formed by concatenating the columns of M, and that, given the $m_A\times n_A$ matrix A with (i,j) entry a_{ij}, and the $m_B\times n_B$ matrix B, the Kronecker product $A\otimes B$ is the $(m_Am_B)\times(n_An_B)$ matrix $(a_{ij}B)_{1\le i\le m_A,1\le j\le n_A}$. We can now express the second term of (7) as $\mathrm{vec}(B^T)^T(\Lambda\otimes R)\mathrm{vec}(B^T)$, where $\Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_q)$ and R is the $K\times K$ matrix with (i,j) entry $\int(L\theta_i)(t)(L\theta_j)(t)dt$; and, using the standard identity

$$\mathrm{vec}(ABC^T)=(C\otimes A)\mathrm{vec}(B), \tag{8}$$

the first term of (7) equals

$$\|\mathrm{vec}(Y^T)-\mathrm{vec}[(ZB\Theta^T)^T]\|^2=\|\mathrm{vec}(Y^T)-(Z\otimes\Theta)\mathrm{vec}(B^T)\|^2.$$

Defining $P_\Lambda=\Lambda\otimes R$, we can then rewrite criterion (7) in form (5) with outcome vector $y=\mathrm{vec}(Y^T)$, design matrix $X=Z\otimes\Theta$, penalty matrix $P=P_\Lambda$, and estimand $b=\mathrm{vec}(B^T)$. Thus, for given values of $\lambda_1,\ldots,\lambda_q$, formula (6) leads directly to the estimate (in vector form)

$$\mathrm{vec}(\hat{B}^T)=\left[(Z\otimes\Theta)^T(Z\otimes\Theta)+P_\Lambda\right]^{-1}(Z\otimes\Theta)^T\mathrm{vec}(Y^T).$$

Another application of (8), along with other standard Kronecker product results, leads to an alternative expression that may be more convenient (cf. eq. (13.25) of RS):

$$\mathrm{vec}(\hat{B}^T)=\left[(Z^TZ)\otimes(\Theta^T\Theta)+P_\Lambda\right]^{-1}(Z\otimes\Theta)^T\mathrm{vec}(Y^T). \tag{9}$$

To provide some intuition about viewing $Z\otimes\Theta$ as the design matrix, we note that, whereas ordinary linear regression models the linear dependence of a scalar response on each predictor, here we are modeling the smooth dependence—captured by K basis coefficients—of n sampled points of a response function on each predictor. Thus each entry z_{ij} of the "base" design matrix Z gives rise to an $n\times K$ submatrix, $z_{ij}\Theta$, of the derived design matrix $Z\otimes\Theta$.

It is worth noting that if we replace Θ with I_n, which may be thought of as a degenerate case of restriction to the span of a basis, then (9) becomes a special case of the (n-dimensional) multivariate ridge regression estimate of Brown and Zidek (1980).

3.3 Responses in basis coefficient form

When the responses are given in the form of an $N \times K$ matrix $C = (c_1 \dots c_N)^T$ of basis coefficients, B is estimated as the minimizer of

$$\int \|C\theta(t) - ZB\theta(t)\|^2 dt + \sum_{k=1}^{q} \lambda_k \int [L(b_k^T \theta)(t)]^2 dt. \tag{10}$$

Comparing this with (7), we see that the double SSE has been replaced by the integral over t of the SSE at point t, i.e. of $\sum_{i=1}^{N} [c_i^T \theta(t) - z_i^T B\theta(t)]^2$.

As in the raw response case, we can minimize (10) by formulating it as a generalized ridge regression criterion. Let $J_{\theta\theta}$ be the $K \times K$ matrix with (i, j) entry $\int \theta_i(t)\theta_j(t)dt$. Appendix B shows that criterion (10) can be expressed as

$$\|\mathrm{vec}(J_{\theta\theta}^{1/2} C^T) - (Z \otimes J_{\theta\theta}^{1/2})\mathrm{vec}(B^T)\|^2 + \mathrm{vec}(B^T)^T P_\Lambda \mathrm{vec}(B^T), \tag{11}$$

which, like (7), has the generalized ridge form (5) with $b = \mathrm{vec}(B^T)$ and $P = P_\Lambda$, but in this case the outcome vector is

$$y = \mathrm{vec}(J_{\theta\theta}^{1/2} C^T) = \begin{pmatrix} J_{\theta\theta}^{1/2} c_1 \\ \vdots \\ J_{\theta\theta}^{1/2} c_N \end{pmatrix} \tag{12}$$

and the design matrix is $X = Z \otimes J_{\theta\theta}^{1/2}$. Formula (6) thus yields the estimate

$$\mathrm{vec}(\hat{B}^T) = \left[(Z \otimes J_{\theta\theta}^{1/2})^T (Z \otimes J_{\theta\theta}^{1/2}) + P_\Lambda \right]^{-1} (Z \otimes J_{\theta\theta}^{1/2})^T \mathrm{vec}(J_{\theta\theta}^{1/2} C^T). \tag{13}$$

One can gain some intuition into this solution by viewing the first (SSE) term of (10), (11) as the SSE for a multivariate linear model with basis coefficients as responses:[4]

$$J_{\theta\theta}^{1/2} c_i = J_{\theta\theta}^{1/2} B^T z_i + u_i = (z_i^T \otimes J_{\theta\theta}^{1/2})\mathrm{vec}(B^T) + u_i \text{ for } i = 1, \dots, N, \tag{14}$$

where the error vectors $u_i = (u_{i1}, \dots, u_{iK})^T$ are assumed independent of each other with common mean zero and covariance matrix Σ.

It is instructive to contrast the above design matrix $Z \otimes J_{\theta\theta}^{1/2}$ with the raw-response design matrix $Z \otimes \Theta$ (see the end of Section 3.2). The submatrices $z_{ij} J_{\theta\theta}^{1/2}$

[4]More precisely, the responses are coefficients with respect to the orthonormal basis given by the components of the $\mathscr{R}^K$-valued function $t \mapsto J_{\theta\theta}^{-1/2}\theta(t)$. Similarly, outcome vector (12) is the concatenation of the N response functions' coefficients with respect to this orthonormal basis.

DOI: 10.2202/1557-4679.1246

of the former, like those of the latter, consist of K columns corresponding to the coefficient functions' expansion with respect to K basis functions. However, whereas the latter submatrices have n rows, the former have K—expressing the assumption that, even if the functional response data consist of a large number n of points, these contain no more than K "pieces of information" given by K basis coefficients. When $J_{\theta\theta} = I_K$, (13) reduces again to the multivariate ridge regression estimate of Brown and Zidek (1980), but in this case the multivariate response is the basis coefficient vector.

In Appendix B we provide further insight into criterion (10) by showing that, in the orthonormal basis ($J_{\theta\theta} = I_K$) case, its first term reduces to $\|C - ZB\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm—i.e., the sum of squared differences between corresponding entries of the $N \times K$ matrices C (the observed basis coefficients) and ZB (the fitted values of the basis coefficients).

