Electronic Journal of Statistics
Vol. 12 (2018) 3554–3600
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1487

P-splines with an ℓ_1 penalty for repeated measures*

Brian D. Segal, Michael R. Elliott, Thomas Braun, and Hui Jiang

Department of Biostatistics
University of Michigan
Ann Arbor, MI
e-mail:
bdsegal@umich.edu; mrelliot@umich.edu; tombraun@umich.edu; jianghui@umich.edu

Abstract: P-splines are penalized B-splines, in which finite order differences in coefficients are typically penalized with an ℓ_2 norm. P-splines can be used for semiparametric regression and can include random effects to account for within-subject correlations. In addition to ℓ_2 penalties, ℓ_1-type penalties have been used in nonparametric and semiparametric regression to achieve greater flexibility, such as in locally adaptive regression splines, ℓ_1 trend filtering, and the fused lasso additive model. However, there has been less focus on using ℓ_1 penalties in P-splines, particularly for estimating conditional means.

In this paper, we demonstrate the potential benefits of using an ℓ_1 penalty in P-splines with an emphasis on fitting non-smooth functions. We propose an estimation procedure using the alternating direction method of multipliers and cross validation, and provide degrees of freedom and approximate confidence bands based on a ridge approximation to the ℓ_1 penalized fit. We also demonstrate potential uses through simulations and an application to electrodermal activity data collected as part of a stress study.

MSC 2010 subject classifications: Primary 62G08; secondary 62P10.
Keywords and phrases: Additive models, semiparametric regression, clustered data.

Received July 2017.

Contents

*We thank Dr. Margaret Hicken for sharing the data for the stress study example.

1. Introduction

Many nonparametric regression methods, including smoothing splines and regression splines, obtain point estimates by minimizing a penalized negative log-likelihood function of the form $l_{\text{pen}} = -l(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})$, where l is a log-likelihood, P is a penalty term, $\lambda > 0$ is a smoothing parameter, and $\boldsymbol{\beta}$ are the coefficients to be estimated. Typically, quadratic (ℓ_2 norm) penalties are used, which lead to straightforward computation and inference. In particular, ℓ_2 penalties typically lead to ridge estimators, which have both closed form solutions and are linear smoothers. The ℓ_2 penalty also has connections to mixed models, which allows the smoothing parameters to be estimated as variance components [13, 34, 44, 52].

However, nonparametric regression methods that use an ℓ_1-type penalty, such as ℓ_1 trend filtering [20] and locally adaptive regression splines [23], are better able to adapt to local differences in smoothness and achieve the minimax rate of convergence for weakly differentiable functions of bounded variation [39],

whereas ℓ_2 penalized methods do not [5]. The trade-off is that ℓ_1 penalties generally lead to more difficult computation and inference because the objective function is convex but non-differentiable, and the fit is no longer a linear smoother.

In this article, we propose P-splines with an ℓ_1 penalty as a framework for generalizing ℓ_1 trend filtering within the context of repeated measures data and semiparametric (additive) models [15]. In Section 2, we discuss the connection between P-splines and ℓ_1 trend filtering which motivates the methodological development. In Section 3, we present our proposed model, and in Section 4, we discuss related work. In Section 5, we propose an estimation procedure using the alternating direction method of multipliers (ADMM) [see 2] and cross validation (CV). In Section 6, we derive the degrees of freedom and propose computationally stable and fast approximations, and in Section 7, we develop approximate confidence bands based on a ridge approximation to the ℓ_1 fit. In Section 8, we study our method through simulations and evaluate its performance in fitting non-smooth functions. In section 9, we demonstrate our method in an application to electrodermal activity data collected as part of a stress study. We close with a discussion in Section 10.

We have implemented our method in the R package `psplinesl1`. The `psplinesl1` package and all code for the simulations and analyses in this paper are available as supplementary material [33].

2. P-splines and ℓ_1 trend filtering

In this section, we give brief background on P-splines and ℓ_1 trend filtering, and show the relation between them when the data are independent and identically distributed (i.i.d.) normal.

P-splines [8] are penalized B-splines [see 4]. B-splines are flexible bases that are notable in part because they have compact support, which leads to banded design matrices and fast computation. This compact support can be seen in Figure 1, which shows eight evenly spaced first degree and third degree B-spline bases on $[0,1]$. We can define an order M (degree $M-1$) B-spline basis with $j=1,\ldots,p$ basis functions recursively as [4]

$$\phi_j^m(x) = \frac{x-t_j}{t_{j+m-1}-t_j}\phi_j^{m-1}(x) + \frac{t_{j+m}-x}{t_{j+m}-t_{j+1}}\phi_{j+1}^{m-1}(x), \quad \begin{array}{l} j=1,\ldots,2M+c-m, \\ 1<m\le M \end{array}$$

$$\phi_j^1(x) = \begin{cases} 1 & t_j \le x < t_{j+1} \\ 0 & \text{otherwise} \end{cases}, \qquad j=1,\ldots,2M+c-1$$

where t_j are the knots, division by zero is taken to be zero, and c is the number of internal knots. For order M B-splines defined on the interval $[x_{\min}, x_{\max}]$, in order to obtain $j=1,\ldots,p$ basis functions, we set $2M$ boundary knots (M knots on each side) and $c=p-M$ interior knots. In general, one can set $t_1 \le t_2 \le \cdots \le t_M = x_{\min} < t_{M+1} < \cdots < t_{M+c} < x_{\max} = t_{M+c+1} \le t_{M+c+2} \le \cdots \le t_{2M+c}$.

In order to ensure continuity at the boundaries, we set $t_1 < t_2 < \cdots < t_{M-1} < t_M = x_{\min}$ and $x_{\max} = t_{M+c+1} < t_{M+c+2} < \cdots < t_{2M+c}$. We also use equally spaced interior knots, which is important for the P-spline penalty, and drop the superscript on ϕ designating order when the order does not matter or is stated in the text.

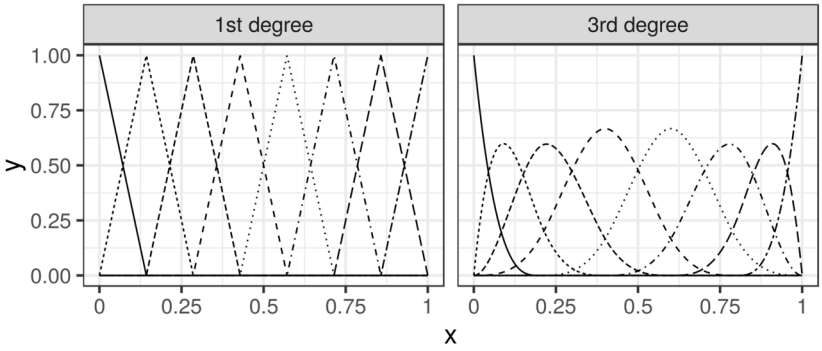


FIG 1. *Eight evenly spaced B-spline bases on* $[0, 1]$

B-spline bases can be used to fit nonparametric models of the form $y(x) = f(x) + \epsilon(x)$, where $y(x)$ is the outcome y at point x, $f(x)$ is the mean response function at x, and $\epsilon(x)$ is the error at x. To that end, let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be an $n \times 1$ vector of outcomes and $\boldsymbol{x} = (x_1 \ldots, x_n)^T$ be a corresponding $n \times 1$ vector of covariates. Also, let $\phi_1, \ldots, \phi_p$ be B-spline basis functions and let F be an $n \times p$ design matrix such that $F_{ij} = \phi_j(x_i)$, i.e., the j^{th} column of F is the j^{th} basis function evaluated at $x_1, \ldots, x_n$. Equivalently, the i^{th} row of F is the i^{th} data point evaluated by $\phi_1, \ldots, \phi_p$. For i.i.d. normal $\boldsymbol{y}$, a simple linear P-spline model with the standard ℓ_2 penalty can be written as

$$\hat{\beta}_0, \hat{\boldsymbol{\beta}} = \underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \|\boldsymbol{y} - \beta_0 \mathbf{1} - F\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2} \|D^{(k+1)}\boldsymbol{\beta}\|_2^2, \tag{1}$$

where β_0 is the intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameter estimates, $\mathbf{1}$ is an $n \times 1$ vector with each element equal to 1, $\lambda > 0$ is a smoothing parameter, and $D^{(k+1)} \in \mathbb{R}^{(p-k-1) \times p}$ is the $k+1$ order finite difference matrix. For example, for $k = 1$

$$D^{(2)} = \begin{bmatrix} 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(p-2) \times p} \tag{2}$$