3.4 Relationship between the raw-response and basis coefficient estimators

Estimate (13) can be derived as a limiting case of the raw-response solution given above, after a bit of rescaling. To simplify the following, assume $\mathscr{T} = [0,1]$. If we replace the first term of (7) with $\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{n}[y_i(t_j) - (ZB\Theta^T)_{ij}]^2$, then the raw-response estimate (9) becomes

$$\text{(15)} \qquad \mathrm{vec}(\hat{B}^T) = \left[(Z^TZ) \otimes \left(\frac{1}{n}\Theta^T\Theta\right) + P_\Lambda\right]^{-1} \mathrm{vec}\left(\frac{1}{n}\Theta^T Y^T Z\right).$$

If we assume a uniform grid of points $t_j = j/n$ for $j = 1,\ldots,n$ then, as $n \to \infty$, $\frac{1}{n}\Theta^T\Theta \to J_{\theta\theta}$ and $\frac{1}{n}\Theta^T Y^T \to M$, where $M = [\int \theta_k(t) y_i(t)dt]_{1\le k\le K, 1\le i\le N}$. We can alternatively write $M = [\int \theta_k(t)(P_\theta y_i)(t)dt]_{1\le k\le K, 1\le i\le N}$, where P_θ is the projection in $L^2([0,1])$ onto the span of the basis functions $\theta_1,\ldots,\theta_K$. It follows that if, for $i = 1,\ldots,N$, the projection of $y_i(\cdot)$ onto this span is given by $c_i^T\theta(\cdot)$ where $c_i \in \mathscr{R}^K$, then $M = J_{\theta\theta}C^T$ where C is the $N \times K$ matrix with ith row c_i^T. Consequently, in the limit as $n \to \infty$, (15) becomes

$$\text{(16)} \qquad \mathrm{vec}(\hat{B}^T) = \left[(Z^TZ) \otimes J_{\theta\theta} + P_\Lambda\right]^{-1} \mathrm{vec}(J_{\theta\theta}C^TZ),$$

which is readily shown to equal (13) (cf. RS's (p. 238) equivalent formula for $\mathrm{vec}(\hat{B})$).

The above argument says that, for a given basis, fitting the raw-response model for functions sampled on a dense grid is essentially equivalent to projecting the responses onto the basis and then fitting the basis-coefficient model. In this sense, reducing densely sampled responses to their basis expansion entails no real loss of information.

4 Penalized generalized least squares

It is well known that, for a linear model with errors that are not independent and identically distributed, the best linear unbiased estimate of the coefficient vector is given by GLS using the inverse of the error covariance matrix. Two complications arise in our setting. First, it is not clear whether the optimality of unpenalized GLS would carry over to function-on-scalar regression with penalized basis functions (but see Lin et al., 2004, Section 5, for a minimum-variance result that may be relevant). Second, GLS presupposes a known covariance matrix, but in practice the covariance must be estimated, so that the resulting estimator is technically known as *feasible* GLS. For brevity, however, we shall refer to our penalized feasible GLS procedure as P-GLS.

4.1 Algorithm

The key ingredient of P-GLS is an estimate of the covariance matrix of the outcomes given by (12), conditional on the scalar predictors, i.e., the $NK \times NK$ covariance matrix of the error vectors $\begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}$ referred to in (14). Under the assumptions given below (14), the covariance matrix can be written as $I_N \otimes \Sigma$ (cf. the "seemingly unrelated regression" of Zellner, 1962), and the problem reduces to estimating the $K \times K$ matrix $\Sigma = \mathrm{Cov}(u_i)$. This can be done using the $N \times K$ matrix $\hat{U} = (\hat{u}_1 \dots \hat{u}_N)^T$ of P-OLS residuals given by

$$\mathrm{vec}(\hat{U}^T) = \mathrm{vec}(J_{\theta\theta}^{1/2} C^T) - (Z \otimes J_{\theta\theta}^{1/2}) \mathrm{vec}(\hat{B}_{OLS}^T), \tag{17}$$

where $\hat{B}_{OLS}$ is the P-OLS estimate (16). Let $\hat{U}^*$ denote the matrix formed from $\hat{U}$ by centering each column. The covariance can then be estimated by $\hat{\Sigma} = \hat{U}^{*T} \hat{U}^* / d$ for a suitable d. It would be natural to take $d = N - df$ where df is the residual degrees of freedom, but in this context, defining the latter quantity is not straightforward. It therefore seems most reasonable to take $d = N$, which yields the maximum likelihood estimate (MLE) under normality.

Given our covariance estimate, the P-GLS criterion is obtained by replacing the first (SSE) term of (11) with

$$\left[\mathrm{vec}(J_{\theta\theta}^{1/2} C^T) - (Z \otimes J_{\theta\theta}^{1/2}) \mathrm{vec}(B^T)\right]^T (I_N \otimes \hat{\Sigma})^{-1} \left[\mathrm{vec}(J_{\theta\theta}^{1/2} C^T) - (Z \otimes J_{\theta\theta}^{1/2}) \mathrm{vec}(B^T)\right]. \tag{18}$$

DOI: 10.2202/1557-4679.1246

As above, one can use a generalized ridge regression representation to derive the minimizer

$$\text{(19)} \qquad \mathrm{vec}(\hat{B}^T) = \left[(Z^T Z) \otimes (J_{\theta\theta}^{1/2} \hat{\Sigma}^{-1} J_{\theta\theta}^{1/2}) + P_\Lambda\right]^{-1} \mathrm{vec}(J_{\theta\theta}^{1/2} \hat{\Sigma}^{-1} J_{\theta\theta}^{1/2} C^T Z)$$

(see Appendix C). Note that this estimate reduces to (16) when $\hat{\Sigma} = I$.

As in Krafty et al. (2008), we may wish to estimate B by an iterative P-GLS procedure:

1. Compute the P-OLS estimate $\mathrm{vec}(\hat{B}_{OLS}^T)$ by (16).
2. Use the residuals (17) to obtain a covariance matrix estimate $\hat{\Sigma}$, and insert $\hat{\Sigma}$ into (19) to derive a provisional P-GLS estimate $\mathrm{vec}(\hat{B}_{GLS}^T)$.
3. Return to step 2 (now using the P-GLS residuals), and repeat until convergence of $\hat{B}_{GLS}$.

One goal of the simulations in Section 7 is to evaluate whether iterating to convergence improves the performance of P-GLS.

4.2 Other approaches to covariance estimation

The MLE $\hat{\Sigma}$ can be inverted only if $N > K$, and even if this inequality holds, the estimate may become unstable for large K. Krafty et al. (2008), in a repeated-measures setting, assume a covariance matrix of the form $\Gamma + \sigma^2 I$, where σ^2 might be interpreted as measurement error, and employ a Kullback-Leibler criterion to regularize the covariance estimate. Their method requires cross-validation over two tuning parameters. This covariance model is especially appropriate when the data are sparse and noisy and when covariance estimation is of intrinsic interest (Yao et al., 2005). In the RS framework, however, the basis coefficients are generally taken to represent denoised functional data, obviating the need for such a computationally intensive approach. A less computationally demanding regularized estimate of the covariance matrix, using the optimal shrinkage method of Schäfer and Strimmer (2005), was tested in simulations (not shown) but did not appear to improve performance.

In the nonparametric regression literature, some authors have used mixed model software to perform smoothing and estimate correlation structure simultaneously (e.g., Wang, 1998; Durban and Currie, 2003; Krivobokova and Kauermann, 2007). An analogous approach might be attempted for function-on-scalar regression, as an alternative to P-GLS. However, this would entail imposing one of several standard parametric correlation structures; and for some examples, such as the periodic functional data considered in Section 8.1, none of these structures may be appropriate.

5 Smoothness selection

Selection of the smoothing parameters $\lambda_1, \ldots, \lambda_q$ is a crucial step,[5] and it is here that our approach attains notable computational efficiency, as this section will explain.

5.1 Leave-one-function-out cross-validation for P-OLS

In the P-OLS setting, the smoothing parameters are usually chosen by a cross-validation (CV) procedure in which one function is left out at a time (Rice and Silverman, 1991). The criterion to be minimized is the cross-validated integrated squared error

$$\frac{1}{N}\sum_{i=1}^{N}\int_{\mathscr{T}}[y_i(t)-\hat{y}_i^{(-i)}(t)]^2dt, \tag{20}$$

where $\hat{y}_i^{(-i)}(\cdot)$ is the predicted value for the ith functional response, based on the model fitted to the other $N-1$ functional responses. Letting c_i and $\hat{c}_i^{(-i)}$ denote the vectors of basis coefficients determining the two functions in (20), the CV criterion (20) is equal to

$$\frac{1}{N}\sum_{i=1}^{N}\|J_{\theta\theta}^{1/2}(c_i-\hat{c}_i^{(-i)})\|^2. \tag{21}$$

Direct computation of (21) would require fitting the model to almost the entire data set N times, but this can be avoided by a trick that we shall explain in reference to the generic penalized regression criterion (5). Suppose the outcome vector is partitioned into N groups, say $y=(y_1^T,\ldots,y_N^T)^T$. Let H be the "hat matrix" such that minimizing (5) yields fitted values $\hat{y}=(\hat{y}_1^T,\ldots,\hat{y}_N^T)^T=Hy$, and partition H into blocks determined by the N groups: $H=(H_{ij})_{1\le i\le N,1\le j\le N}$. Consider a CV procedure in which the same model is refitted with each group deleted in turn, and let $\hat{y}_i^{(-i)}$ denote the fitted values for the ith group based on the group-i-deleted model. It can be shown by the Sherman-Morrison-Woodbury formula (e.g., Golub and Van Loan, 1996) that

$$y_i-\hat{y}_i^{(-i)}=(I-H_{ii})^{-1}(y_i-\hat{y}_i). \tag{22}$$

If the N groups have one element each, this reduces to $y_i-\hat{y}_i^{(-i)}=(y_i-\hat{y}_i)/(1-h_{ii})$ where h_{ii} is the ith diagonal element of H. This last identity provides a well-known computational shortcut for ordinary leave-one-out CV. The more general

[5] Since smoothness is controlled by $\lambda_1,\ldots,\lambda_q$, the precise choice of the number of basis functions K is generally seen as much less critical, as long as it is large enough to capture the detail of the function(s) being estimated. Hence K is often chosen informally (e.g., Ruppert, 2002; Ruppert et al., 2003, pp. 125–127; Wood, 2006a, p. 161).

DOI: 10.2202/1557-4679.1246

identity (22) has been used previously for leave-one-function-out CV (Hoover et al., 1998), as well as for multifold CV (Zhang, 1993); our generalized ridge regression reformulation of RS's development is what makes it available in the present setting as well. Here the left side of (22) equals $J_{\theta\theta}^{1/2}(c_i - \hat{c}_i^{(-i)})$. Using the results of Section 3.3 to evaluate the right side of (22), criterion (21) becomes

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\left[I_K - (z_i^T \otimes J_{\theta\theta}^{1/2})\left\{(Z^T Z)\otimes J_{\theta\theta} + P_\Lambda\right\}^{-1}(z_i \otimes J_{\theta\theta}^{1/2})\right]^{-1} J_{\theta\theta}^{1/2}(c_i - \hat{c}_i)\right\|^2.$$

Whereas repeated model fits would require inverting N $K(N-1)\times K(N-1)$ matrices, the most expensive part of evaluating the above expression is inverting N $K\times K$ matrices.

We remark that efficiency might be further improved by using k-fold rather than leave-one-out CV, say with $k = 5$ or 10. It should also be noted that smoothness selection for generalized ridge regression is usually accomplished by optimizing not CV but either generalized cross-validation (GCV) or restricted maximum likelihood (REML) (Reiss and Ogden, 2009). However, the latter two criteria presuppose that the error covariance either is a multiple of the identity, or else is taken into account—either by using P-GLS, or by simultaneous estimation of the dependence structure as mentioned in Section 4.2.

In the single smoothing parameter case, i.e. when $\lambda_1 = \ldots = \lambda_q = \lambda$, the CV criterion can be computed rapidly for different values of λ by using Demmler-Reinsch orthogonalization (e.g., Ruppert et al., 2003). As an alternative to the usual grid search, the generic minimizer implemented in the R function `optimize` (Brent, 1973) appears to work quite well for finding the minimum of the CV score as a function of λ. Minimizing the CV score as a function of multiple smoothing parameters seems much more difficult, and our current implementation works only for the common smoothing parameter case. The disadvantages of this restriction may be overcome to some degree by scaling each predictor to have unit mean square; see also Section 8.2.

5.2 P-GLS

As noted above, the automatic smoothing parameter selection criteria GCV and REML are available for P-GLS. REML appears to be more popular for regression with functional responses (e.g., Brumback and Rice, 1998; Guo, 2002; Krafty et al., 2008), and is a particularly natural choice when the model includes random effects. The demonstration by Krivobokova and Kauermann (2007) that REML is more robust than GCV to misspecification of the correlation structure provides further support for favoring REML in the present setting. In our implementation, smoothing parameters are optimized efficiently, within each iteration of the algorithm of

Section 4.1, by calling the `gam` function in the `mgcv` package (Wood, 2006a), to which a REML option has recently been added (Wood, 2010). This function is, to the best of our knowledge, the most stable and efficient publicly available software for estimation of separate smoothing parameters $\lambda_1, \ldots, \lambda_q$ in models of generalized ridge regression type.

6 Inference

6.1 Pointwise confidence intervals

To derive approximate standard errors for the function estimates $\hat{\beta}_i(t)$, observe that

$$\hat{\beta}_i(t) = \theta(t)^T \hat{B}^T e_i = [e_i^T \otimes \theta(t)^T]\mathrm{vec}(\hat{B}^T)$$

where e_i denotes the vector in $\mathscr{R}^q$ with 1 in the ith position and 0 elsewhere, and thus $\mathrm{Var}[\hat{\beta}_i(t)] = [e_i^T \otimes \theta(t)^T]\mathrm{Var}[\mathrm{vec}(\hat{B}^T)][e_i \otimes \theta(t)]$. The problem reduces, then, to estimating the variance of $\mathrm{vec}(\hat{B}^T)$. Different approaches to this task have been proposed for P-OLS and for P-GLS. In either case, however, the variance estimator can be explained more clearly by referring to the generic expression (6) with $\hat{b} = \mathrm{vec}(\hat{B}^T)$.

For P-OLS, (6) suggests the variance estimate

$$\widehat{\mathrm{Var}}[\mathrm{vec}(\hat{B}^T)] = (X^T X + P)^{-1} X^T \widehat{\mathrm{Var}}(y|X) X (X^T X + P)^{-1}$$

with y, X and P as given in the text immediately preceding (13). As in Section 4.1 we can take $\widehat{\mathrm{Var}}(y|X) = I_N \otimes \hat{\Sigma}$, where $\hat{\Sigma}$ is the MLE derived from the residuals in the basis-coefficient domain. Plugging in the values of X and P yields

$$\widehat{\mathrm{Var}}[\mathrm{vec}(\hat{B}^T)] = (Z_J^{*T} Z_J^* + P_\Lambda)^{-1} Z_J^{*T} (I_N \otimes \hat{\Sigma}) Z_J^* (Z_J^{*T} Z_J^* + P_\Lambda)^{-1}, \tag{23}$$

where $Z_J^* = Z \otimes J_{\theta\theta}^{1/2}$.