In general, as described by Tibshirani [39], $D^{(k+1)} = D^{(1)} D^{(k)}$ where $D^{(1)}$ is the $(p-k-1) \times (p-k)$ upper left matrix of:

$$D^{(1)} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(p-1)\times p}. \tag{3}$$

Our proposed model builds on one in which the ℓ_2 penalty in (1) is replaced with an ℓ_1 penalty:

$$\hat{\beta}_0, \hat{\boldsymbol{\beta}} = \underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - \beta_0 \mathbf{1} - F\boldsymbol{\beta}\|_2^2 + \lambda \|D^{(k+1)}\boldsymbol{\beta}\|_1. \tag{4}$$

Letting $f(x) = \sum_{j=1}^{p} \beta_j \phi_j^M(x)$, for order $M = 4$ B-splines, Eilers and Marx [8] show that

$$\int_{x_{\min}}^{x_{\max}} \left(\frac{d^2}{dx^2} f(x)\right)^2 dx = c_1 \|D^{(2)}\boldsymbol{\beta}\|_2^2 + c_2 \sum_{j=4}^{p} \nabla^2 \beta_j \nabla^2 \beta_{j-1}$$

where ∇^2 is the second-order backwards difference and c_1 and c_2 are constants. As shown in Appendix C, a similar result holds for P-splines with an ℓ_1 penalty. In particular, for $0 \le k < M - 1$,

$$\int_{x_{\min}}^{x_{\max}} \left|\frac{d^{k+1}}{dx^{k+1}} f(x)\right| dx \le C_{M,k+1} \|D^{(k+1)}\boldsymbol{\beta}\|_1$$

where $C_{M,k+1}$ is a constant given in Appendix C that depends on the order M of the B-splines and order $k+1$ of the finite difference. In other words, controlling the ℓ_1 norm of the $(k+1)^{th}$ order finite differences in coefficients also controls the total variation of the k^{th} derivative of the function.

ℓ_1 trend filtering is similar to (4). In the case where $x_1 < x_2 < \cdots < x_n$ are unique and equally spaced, ℓ_1 trend filtering solves the following problem (the intercept is handled implicitly):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + \lambda \|D^{(k+1)}\boldsymbol{\beta}\|_1. \tag{5}$$

Problem (5) differs from (4) in that (5) has one parameter per data point, and the design matrix is the identity matrix. $D^{(k+1)}$ is also resized appropriately by replacing p with n in the dimensions of (2) and (3). However, under certain conditions noted in Observation 1, (4) and (5) are identical.

Observation 1 (Continuous representation)**.** *For second order (first degree) B-splines with n basis functions, equally spaced data $x_1 < x_2 < \cdots < x_n$ with knots at $t_1 < x_1, t_2 = x_1, t_3 = x_2, \ldots, t_n = x_{n-1}, t_{n+1} = x_n, t_{n+2} > x_n$, and centered outcomes such that $y(0) = 0$, P-splines with an ℓ_1 penalty are a continuous analogue to ℓ_1 trend filtering.*

Proof of Observation 1. Under these conditions, for $i = 1, \ldots, n$

$$\phi_j^2(x_i) = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}.$$

To see this, note that

$$\begin{aligned}\phi_j^2(x_i) &= \frac{x_i - t_j}{t_{j+1} - t_j}\phi_j^1(x_i) + \frac{t_{j+2} - x_i}{t_{j+2} - t_{j+1}}\phi_{j+1}^1(x_i) \\ &= \frac{t_{i+1} - t_j}{t_{j+1} - t_j}\phi_j^1(t_{i+1}) + \frac{t_{j+2} - t_{i+1}}{t_{j+2} - t_{j+1}}\phi_{j+1}^1(t_{i+1}). \end{aligned} \tag{6}$$

Now,

$$\phi_j^1(t_{i+1}) = \begin{cases} 1 & t_j \le t_{i+1} < t_{j+1} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \phi_{j+1}^1(t_{i+1}) = \begin{cases} 1 & t_{j+1} \le t_{i+1} < t_{j+2} \\ 0 & \text{otherwise} \end{cases}.$$

We have $\phi_j^1(t_{i+1}) = 1$ for $i = j-1$ and 0 otherwise, but for $i = j-1$, we have $t_{i+1} - t_j = t_j - t_j = 0$. We also have $\phi_{j+1}^1(t_{i+1}) = 1$ for $i = j$ and 0 otherwise, and for $i = j$, we have $t_{j+2} - t_{i+1} = t_{j+2} - t_{j+1} > 0$. It follows that for $i = 1 \ldots, n$, (6) evaluates to 1 if $i = j$ and 0 otherwise.

Let F be the design matrix in (4), where $F_{ij} = \phi_j^2(x_i)$. Then from the previous result, we have $F = I_n$, where I_n is the $n \times n$ identity matrix. This, together with the assumption that $\beta_0 = y(0) = 0$, implies that the objective functions (4) and (5) are identical, which proves Observation 1. □

We note that Tibshirani [39] shows that ℓ_1 trend filtering has a continuous representation when expressed in the standard lasso form, and Observation 1 gives a continuous representation of ℓ_1 trend filtering when expressed in generalized lasso form.

ℓ_1 trend filtering can be applied to irregularly spaced data, such as with the algorithm developed by Ramdas and Tibshirani [26]. It might also be possible to extend ℓ_1 trend filtering to repeated measures data to account for within-subject correlations. However, due to Observation 1, we think it is beneficial to view ℓ_1 trend filtering as a special case of P-splines with an ℓ_1 penalty. We think this approach has the potential to be a general framework, because higher order B-splines could be used in combination with different order difference matrices just as can be done with P-splines that use the standard ℓ_2 penalty. Furthermore, expressing ℓ_1 trend filtering as P-splines with an ℓ_1 penalty may facilitate the development of confidence bands (see Section 7), which could help to fill a gap in the ℓ_1 penalized regression literature.

In addition, there are connections between P-splines with an ℓ_1 penalty and locally adaptive regression splines. In particular, as Tibshirani [39] shows, the continuous analogue of ℓ_1 trend filtering is identical to locally adaptive regression splines [23] for $k = 0, 1$, and asymptotically equivalent for $k \ge 2$.