An analogous expression could be derived for P-GLS. Note, however, that this estimate ignores the added variation due to the need to estimate Λ. In addition, confidence intervals based on (23) may have poor coverage since the roughness penalty introduces bias in the estimation of $\mathrm{vec}(B^T)$ (Wood, 2006a, 2006b). The latter problem can be remedied by instead using Bayesian confidence intervals, or credible intervals, as developed by Wahba (1983) and Silverman (1985). In Appendix C we obtain the posterior covariance estimate from which such intervals can be derived:

$$\widehat{\mathrm{Var}}[\mathrm{vec}(B^T)|Y] = \hat{\sigma}^2 \left[(Z^T Z) \otimes (J_{\theta\theta}^{1/2} \hat{\Sigma}^{-1} J_{\theta\theta}^{1/2}) + P_\Lambda \right]^{-1}, \tag{24}$$

DOI: 10.2202/1557-4679.1246

where $\hat{\sigma}^2$ is a residual variance estimate given there. Appendix C also explains why this approach works only for P-GLS but not for P-OLS. In summary, then, we base interval estimation on the frequentist formula (23) for P-OLS, and on the Bayesian formula (24) for P-GLS.

6.2 Hypothesis testing

RS (p. 227) propose to test the effects of a set of scalar predictors in pointwise fashion by means of F-statistics. Suppose we wish to test a null model with design matrix Z_0 against the alternative model (1). The statistic at point t is given by

$$F(t) = \frac{[\|y(t) - Z_0\hat{\beta}_0(t)\|^2 - \|y(t) - Z\hat{\beta}(t)\|^2]/(m - m_0)}{\|y(t) - Z\hat{\beta}(t)\|^2/(N - m)},$$

where $\hat{\beta}_0$, $\hat{\beta}$ are the function estimates, and m_0, m are the model degrees of freedom, for the null and alternative models respectively. Given the dependence among models at different ts, $F(t)$ may not have the $F_{m-m_0,N-m}$ distribution under the null model. But in any case, inference is not usually performed by referring $F(t)$ at a particular t to an F distribution. More often, one conducts simultaneous testing by comparing the observed $\{F(t) : t \in \mathscr{T}\}$ to the permutation distribution of $\sup_{t\in\mathscr{T}} F(t)$. In practice, one approximates this distribution by Monte Carlo simulation. The null model can then be rejected at the $100\alpha\%$ level if, for some t, $F(t)$ exceeds the $100(1-\alpha)$ percentile of the permuted-data values of $\sup_{t\in\mathscr{T}} F(t)$.

6.3 Overall and pointwise model selection

If permutation tests confirm that each of several scalar predictors has a significant effect on the functional outcome, it is natural to ask which of these is the most predictive. More generally we may be interested in choosing the best among several possibly non-nested function-on-scalar regression models. The P-OLS method offers the most straightforward approach to model selection: one can simply select the model with the lowest cross-validated integrated squared error (20).

In at least some applications, however, it may make sense to allow for a different "best" model within different subsets of $\mathscr{T}$, the response functions' domain. It is natural to perform pointwise model selection criterion for each $t \in \mathscr{T}$ using the cross-validated *pointwise* squared error $\frac{1}{N}\sum_{i=1}^{N}[y_i(t) - \hat{y}_i^{(-i)}(t)]^2$, i.e. the quantity whose integral over $\mathscr{T}$ equals (20). Note that since the functional linear model "borrows strength" across values of t and thus obtains a smooth estimate of $\hat{y}_i^{(-i)}(\cdot)$, the proposed pointwise CV criterion should be more stable than naïve ordinary CV based on fitting the model separately at each t. In practice one would use the

equivalent expression $\frac{1}{N}\sum_{i=1}^{N}[(c_i - \hat{c}_i^{(-i)})^T \theta(t)]^2$, which can be computed without repeated model fits by the methods of Section 5.1.

7 Comparative simulations

We conducted a simulation study using the three-group one-way functional ANOVA model[6] $y_i(t) = \mu(t) + \beta_{gp(i)}(t) + \varepsilon_i(t)$ ($t \in [0,1]$), where $gp(i)$ denotes the group (1, 2, or 3) to which the ith functional response belongs. The mean function $\mu(t) = 0.4\arctan(10x-5)+0.6$, and the group effect functions $\beta_1(t) = -0.5e^{-10t} -0.04\sin(8t)-0.3t+0.5$, $\beta_2(t) = -(t-0.5)^2 - 0.15\sin(13t)$, and $\beta_3(t) = -\beta_1(t) - \beta_2(t)$, are shown in the top panels of Figure 2. The error functions $\varepsilon_i(\cdot)$ were simulated from a mean-zero Gaussian process with covariance $V(s,t) = \sigma_1^2 0.15^{|s-t|} + \sigma_2^2\delta_{st}$, where $\delta_{st} = 1$ if $s = t$ and $= 0$ otherwise, sampled at $t = m/200$ for $m = 0,\ldots,200$ (cf. Section 4.2 above). Note that, although we adopt a fixed-effects modeling approach in this paper, for purposes of simulating functional responses it is more natural to think of $V(s,t)$ as arising from a mixed model in which the error is decomposed as $\varepsilon_i(t) = \varepsilon_{i1}(t) + \varepsilon_{i2}(t)$, where $\mathrm{Cov}[\varepsilon_{i1}(s), \varepsilon_{i1}(t)] = \sigma_1^2 0.15^{|s-t|}$, $\mathrm{Cov}[\varepsilon_{i2}(s), \varepsilon_{i2}(t)] = \sigma_2^2\delta_{st}$, and $\varepsilon_{11},\ldots,\varepsilon_{N1}$ are independent of $\varepsilon_{12},\ldots,\varepsilon_{N2}$. In mixed model terms, $\mu(t) + \beta_{gp(i)}(t) + \varepsilon_{i1}(t)$ is the underlying true function for observation i; σ_1^2 represents the variation among the true functions in each group; and σ_2^2 represents the noise or error variance at each point of these functions, possibly due to measurement error. We used samples of $N_g = 10, 60$ for each of the three groups, and two levels of the among-function standard deviation σ_1 (0.05 and 0.15); σ_2 was fixed at 0.05. For each combination of N_g and σ_1, we simulated 500 data sets, and fitted the model by four methods:

1. P-OLS with a common smoothing parameter λ for all four coefficient functions being estimated, chosen by cross-validation;

2. P-GLS with a common λ chosen by REML and with Σ estimated only once, i.e., only one step of the iterative algorithm;

3. same as method 2, but with iteration to convergence;

4. same as method 2, but with separate smoothing parameters λ_1–λ_4.

The raw responses were smoothed with a 20-knot cubic B-spline basis, by means of the R function `smooth.spline`, to produce functions similar to those shown at

[6]This is not to be confused with the very different type of functional ANOVA studied, for instance, by Hooker (2007).

DOI: 10.2202/1557-4679.1246

the bottom of Figure 2; model fitting was then performed on the responses in spline coefficient form. We imposed the constraint $\beta_1(t)+\beta_2(t)+\beta_3(t)=0$ at each t by a standard device (Wood, 2006a, pp. 185–186). In simulations not reported here, we tried using GCV rather than REML for the above three versions of P-GLS. The results tended to be slightly worse than with REML.

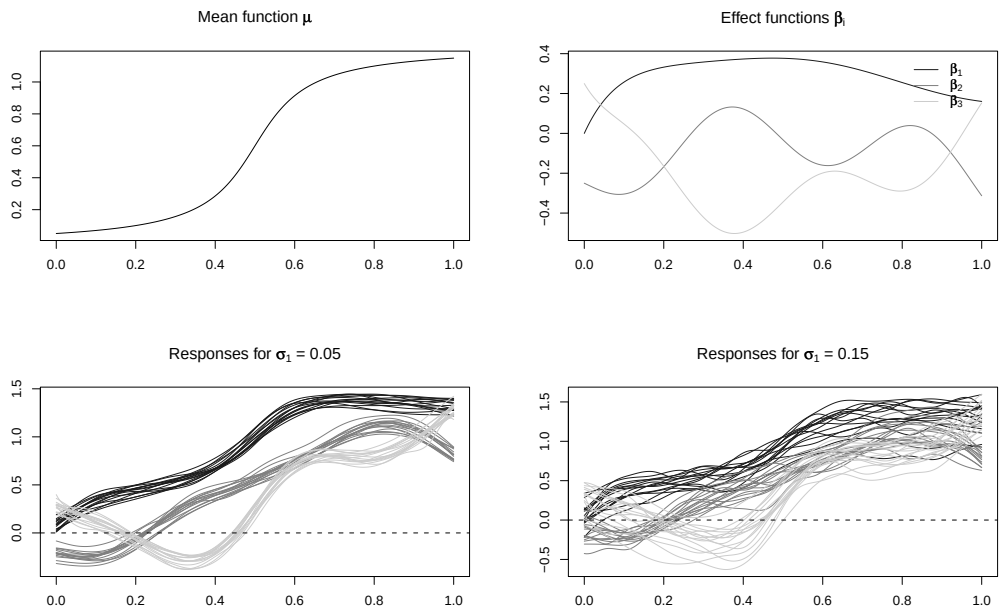

Figure 2: Top: Mean function $\mu(\cdot)$ and group effect functions $\beta_i(\cdot)$ $(i=1,2,3)$ for the simulations. Bottom: Example smoothed response functions for the three groups, color-coded as in the top right panel, with $\sigma_1=0.05$ and with $\sigma_1=0.15$.

Figure 3 presents 1000 times the mean integrated squared error in estimating the four coefficient functions. The box plots have been truncated to facilitate visual comparisons; the scale of each subfigure was chosen so as to include at least the lower 95% of each empirical distribution. As one would expect, the error is lower for μ than for the group effects β_i. Overall, the four methods perform quite similarly. The most striking difference is in the easiest scenario ($N_g=60, \sigma_1=0.05$), in which the three P-GLS methods outperform P-OLS. On the other hand, P-OLS does slightly better than P-GLS in the most difficult scenario ($N_g=10, \sigma_1=0.15$).

Coverage for 95% confidence/credible intervals—in the "across-the-function" sense, i.e., the proportion of the true function lying within the given interval—is shown in Figure 4. The four subfigures use a common scale chosen to include at

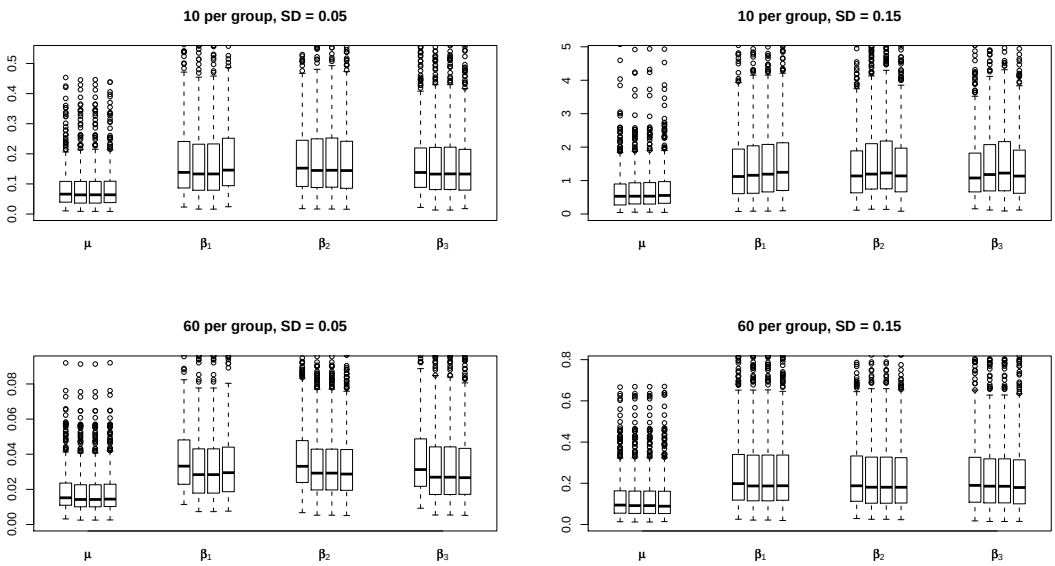

Figure 3: Mean integrated squared error times 1000, with $N_g = 10, 60$ and with $\sigma_1 = .05, .15$. Each set of box plots represents the four methods: P-OLS, one-step P-GLS, iterative P-GLS, and P-GLS with multiple smoothing parameters.

least the upper 95% of each empirical distribution. All methods have median coverage well above the nominal level, except when $N_g = 60, \sigma_1 = 0.05$, in which case the median coverage is quite close to 95%. However, the first quartile of the coverage is seen to lie below 95% in all but one of the box plots. The P-OLS intervals tend to be somewhat wider than their P-GLS counterparts in the $N_g = 10, \sigma_1 = 0.15$ scenario, and slightly narrower otherwise, especially for $N_g = 60, \sigma_1 = 0.05$. The lack of a clear overall pattern may arise from two opposing tendencies. On the one hand, the credible intervals for P-GLS might be wider than the P-OLS confidence intervals due to the former's incorporating a bias correction (see above, Section 6.1); on the other hand, the P-GLS intervals might be too narrow due to not accounting for error in covariance estimation.

8 Real data examples

8.1 Canadian temperatures

The Canadian weather data set is familiar to students of FDA. RS use this data set to illustrate a number of their methods, and it is included in the `fda` package (Ramsay et al., 2009). The functional data consist of mean daily temperature and precipita-