3. Proposed model: Additive mixed model using P-splines with an ℓ_1 penalty

To introduce our model, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ be an $n_i \times 1$ vector of responses for subject $i = 1, \ldots, N$, and let $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T)^T$ be the stacked $n \times 1$ vector

of responses for all N subjects, where $n = \sum_{i=1}^{N} n_i$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in_i})^T$ be a corresponding $n_i \times 1$ vector of covariates for subject i, and $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T)^T$ be the $n \times 1$ stacked vector of all covariate values. In many contexts, x is time. To account for the within-subject correlations of $\boldsymbol{y}_i$, we can incorporate random effects into the P-spline model. To that end, let Z_i be an $n_i \times q_i$ design matrix for the random effects for subject i (possibly including a B-spline basis), and let $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq_i})^T$ be the corresponding $q_i \times 1$ vector of random effect coefficients for subject i. Also, let

$$Z = \begin{bmatrix} Z_1 & & \\ & \ddots & \\ & & Z_N \end{bmatrix}$$

be the $n \times q$ block diagonal random effects design matrix for all subjects, where $q = \sum_{i=1}^{N} q_i$, and let $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_N^T)^T$ be the $q \times 1$ stacked vector of random effects for all subjects. We propose an additive mixed model with $j = 1, \ldots, J$ smooths (tildes denote quantities that will be subject to additional constraints, as described below):

$$\begin{aligned} \underset{\beta_0 \in \mathbb{R}, \boldsymbol{b} \in \mathbb{R}^q, \tilde{\boldsymbol{\beta}}_j \in \mathbb{R}^{p_j}, j=1,\ldots,J}{\text{minimize}} \quad & \frac{1}{2} \|\boldsymbol{y} - \beta_0 \mathbf{1} - \sum_{j=1}^{J} \tilde{F}_j \tilde{\boldsymbol{\beta}}_j - Z\boldsymbol{b}\|_2^2 + \sum_{j=1}^{J} \lambda_j \|\tilde{D}_j^{(k_j+1)} \tilde{\boldsymbol{\beta}}_j\|_1 \\ & + \tau \frac{1}{2} \boldsymbol{b}^T S \boldsymbol{b} \end{aligned} \tag{7}$$

where $\tilde{F}_j$ is a $n \times p_j$ design matrix of B-spline bases for smooth j, $\tilde{D}_j^{(k_j+1)}$ is the $k_j + 1$ finite difference matrix, and $\sigma_b^2 S$ is the covariance matrix of the random effects $\boldsymbol{b}$. For example, if $\boldsymbol{b}$ are random intercepts, then $S = I_N$ and Z would be an $n \times N$ matrix such that $Z_{il} = 1$ if observation i belonged to subject l and zero otherwise. Alternatively, to obtain random curves using smoothing splines and a B-spline basis, we could set

$$S = \begin{bmatrix} S_1 & & \\ & \ddots & \\ & & S_N \end{bmatrix}$$

where $S_{j,il} = \int \phi''_{ji}(t)\phi''_{jl}(t)dt$, and $\phi''_{j1}, \ldots, \phi''_{jp_j}$ are the second derivatives of the B-spline basis functions for the j^{th} smooth. We would then set Z to be the corresponding B-splines evaluated at the input points.

We note that (7) includes varying-coefficient models [17]. For example, as pointed out by Wood [47, p. 169], if $\tilde{F}_1$ are B-splines evaluated at $\boldsymbol{x}$, we could have $\tilde{F}_2 = \text{diag}(\boldsymbol{x'})\tilde{F}_1$, where $\boldsymbol{x'} \neq \boldsymbol{x}$ is another covariate vector and $\text{diag}(\boldsymbol{x'})$ is a diagonal matrix with x'_i at the i^{th} leading diagonal position.

As written, (7) is not generally identifiable. To see this, suppose $\hat{y}(x) = \hat{\beta}_0 + \hat{f}_1(x) + \hat{f}_2(x)$, where neither f_1 nor f_2 are varying-coefficient terms. Then letting $\hat{f}'_1(x) = \hat{f}_1(x) + \delta$ and $\hat{f}'_2(x) = \hat{f}_2(x) - \delta$ for $\delta \in \mathbb{R}$, we also have $\hat{y}(x) =$

$\hat{\beta}_0 + \hat{f}_1'(x) + \hat{f}_2'(x)$. To make (7) identifiable, we follow Wood [47, Section 4.2] and introduce a centering constraint on each non-varying coefficient smooth, i.e. $\int \hat{f}_j(x)dx = 0$ for all smooths $j = 1, \ldots, J$ such that $\tilde{F}_j \neq \text{diag}(\boldsymbol{x}')\tilde{F}_l$ for some $\boldsymbol{x}'$ and $l \neq j$. To this end, let $\mathcal{E} = \{j \in \{1, \ldots, J\} : \tilde{F}_j \neq \text{diag}(\boldsymbol{x}')\tilde{F}_l \text{ for some } \boldsymbol{x}', l \neq j\}$ be the indices of the non-varying coefficient smooths, and let $\bar{\mathcal{E}} = \{j \in \{1, \ldots, J\} : j \notin \mathcal{E}\}$ be its complement. We constrain $\mathbf{1}^T\tilde{F}_j\tilde{\boldsymbol{\beta}}_j = 0$ for $j \in \mathcal{E}$. We accomplish this by defining new $p_j \times (p_j - 1)$ orthonormal matrices Q_j, $j = 1, \ldots, J$, such that $\mathbf{1}^T\tilde{F}_jQ_j = \mathbf{0}$. If desired, one can also define a $q \times (q-1)$ matrix Q_{J+1} such that $\mathbf{1}^TZQ_{J+1} = \mathbf{0}$. However, this last centering constraint is not necessary, because the penalty on the random effect terms pulls the coefficients themselves towards zero, as opposed to the finite order differences in coefficients.

As Wood [47, Section 1.8.1] shows, Q can be obtained by taking the QR decomposition of $\tilde{F}_j^T\mathbf{1}$ and retaining the last $p_j - 1$ columns of the left orthonormal matrix.[1] We can then re-parameterize the p_j constrained parameters $\tilde{\boldsymbol{\beta}}_j$ in terms of the $p_j - 1$ unconstrained parameters $\boldsymbol{\beta}_j$, such that $\tilde{\boldsymbol{\beta}}_j = Q_j\boldsymbol{\beta}_j$. For $j \in \mathcal{E}$, let $F_j = \tilde{F}_jQ_j$ and $D_j = \tilde{D}_j^{(k_j+1)}Q_j$. For $j \in \bar{\mathcal{E}}$, let $F_j = \tilde{F}_j$ and $D_j = \tilde{D}_j^{(k_j+1)}$. If centering the random effects, then we redefine $S := Q_{J+1}^TSQ_{J+1}$ and $Z := ZQ_{J+1}$. Then we can re-write (7) in the identifiable form

$$\underset{\beta_0 \in \mathbb{R}, \boldsymbol{b} \in \mathbb{R}^q, \boldsymbol{\beta}_j \in \mathbb{R}^{p_j'}, j=1,\ldots,J}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \sum_{j=1}^{J} F_j\boldsymbol{\beta}_j - Z\boldsymbol{b}\|_2^2 + \sum_{j=1}^{J}\lambda_j\|D_j\boldsymbol{\beta}_j\|_1 + \tau\frac{1}{2}\boldsymbol{b}^TS\boldsymbol{b} \tag{8}$$

where $p_j' = p_j - 1$ for $j \in \mathcal{E}$ and $p_j' = p_j$ for $j \in \bar{\mathcal{E}}$.

We note that the penalty matrix S given above for random subject-specific splines defines non-zero correlation between nearby within-subject random effect coefficients. This is in contrast to the approach of Ruppert, Wand and Carroll [30] for estimating subject-specific random curves, which focuses on the case in which nearby within-subject coefficients are not correlated. To see this, let $\hat{d}_i(x) = \sum_{j=1}^{q_i}\hat{b}_{ij}\phi_{ij}(x)$ be the estimated difference between the i^{th} subject-specific curve and the marginal mean at point x. The smoothing spline approach above constrains $\int(\hat{d}'')^2(x)dx = \boldsymbol{b}_i^TS_i\boldsymbol{b}_i < C$ for some constant $C > 0$, whereas the approach of Ruppert, Wand and Carroll [30] constrains $\boldsymbol{b}_i^TI_{q_i}\boldsymbol{b}_i = \sum_{j=1}^{q_i}\hat{b}_j^2 < C$. Whereas the non-diagonal penalty matrix S implies correlations between nearby coefficients, the identity matrix in the approach of Ruppert, Wand and Carroll [30] implies zero correlation.