DOI: 10.2202/1557-4679.1246

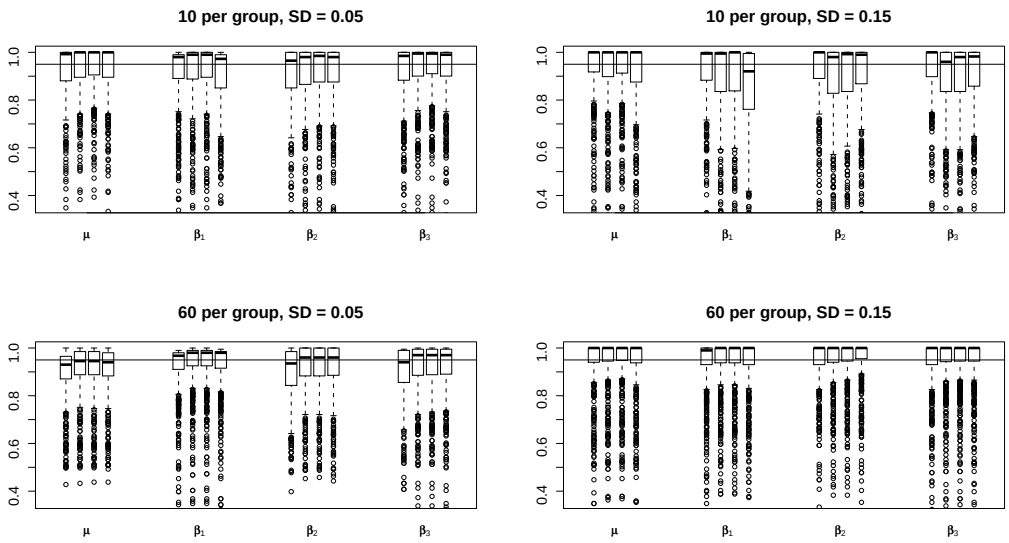

Figure 4: Observed across-the-function coverage for 95% confidence/credible intervals. Each set of box plots represents the four methods: P-OLS, one-step P-GLS, iterative P-GLS, and P-GLS with multiple smoothing parameters.

tion for the years 1960–1994 at 35 stations across Canada. Here we consider P-OLS regression of the temperature curves on two scalar variables: the region in which the station is located (Arctic, Atlantic, Continental, or Pacific), and the station's latitude. Projecting the temperature functions onto the Fourier basis consisting of a constant function along with $f_k(t) = \sin(2\pi kt/365)$ and $g_k(t) = \cos(2\pi kt/365)$ for $k = 1, \ldots, 12$ yields the curves shown at left in Figure 5. Tests of the two predictors based on 300 permutations, as described in Section 6.2, found both to be extremely significant: indeed, in both cases, the real-data F statistic *at each point* lay far into the right tail of the simulated permutation distribution of *maximal* F statistics. These results come as no great surprise. A more interesting question concerns the two predictors' comparative strength. Which tells us more about a site's average temperature: knowing which region it belongs to, or knowing its latitude? This question can be answered separately for each day of the year by the pointwise CV method of Section 6.3. As Figure 5 shows, for the period from May through November, latitude has lower CV, and thus appears to be a stronger determinant of temperature. From January through April, however, regional differences—such as, perhaps, the tendency for coastal areas to have milder winters—appear to be more important than latitude alone.

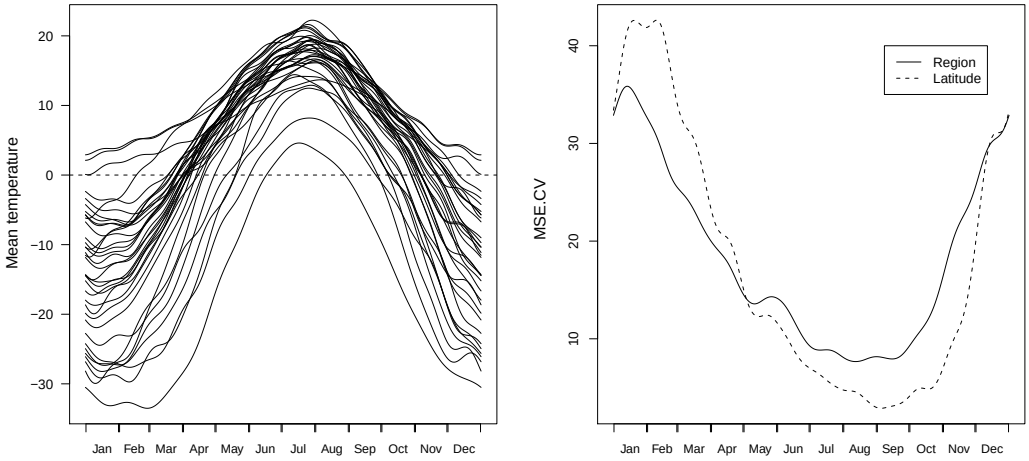

Figure 5: At left, smoothed curves representing daily mean temperatures at the 35 Canadian locations. At right, pointwise cross-validation MSE for regression of the temperature functions on region (solid curve) and on latitude (dashed curve).

Our P-OLS implementation, running in R version 2.9.0 on a MacBook Pro with a 2.16 GHz Intel Core Duo processor, required 2.9 seconds to optimize λ by cross-validation over a grid of 21 candidate values (or 5.1 seconds to find λ via the `optimize` function with the default tolerance level, which gives higher accuracy than would usually be needed). By contrast, the function `fRegress.CV` in the `fda` package, which refits the model for each delete-one-function data set, needed about 32 seconds to compute the CV for a single candidate λ. Thus, calling `fRegress.CV` for each λ in the same grid of values would take over 200 times as long as our method.

8.2 Functional connectivity

We now return to the fMRI data displayed in Figure 1. In traditional fMRI experiments, the blood oxygen level dependent (BOLD) signal, an index of brain activity, is recorded at each of a dense grid of brain locations, known as voxels, while the subject attends to a series of stimuli. By contrast, our data were acquired by resting-state fMRI, in which subjects are scanned while they attend to no stimulus in particular. The objective is not to study the brain's response to particular stimuli, but to investigate functional networks of brain regions whose BOLD time courses move in tandem. A popular analytic strategy involves choosing a "seed" brain region, and computing the correlation between its BOLD time series and those of

DOI: 10.2202/1557-4679.1246

every voxel in the brain. A recent set of seed correlation analyses (Kelly et al., 2009) suggests that as the brain develops from childhood to adulthood, high correlations between distant brain regions become more prevalent. As a tool for studying the relationship between connectivity and distance from a given seed, we have developed subject-specific connectivity-distance functions. Briefly, these functions are derived by applying the inverse hyperbolic tangent (or Fisher z transform) to the correlations of each voxel with the seed region, then estimating a conditional quantile (say, the 95th percentile) of the resulting values as a smooth function of distance from the seed (Koenker et al., 1994). The fitted function is then projected onto a cubic B-spline basis with 40 equally-spaced internal knots. Repeating this for each of N subjects yields functional responses $y_1(\cdot),\ldots,y_N(\cdot)$, which we modeled as $y_i(d) = \beta_1(d) + z_i\beta_2(d) + \varepsilon_i(d)$, where d denotes distance from the seed and z_i is the ith subject's age.

This model was fitted to data from 59 participants, ranging in age from 20 to 49, who were scanned at New York University. Distance was measured from a seed located in the posterior cingulate/precuneus region, and age was centered to mean zero. The first plot in Figure 6 shows an estimate of β_1; this function is highest for short distances, as expected, but it also has a slight peak around 120 mm. The next two plots display estimates of β_2 from P-GLS and P-OLS models, respectively, with a common smoothing parameter λ used for both β_1 and β_2. These estimates resemble each other closely but seem rather too bumpy, evidently because $|\beta_1(d)| \gg |\beta_2(d)|$ for most d, so that the penalty $\lambda \int \beta_1''^2 + \lambda \int \beta_2''^2$ is dominated by its first term. A P-GLS fit with separate smoothing parameters for β_1 and β_2 yields a much smoother estimate of the latter, as seen in the lower left plot.