Similar to the equivalence between Bayesian models and ℓ_2 penalized smoothing splines [43], there is an equivalence between Bayesian models and ℓ_1 penalized splines. In particular, (8) is equivalent to the following distributional

[1] The matrices $\mathbf{1}^T\tilde{F}_j$, $j = 1, \ldots, J$ are of rank 1, so the remaining $p_j - 2$ columns are arbitrary orthonormal vectors. In R [37], when taking the QR decomposition of $\tilde{F}^T\mathbf{1}$, an appropriate matrix Q can be obtained as `Q <- qr.Q(qr(colSums(F_tilde)), complete = TRUE)[, -1]`.

assumptions, which we can use to obtain Bayesian estimates:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{b} &= \beta_0\mathbf{1} + \sum_{j=1}^{J} F_j\boldsymbol{\beta}_j + Z\boldsymbol{b} + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim N\left(\mathbf{0}, \sigma_\epsilon^2 I_n\right) \\
\boldsymbol{b} &\sim N(\mathbf{0}, \sigma_b^2 S^{-1}) \text{ for } \sigma_b^2 = \sigma_\epsilon^2/\tau \\
\boldsymbol{\epsilon} &\perp \boldsymbol{b} \\
(D_j\boldsymbol{\beta}_j)_l &\sim \text{Laplace}(0, a_j) \text{ for } a_j = \sigma_\epsilon^2/(2\lambda_j), l = 1, \ldots, p_j - k_j - 1, j = 1, \ldots, J
\end{aligned} \tag{9}$$

The last distributional assumption is an element-wise Laplace prior on the k_j+1 order differences in coefficients.

In some cases, the random effects penalty matrix S may be positive semidefinite but not invertible. For example, the smoothing spline random curves outlined above lead to a penalty matrix S that is not strictly positive definite, but that is still positive semidefinite. This does not cause problems for the ADMM algorithm, but some changes are required for other algorithms as well as for Bayesian estimation. Following Wood [47, Section 6.6.1], let $S = U\Lambda U^T$ be the eigendecomposition of a positive semidefinite matrix S, where $UU^T = I_q$ and Λ is a diagonal matrix with eigenvalues in descending order in the diagonal positions. Let $\breve{\boldsymbol{b}} = U^T\boldsymbol{b}$ and $\breve{Z} = ZU$, so that $\boldsymbol{b}^T S\boldsymbol{b} = \breve{\boldsymbol{b}}^T\Lambda\breve{\boldsymbol{b}}$ and $\breve{Z}\breve{\boldsymbol{b}} = Z\boldsymbol{b}$. Let q_r be the number of strictly positive eigenvalues of S, where $0 < q_r < q$, and let Λ_r be the $q_r \times q_r$ upper left portion of Λ. We can partition $\breve{\boldsymbol{b}}$ as $\breve{\boldsymbol{b}} = (\breve{\boldsymbol{b}}_r^T, \breve{\boldsymbol{b}}_f^T)^T$, where $\breve{\boldsymbol{b}}_r^T$ is a $q_r \times 1$ vector of penalized coefficients and $\breve{\boldsymbol{b}}_f^T$ is a $q_f \times 1$ vector of unpenalized coefficients, where $q_r + q_f = q$. Then $\breve{\boldsymbol{b}}^T\Lambda\breve{\boldsymbol{b}} = \breve{\boldsymbol{b}}_r^T\Lambda_r\breve{\boldsymbol{b}}_r$, and it follows that $\breve{\boldsymbol{b}}_r \sim N(\mathbf{0}, \sigma_b^2\Lambda_r^{-1})$ and $\breve{\boldsymbol{b}}_f \propto \mathbf{1}$.

However, allowing for unconstrained random effect parameters leads to identifiability issues. Therefore, in practice if $q_f > 0$, we recommend using a normal or Cauchy prior on $\breve{\boldsymbol{b}}_f$. In particular, $\breve{b}_{f,l} \sim N(0, \sigma_f)$ or $\breve{b}_{f,l} \sim \text{Cauchy}(0, \sigma_f)$, $l = 1, \ldots, q_f$ with either a diffuse prior on σ_f and constraints to ensure $\sigma_f > 0$, or a diffuse prior on $\log(\sigma_f)$ without constraints. The Cauchy prior may be a preferable first choice, as it provides a weaker penalty and is similar to the recommendations of Gelman et al. [11] for logistic regression. However, in some cases, such as in Section 9, it is necessary to use a normal prior.

To further improve the computational efficiency of Monte Carlo sampling methods, we can partition $\breve{Z}$ into $\breve{Z} = [\breve{Z}_r, \breve{Z}_f]$ where $\breve{Z}_r$ contains the first q_r columns of $\breve{Z}$ and $\breve{Z}_f$ contains the remaining q_f columns. We then set $\check{\boldsymbol{b}}_r = \Lambda_r^{-1/2}\breve{\boldsymbol{b}}_r$ and $\check{Z}_r = \breve{Z}_r\Lambda_r^{1/2}$, so that $\check{Z}_r\check{\boldsymbol{b}}_r = \breve{Z}_r\breve{\boldsymbol{b}}_r$ and $\check{\boldsymbol{b}}_r \sim N(\mathbf{0}, \sigma_b^2 I)$, which allows for more efficient sampling [47].

4. Related work

There are many nonparametric and semiparametric methods for analyzing repeated measures data. For an overview, please see Fitzmaurice et al. [10, Part III]. However, most existing methods use an ℓ_2 penalty [e.g. 28, 14, 3, 32].

Focusing on the optimization problem, our method puts a generalized lasso penalty [38] on the fixed effects and a quadratic penalty on the random effects. Unlike the elastic net [54], we do not mix the ℓ_1 and ℓ_2 penalties on the same parameters, though this could be done in the future.

The additive model with trend filtering developed by Sadhanala and Tibshirani [31] is similar to our approach. Sadhanala and Tibshirani [31] optimize

$$\begin{aligned} &\underset{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\bar{y}\mathbf{1}-\sum_{j=1}^{J}\boldsymbol{\theta}_j\|_2^2+\lambda\sum_{j=1}^{J}\|D^{(k+1)}\boldsymbol{\theta}_j\|_1 \\ &\text{subject to} \quad \mathbf{1}^T\boldsymbol{\theta}_j=\mathbf{0}, j=1,\ldots,J. \end{aligned} \tag{10}$$

In contrast to (8), (10) has one smoothing parameter and constrains all smooths to be zero-centered. From Observation 1, we see that (10) is equivalent to (8) when there is $J=1$ smooth and no random effects, in which case there would be only one smoothing parameter λ and no varying-coefficient smooths.

Sadhanala and Tibshirani [31] develop the theoretical and computational aspects of additive models with trend filtering, including the extension of the falling factorial basis to additive models. Similar to the B-spline basis, the falling factorial basis allows for linear time multiplication and inversion, which leads to fast computation [45].

When smooths $j=1,\ldots,J$ are expected to have similar degrees of freedom and n is not large enough to require dimension reduction, then (10) with the addition of random effects and the relaxation of the zero-constraints for varying-coefficient smooths may be a viable alternative to (8) that could potentially adapt better to local differences in smoothness because it would have one knot per data point.

While not developed for analyzing repeated measures, the fused lasso additive model (FLAM) [25] is also similar to (8). FLAM optimizes the following problem:

$$\underset{\theta_0\in\mathbb{R},\boldsymbol{\theta}_j\in\mathbb{R}^n,1\le j\le J}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\theta_0\mathbf{1}-\sum_{j=1}^{J}\boldsymbol{\theta}_j\|_2^2+\alpha\lambda\sum_{j=1}^{J}\|D^{(1)}\boldsymbol{\theta}_j\|_1+(1-\alpha)\lambda\sum_{j=1}^{J}\|\boldsymbol{\theta}_j\|_2 \tag{11}$$

where $0\le\alpha\le 1$ specifies the balance between fitting piecewise constant functions ($\alpha=1$) and inducing sparsity on the selected smooths ($\alpha=0$). From Observation 1, we see that (11) is equivalent to our model (8) when: $\alpha=1$, there is $J=1$ smooth, our design matrix has $p=n$ columns, the B-spline bases have appropriately chosen knots, and our model has no random effects. As Petersen, Witten and Simon [25] show, FLAM can be a very useful method for modeling additive phenomenon, and as with the fused lasso [42], jumps in the piecewise linear fits have the advantage of being interpretable.

We also mention the sparse additive model (SpAM) [27] and sparse partially linear additive model (SPLAM) [22]. SpAM fits an additive model and uses a group lasso penalty [51] to induce sparsity on the number of active smooths. SPLAM fits a partially linear additive model and uses a hierarchical group lasso

penalty [53] to induce sparsity in the selected predictors and to control the number of nonlinear features.

One notable difference between our model and that of Sadhanala and Tibshirani [31], as well as FLAM, SpAM, and SPLAM, is that we allow for multiple smoothing parameters. In our applied experience with additive models and standard ℓ_2 penalties, we have found that in practice it can be important to allow for multiple smoothing parameters, particularly when the quantities of interest are the individual smooths as opposed to the overall prediction. This is equivalent to allowing each smooth to have different variance. However, this flexibility comes at a cost: estimating multiple smoothing parameters is currently the greatest challenge in fitting our proposed model. Perhaps due in part to these computational difficulties, several other authors also assume a single smoothing parameter in high-dimensional additive models [e.g. 21, 24].

There are fast and stable methods for fitting multiple smoothing parameters for ℓ_2 penalties paired with exponential family and quasilikelihood loss functions, notably the work of Wood [46] using generalized cross validation (GCV) and Wood [48] using restricted maximum likelihood. Furthermore, Wood, Goude and Shaw [49] extend these methods to larger datasets, and Wood, Pya and Säfken [50] extend these methods to likelihoods outside the exponential family and quasilikelihood form. However, similarly computationally efficient methods do not yet exist for fitting multiple smoothing parameters for ℓ_1 penalties.

In addition to allowing for multiple smoothing parameters, we also propose approximate inferential methods, which is not typically provided for ℓ_1 penalized models. Yuan and Lin [51], Ravikumar et al. [27], Lou et al. [22], and Petersen, Witten and Simon [25] focus on prediction and provide bounds on the prediction risk and related quantities. These are important results, and we think that distributional results for individual parameters and smooths will also be useful to practitioners.

We also note that Eilers [7] and Bollaerts, Eilers and Aerts [1] discuss a variant of P-splines for quantile regression, in which the ℓ_1 norm is used in both the loss and penalty function. However, we are not aware of existing P-spline methods that combine an ℓ_1 penalty with an ℓ_2 loss function.

5. Point estimation

5.1. Regression parameters and random effects

To fit (8), we use the alternating direction method of multipliers (ADMM) [see 2]. ADMM has the advantage of being scalable to large datasets. To formulate (8) for ADMM, we introduce constraint terms $\boldsymbol{w}_j$ and re-write the optimization problem as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \sum_{j=1}^{J} F_j\boldsymbol{\beta}_j - Z\boldsymbol{b}\|_2^2 + \sum_{j=1}^{J}\lambda_j\|\boldsymbol{w}_j\|_1 + \frac{\tau}{2}\boldsymbol{b}^T S\boldsymbol{b} \qquad (12) \\ \text{subject to} \quad & D_j\boldsymbol{\beta}_j - \boldsymbol{w}_j = \mathbf{0},\ j = 1,\ldots,J \end{aligned}$$

The augmented Lagrangian in scaled form (using $\boldsymbol{u}$ to denote the scaled dual variable) is

$$L_\rho(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{w}, \boldsymbol{u}) \propto \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \sum_j F_j\boldsymbol{\beta}_j - Z\boldsymbol{b}\|_2^2 + \sum_j \lambda_j \|\boldsymbol{w}_j\|_1 + \frac{\rho}{2}\sum_j \|D_j\boldsymbol{\beta}_j - \boldsymbol{w}_j + \boldsymbol{u}_j\|_2^2 + \frac{\tau}{2}\boldsymbol{b}^T S\boldsymbol{b}$$

where $\rho > 0$ is the penalty parameter. The dimensions are $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$, $\beta_0 \in \mathbb{R}$, $F_j \in \mathbb{R}^{n\times p_j'}$, $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j'\times 1}$, $Z \in \mathbb{R}^{n\times q}$, $\boldsymbol{b} \in \mathbb{R}^{q\times 1}$, $D_j \in \mathbb{R}^{(p_j-k_j-1)\times p_j'}$, $\boldsymbol{w}_j \in \mathbb{R}^{(p_j-k_j-1)\times 1}$, $\boldsymbol{u}_j \in \mathbb{R}^{(p_j-k_j-1)\times 1}$, and $S \in \mathbb{R}^{q\times q}$, where $p_j' = p_j - 1$ if $j \in \mathcal{E}$ (non-varying coefficient smooths) and $p_j' = p_j$ if $j \in \bar{\mathcal{E}}$ (varying coefficient smooths).

ADMM is an iterative algorithm, and we re-estimate the parameters for updates $m = 1, 2, \ldots$ until convergence.[2] It is straightforward to derive the $m+1$ updates [see 2, Section 6.4.1]:

$$\begin{aligned}
\beta_0^{m+1} &= \frac{1}{n}\mathbf{1}^T\left(\boldsymbol{y} - \sum_j F_j\boldsymbol{\beta}_j^m - Z\boldsymbol{b}^m\right)\\
\boldsymbol{\beta}_j^{m+1} &:= \arg\min_{\boldsymbol{\beta}_j} L_\rho(\beta_0^{m+1}, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{l<j}^{m+1}, \boldsymbol{\beta}_{l>j}^m, \boldsymbol{b}^m, \boldsymbol{w}^m, \boldsymbol{u}^m)\\
&= \left(F_j^TF_j + \rho D_j^TD_j\right)^{-1}\left(F_j^T\boldsymbol{y}^{(j,m)} + \rho D_j^T(\boldsymbol{w}_j^m - \boldsymbol{u}_j^m)\right)\\
\boldsymbol{b}^{m+1} &:= \arg\min_{\boldsymbol{b}} L_\rho(\boldsymbol{\beta}_{j=1,\ldots,J}^{m+1}, \boldsymbol{b}, \boldsymbol{w}^m, \boldsymbol{u}^m)\\
&= (Z^TZ + \tau S)^{-1}Z^T(\boldsymbol{y} - \beta_0^{m+1}\mathbf{1} - \sum_j F_j\boldsymbol{\beta}_j^{m+1})\\
\boldsymbol{w}_j^{m+1} &:= \arg\min_{\boldsymbol{w}_j} L_\rho(\boldsymbol{\beta}_{j=1,\ldots,J}^{m+1}, \boldsymbol{b}^{m+1}, \boldsymbol{w}_j, \boldsymbol{u}^m)\\
&= \psi_{\lambda_j/\rho}(D_j\boldsymbol{\beta}_j^{m+1} + \boldsymbol{u}_j^m)\\
\boldsymbol{u}_j^{m+1} &:= \boldsymbol{u}_j^m + D_j\boldsymbol{\beta}_j^{m+1} - \boldsymbol{w}_j^{m+1}
\end{aligned} \tag{13}$$

where $\boldsymbol{y}^{(j,m)} = \boldsymbol{y} - \beta_0^{m+1}\mathbf{1} - \sum_{l<j} F_l\boldsymbol{\beta}_l^{m+1} - \sum_{l>j} F_j\boldsymbol{\beta}_l^m - Z\boldsymbol{b}^m$ and $\psi_{\lambda/\rho}$ is the element-wise soft thresholding operator, where for a single scalar element x

$$\psi_{\lambda/\rho}(x) = \begin{cases} x - \lambda/\rho & x > \lambda/\rho \\ 0 & |x| \le \lambda/\rho \\ x + \lambda/\rho & x < -\lambda/\rho \end{cases}$$