Although our P-OLS implementation assumes a common λ, multiple smoothing parameter selection can be achieved, in effect, by rescaling. Consider fitting the model $y_i(d) = \beta_1(d) + z_i^*\beta_2^*(d) + \varepsilon_i(d)$, where $z_i^* = cz_i$ and $\beta_2^*(d) = \beta_2(d)/c$ for some $c > 0$, with a common λ, thereby obtaining the estimate $\hat{\beta}_2^*$, and taking $c\hat{\beta}_2^*$ as the estimate for β_2. This is practically equivalent to fitting the original model with separate parameters λ_1, λ_2 related by $\lambda_1 = c^2\lambda_2$. We can thus fit rescaled models with a range of values of $c = \sqrt{\lambda_1/\lambda_2}$, and choose the CV-minimizing c. (For a model having more than one scalar predictor, CV would have to be minimized over a multidimensional grid.) The optimal value $c = .00056$ yielded the age effect function shown in the lower center plot of Figure 6. This is quite similar to the P-GLS estimate with separate smoothing parameters, but the 95% confidence interval is wider. The lower right plot shows expected profiles at the first and third quartiles of the age distribution. All four model fits point to a negative age effect at short distance from the seed, and little or no effect at longer distances. A simultaneous test based on 1500 permutations (not shown) found a significant (at the 5% level)

negative effect of age, for distances up to 24 mm from the seed. Our observation that older subjects' high-connectivity regions tend to be less concentrated around the seed is consistent with the findings of Kelly et al. (2009).

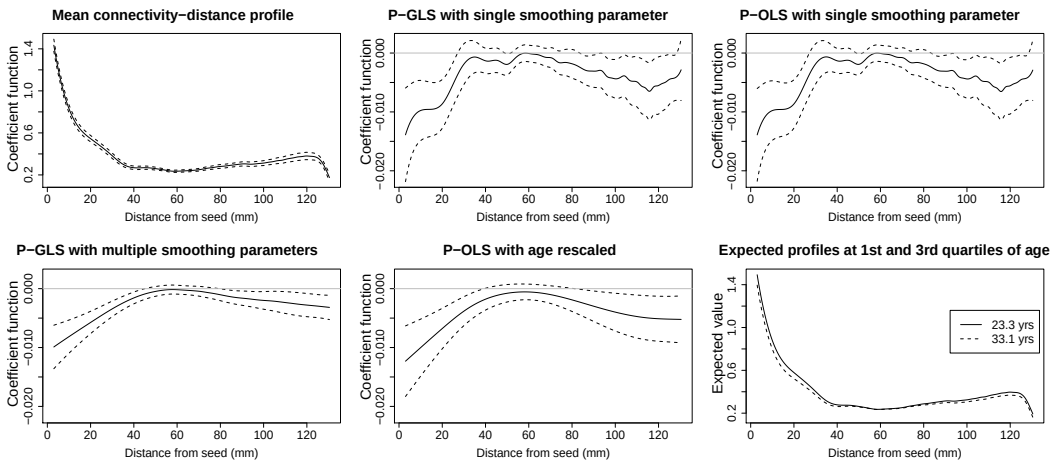

Figure 6: The upper left plot shows the estimate (solid) and 95% credible interval (dashed) for the intercept function β_1 in the fMRI application. The next four plots show the estimates and 95% intervals for the age effect function β_2, derived by P-GLS and P-OLS with a single smoothing parameter; P-GLS with multiple smoothing parameters; and P-OLS with a single smoothing parameter, but with age optimally rescaled. The lower right plot shows expected profiles at the first and third quartiles (solid and dashed, respectively) of the observed age distribution. The upper left and lower right plots are based on the P-GLS model with multiple smoothing parameters, but the other models produced similar results.

9 Conclusion

It is difficult to choose a clear "winner" between the two basic methods we have discussed. P-OLS is more conceptually straightforward and less vulnerable to undercoverage in small samples; moreover, the CV score, computed in the course of smoothing parameter selection, can double as a criterion for selecting among different models, in either an overall or a pointwise sense. On the other hand, our implementation of P-GLS with the `mgcv` package allows for rapid selection of multiple smoothing parameters, the advantages of which are apparent from our analysis of the connectivity data. Our simulation results do not unequivocally favor either P-OLS or any version of P-GLS, although they suggest that P-GLS may become more accurate for functional ANOVA with relatively large, homogeneous groups

DOI: 10.2202/1557-4679.1246

of functional responses. The relative merits of P-OLS and P-GLS should become clearer as these methods are applied to a wider variety of data sets. It is our hope that our fast implementation of function-on-scalar regression will contribute to the more widespread application of models of this type.

As noted above (see footnote 1 and Section 4.2), in this paper we have been primarily concerned with densely sampled functional data of the type emphasized by RS. Rather different approaches have been developed for sparsely and irregularly sampled data with significant measurement error (e.g., Chiou et al., 2003, 2004; Yao et al., 2005; Krafty et al., 2008). It would be of great interest to demarcate more precisely the types of applications for which each approach is best suited.

A number of extensions of our methods are planned, including smooth (as opposed to linear) dependence of the functional response on continuous scalar predictors; generalized linear models with functional responses; and permutation tests appropriate for more complex designs. The methods described here are implemented in the R package `refund` (*re*gression with *fun*ctional *d*ata), available at `http://cran.r-project.org/web/packages/refund`.

Appendix A: Reduction to using the same bases for the responses and coefficient functions

The first term of equation (10) gave the (unpenalized) integrated SSE as $\int \|C\theta(t) - ZB\theta(t)\|^2 dt$. In RS's (Section 13.4) formulation, the functional responses are given by $y(t) = C\phi(t)$, where $\phi(t) = [\phi_1(t), \ldots, \phi_{K_y}]^T$ for some basis functions $\phi_1, \ldots, \phi_{K_y}$ not necessarily coinciding with the basis functions $\theta_1, \ldots, \theta_K$ used to expand the coefficient functions. We show here why, for our purposes, it suffices to assume that the bases do coincide.

The integrated SSE can be written as

$$\int \|C\phi(t) - ZB\theta(t)\|^2 dt = \sum_{i=1}^{N} \int [c_i^T \phi(t) - z_i^T B\theta(t)]^2 dt, \tag{25}$$

i.e., this criterion can be expressed either as an integral of a SSE, or as a sum of integrated squared errors. Let $c_i^{*T}\theta$ be the orthogonal projection (in L^2) of $c_i^T\phi$ onto the span of $\theta_1, \ldots, \theta_K$; it is easily shown that $c_i^* = J_{\theta\theta}^{-1} J_{\theta\phi} c_i$, where where $J_{\theta\phi}$ is the $K \times K_y$ matrix with (i, j) entry $\int \theta_i(t)\phi_j(t)dt$. By the definition of an orthogonal projection, we have the orthogonal decomposition

$$c_i^T\phi - z_i^T B\theta = (c_i^T\phi - c_i^{*T}\theta) + (c_i^{*T}\theta - z_i^T B\theta),$$

and thus, by the Pythagorean identity for inner product spaces, the integral on the right side of (25) can be decomposed as

$$\int [c_i^T \phi(t) - c_i^{*T} \theta(t)]^2 dt + \int [c_i^{*T} \theta(t) - z_i^T B\theta(t)]^2 dt.$$

The first of these two integrals does not depend on B. Consequently, the problem of finding B to minimize the penalized SSE (i.e., (25) plus a roughness penalty) is unchanged if we replace $C\phi(t)$ with $C^*\theta(t)$, where $C^* = (c_1^* \dots c_N^*)^T = CJ_{\theta\phi}^T J_{\theta\theta}^{-1}$. For our development, then, we can assume that $\{\phi_1, \dots, \phi_{K_y}\} = \{\theta_1, \dots, \theta_K\}$.