To initialize the algorithm, we set $\beta_0 := \bar{y}$, $\boldsymbol{b} := \mathbf{0}$, and $\boldsymbol{\beta}_j := \mathbf{0}$, $\boldsymbol{w}_j := \mathbf{0}$, and $\boldsymbol{u}_j := \mathbf{0}$, for $j = 1, \ldots, J$.

[2]We use m to denote the iteration of the ADMM algorithm. This is unrelated to our use of m in Section 2 to denote the order of the B-spline basis.

As an alternative to the closed-form update (13) for the random effects, it is also possible to update the random effects via a linear mixed effects (LME) model that is embedded into the ADMM algorithm. In particular, an LME model is fit to the residuals $\boldsymbol{y} - \beta_0^{m+1}\mathbf{1} - \sum_j F_j \boldsymbol{\beta}_j^{m+1}$, and $\boldsymbol{b}^{m+1}$ are updated as the best linear unbiased predictors (BLUPs). This update occurs at each step of the ADMM algorithm and replaces the update given by (13). The LME update has the additional benefit of simultaneously estimating the variance of the random effects σ_b^2. In simulations, we have found that using an LME update leads to more accurate estimates of σ_b^2, which is important for subsequent estimates of degrees of freedom and confidence intervals.

For stopping criteria, we use the primal and dual residuals (r^m and s^m, respectively):

$$r^m = \begin{bmatrix} D_1\boldsymbol{\beta}_1^m - \boldsymbol{w}_1^m \\ \vdots \\ D_J\boldsymbol{\beta}_J^m - \boldsymbol{w}_J^m \end{bmatrix} \in \mathbb{R}^{(p-k-J)\times 1}$$

$$s^m = -\rho \begin{bmatrix} D_1^T\left(\boldsymbol{w}_1^m - \boldsymbol{w}_1^{m-1}\right) \\ \vdots \\ D_J^T\left(\boldsymbol{w}_J^m - \boldsymbol{w}_J^{m-1}\right) \end{bmatrix} \in \mathbb{R}^{p\times 1}$$

where $k = \sum_{j=1}^J k_j$, $p = \sum_{j=1}^J p_j - |\mathcal{E}|$, and $|\mathcal{E}|$ is the cardinality of $\mathcal{E}$.

Following the guidance of [2], we stop when $\|r^m\|_2 \leq \epsilon^{\text{pri}}$ and $\|s^m\|_2 \leq \epsilon^{\text{dual}}$, where

$$\epsilon^{\text{pri}} = \epsilon^{\text{abs}}\sqrt{p-k-J} + \epsilon^{\text{rel}} \max\left\{ \left\| \begin{matrix} D_1\boldsymbol{\beta}_1^m \\ \vdots \\ D_J\boldsymbol{\beta}_J^m \end{matrix} \right\|_2, \left\| \begin{matrix} \boldsymbol{w}_1^m \\ \vdots \\ \boldsymbol{w}_J^m \end{matrix} \right\|_2 \right\}$$

$$\epsilon^{\text{dual}} = \epsilon^{\text{abs}}\sqrt{p} + \epsilon^{\text{rel}}\rho \left\| \begin{matrix} D_1^T\boldsymbol{u}_1^m \\ \vdots \\ D_J^T\boldsymbol{u}_J^m \end{matrix} \right\|_2.$$

By default, we set $\epsilon^{\text{rel}} = \epsilon^{\text{abs}} = 10^{-4}$ and the maximum number of iterations at $1,000$.

5.2. *Smoothing parameters*

To estimate $\lambda_1, \ldots, \lambda_J$ we compute cross validation (CV) error for a path of values one smoothing parameter at a time. In the CV, we split the sample at the subject level, as opposed to individual observations, and ensure that there are at least two subjects in each fold per unique combination of factor covariates. First, we estimate a path for τ with $\lambda_1, \ldots, \lambda_J$ set to 0. Then we fix τ at the value that minimizes CV error and compute a path for λ_1, setting it to the value that minimizes CV error, and so on.

We fit a path for each λ_j from $\lambda_j^{\max}$ to $10^{-5}\lambda_j^{\max}$ evenly spaced on the log scale, where $\lambda_j^{\max}$ is the smallest value at which $D_j\boldsymbol{\beta}_j = \mathbf{0}$. As shown in Ap-

pendix B, $\lambda_j^{\max} = \|(D_j D_j^T)^{-1} D_j (F_j^T F_j)^{-1} F_j^T \boldsymbol{r}_j\|_\infty$, where $\boldsymbol{r}_j = \boldsymbol{y} - \beta_0 \mathbf{1} - \sum_{\ell \neq j} F_\ell \boldsymbol{\beta}_\ell - Z\boldsymbol{b}$ are the j^{th} partial residuals and for a vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|_\infty = \max_j |a_j|$.

We also use warm starts, passing starting values separately for each fold, though warm starts appear to be minimally beneficial with ADMM. We set $\rho = \min(\max(\lambda_1, \ldots, \lambda_J), c)$ at each iteration for some constant $c > 0$ (e.g. $c = 5$). When the number of smooths J is small (e.g. $J \leq 2$) a grid search is also feasible.

To estimate τ, we can either use CV and the close-form update given by (13), or an LME update that is embedded in the ADMM algorithm, as described in Section 5.1. In simulations, we have found that the overall computation time to estimate the smoothing parameters is greater when using the LME update, and that the estimates of $\lambda_1, \ldots, \lambda_J$ do not appear sensitive to updates for $\boldsymbol{b}$. However, the final estimates of σ_b^2, and consequently the width of confidence intervals can be improved by using the LME update. Consequently, we recommend using cross validation to estimate τ for the purposes of then estimating $\lambda_1, \ldots, \lambda_J$, but using an LME update when estimating the final model.

With both the closed-form and LME update, we cannot use the training sample to estimate the random effect parameters $\boldsymbol{b}$ for the test sample, because these parameters are subject-specific and the test subjects are not included in the training sample. Instead, we use the training sample to obtain estimates for the fixed effect parameters β_0, $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$ and then use the test sample to estimate the random effects.

To make our approach clear, we first fix notation. Let $\mathcal{T}^r \subseteq \{1, \ldots, n\}$ be the row indices for the observations in the test sample for both the fixed and random effect design matrices F_j, $j = 1, \ldots, J$, and Z. Also, let $\mathcal{T}^c \subseteq \{1, \ldots, q\}$ be the column indices of Z for observations in the test sample, and let $\mathcal{T} = (\mathcal{T}^r, \mathcal{T}^c)$ be the tuple of row and column indices designating the test sample. Let matrices $F_{j,\mathcal{T}}$ and $F_{j,-\mathcal{T}}$ be matrix F_j with only rows indexed by $\mathcal{T}^r$ retained and removed, respectively. Similarly, let matrices $Z_\mathcal{T}$ and $Z_{-\mathcal{T}}$ be matrix Z with only rows and columns indexed by $\mathcal{T}^r$ and $\mathcal{T}^c$, respectively, retained and removed, respectively. Let matrices $S_\mathcal{T}$ and $S_{-\mathcal{T}}$ be matrix S with only rows and columns indexed by $\mathcal{T}^c$ retained and removed, respectively. Also, let $\boldsymbol{y}_\mathcal{T}$ and $\boldsymbol{y}_{-\mathcal{T}}$ be vector $\boldsymbol{y}$ with elements indexed by $\mathcal{T}^r$ retained and removed, respectively.

We obtain out-of-sample marginal estimates as $\hat{\boldsymbol{\mu}}_\mathcal{T} = \hat{\beta}_0 \mathbf{1} + \sum_{j=1}^J F_{j,\mathcal{T}} \hat{\boldsymbol{\beta}}_j$, where $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_j$, $j = 1, \ldots, J$ are estimated with $\boldsymbol{y}_{-\mathcal{T}}$, $F_{j,-\mathcal{T}}$, and $Z_{-\mathcal{T}}$. If using the closed-form update (13), we estimate subject-specific random effects as $\hat{\boldsymbol{b}}_\mathcal{T} = \left(Z_\mathcal{T}^T Z_\mathcal{T}^T + \tau S_\mathcal{T}\right)^{-1} Z_\mathcal{T}^T (\boldsymbol{y}_\mathcal{T} - \hat{\boldsymbol{\mu}}_\mathcal{T})$ and obtain the out-of-sample prediction residuals as $\boldsymbol{r}_\mathcal{T} = \boldsymbol{y}_\mathcal{T} - \hat{\boldsymbol{\mu}}_\mathcal{T} - Z_\mathcal{T} \hat{\boldsymbol{b}}_\mathcal{T}$. Letting $\mathcal{T}_k$ be the tuple of indices for test sample (fold) $k = 1, \ldots, K$, we obtain the CV error as $\sum_{k=1}^K \|\boldsymbol{r}_{\mathcal{T}_k}\|_2^2$.

6. Degrees of freedom

In this section, we obtain the degrees of freedom, with the primary goal of estimating variance (see Section 7.1). However, we note that the degrees of

freedom does not always align with a model's complexity in terms of its tendency to overfit the data [19].

In each of the approaches described in this section, the degrees of freedom (df) is a function of the smoothing parameters $\lambda_1, \dots \lambda_J$ and τ. We always obtain the fixed effects smoothing parameters $\lambda_1, \dots, \lambda_J$ from CV, but when using an LME update for the random effects $\boldsymbol{b}$ as described in Sections 5.1 and 5.2, we do not directly obtain τ. Consequently, we cannot directly apply the results in this section to estimate df. However, from (9), we have that $\tau = \sigma_b^2/\sigma_\epsilon^2$. Writing $\mathrm{df} = \mathrm{df}(\tau)$, and letting $\boldsymbol{r} = \boldsymbol{y} - \sum_{j=1}^{J} F_j \hat{\boldsymbol{\beta}}_j - Z\hat{\boldsymbol{b}}$ be an $n \times 1$ vector of residuals and $\hat{\sigma}_\epsilon^2 = \|\boldsymbol{r}\|_2^2/(n - df(\tau))$ be an estimate of variance, we have that

$$\hat{\tau} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_\epsilon^2} = \frac{\hat{\sigma}_b^2}{\|\boldsymbol{r}\|_2^2}\left(n - \mathrm{df}(\hat{\tau})\right).$$

Therefore, letting

$$\psi(\tau) = \tau - \frac{\hat{\sigma}_b^2}{\|\boldsymbol{r}\|_2^2}\left(n - \mathrm{df}(\tau)\right),$$

we numerically solve for $\hat{\tau}$ such that $\psi(\hat{\tau}) = 0$ and set $\mathrm{df} = \mathrm{df}(\hat{\tau})$.

6.1. Stein's method

Let $g(\boldsymbol{y}) = \hat{\boldsymbol{y}}$, where $g : \mathbb{R}^n \to \mathbb{R}^n$ is the model fitting procedure. For $\boldsymbol{y} \sim N(\mu, \sigma^2 I)$, the degrees of freedom is defined as [see 6, 16]

$$\mathrm{df} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}(g_i(y), y_i). \tag{14}$$

As Tibshirani [39] notes, (14) is motivated by the fact that the risk $\mathrm{Risk}(g) = \mathbb{E}\|g(\boldsymbol{y}) - \boldsymbol{\mu}\|_2^2$ can be decomposed as

$$\mathrm{Risk}(g) = \mathbb{E}\|g(\boldsymbol{y}) - \boldsymbol{y}\|_2^2 - n\sigma^2 + 2\sum_{i=1}^{n} \mathrm{Cov}(g_i(\boldsymbol{y}), y_i).$$

Therefore, the degrees of freedom (14) corresponds to the difference between risk and expected training error. Furthermore, if g is continuous and weakly differentiable, then $\mathrm{df} = \mathbb{E}[\nabla \cdot g(y)]$ [35] where $\nabla \cdot g = \sum_{i=1}^{n} \partial g_i/\partial y_i$ is the divergence of g. Therefore, an unbiased estimate of df (also used in Stein's unbiased risk estimate [35]) is

$$\hat{\mathrm{df}} = \sum_{i=1}^{n} \partial g_i/\partial y_i. \tag{15}$$

To obtain an estimate of degrees of freedom, we transform the generalized lasso component of our model to standard form, similar to the approach of Petersen, Witten and Simon [25]. To do so, we use the following matrices described

by Tibshirani [40]. Let

$$\tilde{D}_j^* = \begin{bmatrix} \tilde{D}_{j,1}^{(0)} \\ \vdots \\ \tilde{D}_{j,1}^{(k_j)} \\ \tilde{D}_j^{(k_j+1)} \end{bmatrix} \in \mathbb{R}^{p_j \times p_j}$$

be an augmented finite difference matrix, where $\tilde{D}_{j,1}^{(i)}$ is the first row of the finite difference matrix $\tilde{D}_j^{(i)}$, and $\tilde{D}_j^{(0)} = I_{p_j}$ is the identity matrix. As shown by Tibshirani [40], the inverse of $\tilde{D}_j^*$ is given by $M_j = M_j^{(0)} M_j^{(1)} \cdots M_j^{(k)}$ where[3]

$$M_j^{(i)} = \begin{bmatrix} I_i & \\ & L_{(p_j-i)\times(p_j-i)} \end{bmatrix} \in \mathbb{R}^{p_j \times p_j},$$

where $L_{(p_j-i)\times(p_j-i)}$ is the $(p_j - i) \times (p_j - i)$ lower diagonal matrix of 1s.