Appendix B: Derivation of alternative expressions for criterion (10)

Here we derive expression (11), and a further simplification in the orthonormal basis case. We have

$$\begin{aligned}
\int \|C\theta(t) - ZB\theta(t)\|^2 dt &= \int \mathrm{tr}[(C-ZB)^T (C-ZB)\theta(t)\theta(t)^T] dt \\
&= \mathrm{tr}[(C-ZB)^T (C-ZB) J_{\theta\theta}] \\
&= \mathrm{tr}[\{(C-ZB)J_{\theta\theta}^{1/2}\}^T \{(C-ZB)J_{\theta\theta}^{1/2}\}] \\
&= \|\mathrm{vec}(J_{\theta\theta}^{1/2} C^T - J_{\theta\theta}^{1/2} B^T Z^T)\|^2 \qquad (26) \\
&= \|CJ_{\theta\theta}^{1/2} - ZBJ_{\theta\theta}^{1/2}\|_F^2. \qquad (27)
\end{aligned}$$

By (8), (26) equals the first term of (11). The second (penalty) term of (11) is the same as in the raw response case (Section 3.2). We therefore conclude that (10) equals (11), as claimed in Section 3.3. In the orthonormal case we take $J_{\theta\theta}^{1/2} = I_K$ in (27), and immediately obtain the simpler expression $\|C - ZB\|_F^2$.

Appendix C: Derivation of the P-GLS estimate and posterior covariance matrix

The Cholesky decomposition $\hat{\Sigma}^{-1} = L^T L$ allows us to write the P-GLS criterion, i.e., the SSE (18) plus the second (penalty) term of (11), as

$$\left\| \mathrm{vec}(LJ_{\theta\theta}^{1/2} C^T) - [Z \otimes (LJ_{\theta\theta}^{1/2})] \mathrm{vec}(B^T) \right\|^2 + \mathrm{vec}(B^T)^T P_\Lambda \mathrm{vec}(B^T). \qquad (28)$$

This criterion can be thought of as arising from "prewhitening" each observation via premultiplication by L, and then performing P-OLS. Expression (28) has form (5) with $X = Z \otimes (LJ_{\theta\theta}^{1/2})$, $y = \mathrm{vec}(LJ_{\theta\theta}^{1/2} C^T)$, and (as above) $b = \mathrm{vec}(B^T)$ and $P = P_\Lambda$,

DOI: 10.2202/1557-4679.1246

so that (6) gives the minimizer

$$\begin{aligned}\mathrm{vec}(\hat{B}^T) &= \left[\{Z\otimes(LJ_{\theta\theta}^{1/2})\}^T\{Z\otimes(LJ_{\theta\theta}^{1/2})\}+P_\Lambda\right]^{-1}\\ &\quad\times[Z\otimes(LJ_{\theta\theta}^{1/2})]^T\mathrm{vec}(LJ_{\theta\theta}^{1/2}C^T),\end{aligned}$$

which is easily shown to equal (19).

For posterior covariance estimation, the key result (Wood, 2006a, pp. 190–191; Wood, 2006b, p. 450) is that if (4) holds with $\varepsilon\sim N(\mathbf{0},\sigma^2 I)$ and $\hat{b}$ is given by (6), then b has posterior distribution

$$b|y\sim N[\hat{b},\sigma^2(X^TX+P)^{-1}].$$

Given a reasonably accurate estimate of Σ, these assumptions hold approximately with X, y, b, P as above, leading to the estimate

$$\widehat{\mathrm{Var}}[\mathrm{vec}(B^T)|Y]=\hat{\sigma}^2\left[\{Z\otimes(LJ_{\theta\theta}^{1/2})\}^T\{Z\otimes(LJ_{\theta\theta}^{1/2})\}+P_\Lambda\right]^{-1}, \tag{29}$$

for a residual variance estimate $\hat{\sigma}^2$. Given the hat matrix H such that $\hat{y}\equiv X\hat{b}=Hy$, we can use the standard estimate $\hat{\sigma}^2=\|y-\hat{y}\|^2/[NK-\mathrm{tr}(H)]$ (e.g., Wood, 2006a). A bit of algebra reduces (29) to (24).

It is important to note that since the residual error variance cannot be taken as $\sigma^2 I$ without prewhitening the data, the above derivation is not valid for P-OLS.

References

Antoniadis, A., and Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics and Data Analysis* **51,** 4793–4813.

Brent, R. (1973). *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Brown, P. J., and Zidek, J. V. (1980). Adaptive multivariate ridge regression. *Annals of Statistics* **8,** 64–74.

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

Bugli, C., and Lambert, P. (2006). Functional ANOVA with random functional effects: an application to event-related potentials modelling for electroencephalograms analysis. *Statistics in Medicine* **25,** 3718–3739.

Chiou, J. M., Müller, H. G., and Wang, J. L. (2003). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society, Series B* **65,** 405–423.

Chiou, J. M., Müller, H. G., and Wang, J. L. (2004). Functional response models. *Statistica Sinica* **14,** 675–693.

Crainiceanu, C. M., and Ruppert, D. (2004). Restricted likelihood ratio tests in nonparametric longitudinal models. *Statistica Sinica* **14,** 713–729.

Durban, M., and Currie, I. D. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics* **18,** 251–262.

Eubank, R. L., Huang, C., Muñoz-Maldonado, Y., Wang, N., Wang, S., and Buchanan, R. J. (2004). Smoothing spline estimation in varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **66,** 653–667.

Fan, J., and Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B* **62,** 303–322.

Golub, G. H., and Van Loan, C. F. (1996). *Matrix Computations*, 3rd ed. Baltimore: Johns Hopkins University Press.

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

Hooker, G. (2007). Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics* **16,** 709–732.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85,** 809–822.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* **14,** 763–788.

Kelly, A. M. C., Di Martino, A., Uddin, L. Q., Shehzad, Z., Gee, D. G., Reiss, P. T., Margulies, D. S., Castellanos, F. X., and Milham, M. P. (2009). Development of anterior cingulate functional connectivity from late childhood to early adulthood. *Cerebral Cortex* **19,** 640–657.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81,** 673–680.

Krafty, R. T., Gimotty, P. A., Holtz, D., Coukos, G., and Guo, W. (2008). Varying coefficient model with unknown within-subject covariance for analysis of tumor growth curves. *Biometrics* **64,** 1023–1031.

DOI: 10.2202/1557-4679.1246

Krivobokova, T., and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association* **102,** 1328–1337.

Lin, X., Wang, N., Welsh, A. H., and Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika* **91,** 177–193.

Morris, J. S., and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* **68,** 179–199.

Ogden, R. T., and Greene, E. (2010). Wavelet modeling of functional random effects with application to human vision data. *Journal of Statistical Planning and Inference* **140,** 3797–3808.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. New York: Springer.

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

Reiss, P. T., and Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society, Series B* **71,** 505–523.

Rice, J. A., and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* **53,** 233–243.

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

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

Schäfer, J., and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4,** article 32.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B* **45,** 133–150.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* **93,** 341–348.

Wood, S. N. (2006a). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall.

Wood, S. N. (2006b). On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics* **48,** 445–464.

Wood, S. N. (2010) Fast stable REML estimation of semiparametric GLMs. *Journal of the Royal Statistical Society, Series B*, to appear.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100,** 577–591.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57,** 348–368.

Zhang, P. (1993). Model selection via multifold cross validation. *Annals of Statistics* **21,** 299–313.

DOI: 10.2202/1557-4679.1246