Assuming our outcome $\boldsymbol{y}$ is centered, so that $\beta_0 = y(0) = 0$, and letting $V_j = \tilde{F}_j M_j$, $D_j^* = \tilde{D}_j^* Q_j$ for $j \in \mathcal{E}$ and $D_j^* = \tilde{D}_j^*$ for $j \in \bar{\mathcal{E}}$, and $\boldsymbol{\alpha}_j = D_j^* \boldsymbol{\beta}_j$, we can write the penalized negative log likelihood (8) as

$$l_{\text{pen}} = \frac{1}{2}\|\boldsymbol{y} - \sum_j V_j \boldsymbol{\alpha}_j - Z\boldsymbol{b}\|_2^2 + \sum_{j=1}^{J} \lambda_j \sum_{l=k_j+2}^{p_j} |\alpha_{jl}| + \frac{1}{2}\tau \boldsymbol{b}^T S \boldsymbol{b}. \tag{16}$$

To avoid difficulties later differentiating with respect to the ℓ_1 norm, we remove the non-active ℓ_1 penalized coefficients from (16). We also form the concatenated design matrix $V = [V_1, \ldots, V_J]$ and will need to index the active set of V. To these ends, let $\mathcal{A}_j = \{l \in \{k_j + 2, \ldots, p_j'\} : \hat{\alpha}_{j,l} \neq 0\}$ be the active set of the penalized coefficients for smooth j, and let $\mathcal{A}_j^* = \{1, \ldots, k_j + 1\} \cup \mathcal{A}_j$ be the active set for smooth j augmented with the unpenalized coefficients. Also, for a set $\mathcal{A}_j$ and constant $c \in \mathbb{R}$, let $\mathcal{A}_j + c = \{i + c : i \in \mathcal{A}_j\}$ be the set of elements in $\mathcal{A}_j$ shifted by c. Now let $\mathcal{A}^* = \bigcup_{j=1}^{J} (\mathcal{A}_j^* + \sum_{l=0}^{j-1} p_l')$ be the augmented active set of V, where $p_0' = 0$ and $p_j', j = 1, \ldots, J$ are the number of columns in V_j (equivalently F_j). Finally, let $V_{\mathcal{A}^*}$ be matrix V subset to retain only those columns indexed by $\mathcal{A}^*$. Similarly, let $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_1^T, \ldots, \hat{\boldsymbol{\alpha}}_J^T)^T$ be the concatenated vector of estimated coefficients, and let $\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}$ be vector $\hat{\boldsymbol{\alpha}}$ subset to retain only elements indexed by $\mathcal{A}^*$. Then we can write the estimated penalized loss (16) as

$$\hat{l}_{\text{pen}} = \frac{1}{2}\left\| \boldsymbol{y} - [V_{\mathcal{A}^*}, Z] \begin{pmatrix} \hat{\boldsymbol{\alpha}}_{\mathcal{A}^*} \\ \hat{\boldsymbol{b}} \end{pmatrix} \right\|_2^2 + \sum_{j=1}^{J} \lambda_j \sum_{l=k_j+2}^{p_j} |\hat{\alpha}_{jl}| + \frac{1}{2}\tau \hat{\boldsymbol{b}}^T S \hat{\boldsymbol{b}} \tag{17}$$

Taking the derivative of (17) and keeping in mind that the first k_j+1 elements of each $\hat{\boldsymbol{\alpha}}_j$ are unpenalized and $|\hat{\alpha}_{jl}| > 0$ for all $l \in \mathcal{A}_j$, we have

$$\mathbf{0}_{(|\mathcal{A}^*|+q)\times 1} = \frac{\partial l_{\text{pen}}}{\partial (\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T} = \begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix} \left([V_{\mathcal{A}^*}, Z] \begin{pmatrix} \hat{\boldsymbol{\alpha}}_{\mathcal{A}^*} \\ \hat{\boldsymbol{b}} \end{pmatrix} - \boldsymbol{y} \right) + \begin{pmatrix} \boldsymbol{\eta} \\ \tau S \hat{\boldsymbol{b}} \end{pmatrix} \tag{18}$$

[3]We denote the inverse matrix as M_j. This is unrelated to our use of M in Section 2 to denote the order of the B-spline basis.

where

$$\boldsymbol{\eta} = \begin{bmatrix} \mathbf{0}_{k_1+1} \\ \lambda_1 \operatorname{sign}(\hat{\boldsymbol{\alpha}}_{\mathcal{A}_1}) \\ \mathbf{0}_{k_2+1} \\ \lambda_2 \operatorname{sign}(\hat{\boldsymbol{\alpha}}_{\mathcal{A}_2+p_1}) \\ \vdots \\ \mathbf{0}_{k_J+1} \\ \lambda_J \operatorname{sign}(\hat{\boldsymbol{\alpha}}_{\mathcal{A}_J+\sum_{j=1}^{J-1} p_j}) \end{bmatrix},$$

$\mathbf{0}_{k_j+1}$ is a $(k_j+1)\times 1$ vector of zeros, and the sign operator is taken element-wise.

From Tibshirani and Taylor [41, Lemmas 6 and 9], we know that within a small neighborhood of $\boldsymbol{y}$, the active set $\mathcal{A}$ and the sign of the fitted terms $\hat{\alpha}_{\mathcal{A}}$ are constant with respect to $\boldsymbol{y}$ except for $\boldsymbol{y}$ in a set of measure zero. Therefore, $\partial\boldsymbol{\eta}/\partial\boldsymbol{y} = 0_{|\mathcal{A}^*|\times n}$, where $0_{|\mathcal{A}^*|\times n}$ is an $|\mathcal{A}^*| \times n$ matrix of zeros and $|\mathcal{A}^*|$ is the cardinality of $\mathcal{A}^*$. Then taking the derivative of (18) with respect to $\boldsymbol{y}$, we have

$$0_{(|\mathcal{A}^*|+q)\times n} = \frac{\partial^2 l_{\text{pen}}}{\partial(\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T \partial \boldsymbol{y}} = \begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix} [V_{\mathcal{A}^*}, Z] \begin{bmatrix} \partial\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}/\partial\boldsymbol{y} \\ \partial\hat{\boldsymbol{b}}/\partial\boldsymbol{y} \end{bmatrix} - \begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix} + \begin{bmatrix} 0_{|\mathcal{A}^*|\times n} \\ \tau S(\partial\hat{\boldsymbol{b}}/\partial\boldsymbol{y}) \end{bmatrix}.$$

Solving for the derivatives of the estimated coefficients, we have

$$\begin{bmatrix} \partial\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}/\partial\boldsymbol{y} \\ \partial\hat{\boldsymbol{b}}/\partial\boldsymbol{y} \end{bmatrix} = \left(\begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix} [V_{\mathcal{A}^*}, Z] + \begin{bmatrix} 0_{|\mathcal{A}^*|\times|\mathcal{A}^*|} & 0_{|\mathcal{A}^*|\times q} \\ 0_{q\times|\mathcal{A}^*|} & \tau S \end{bmatrix} \right)^{-1} \begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix}.$$

Now let $A = [V_{\mathcal{A}^*}, Z]$ and

$$\Omega = \begin{bmatrix} 0_{|\mathcal{A}^*|\times|\mathcal{A}^*|} & 0_{|\mathcal{A}^*|\times q} \\ 0_{q\times|\mathcal{A}^*|} & \tau S \end{bmatrix}.$$

Then since $\hat{\boldsymbol{y}} = A(\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T$ we have

$$\begin{aligned} \frac{\partial\hat{\boldsymbol{y}}}{\partial\boldsymbol{y}} &= \frac{\partial\hat{\boldsymbol{y}}}{\partial(\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T} \frac{\partial(\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T}{\partial\boldsymbol{y}} \\ &= A\left(A^TA + \Omega\right)^{-1} A^T. \end{aligned}$$

From Tibshirani and Taylor [41, Lemmas 1 and 8], we know that $g(\boldsymbol{y}) = \hat{\boldsymbol{y}}$ is continuous and weakly differentiable. Also, $\nabla g = \operatorname{tr}(\partial\hat{\boldsymbol{y}}/\partial\boldsymbol{y})$. Therefore, we can use Stein's formula (15) to estimate the degrees of freedom as

$$\hat{\mathrm{df}} = 1 + \operatorname{tr}\left(A(A^TA + \Omega)^{-1} A^T\right) = 1 + \operatorname{tr}\left((A^TA + \Omega)^{-1}A^TA\right), \tag{19}$$

where we add 1 for the intercept. We note that this result is similar to the degrees of freedom for the elastic net [see the remark on page 18 of 41] as well as for FLAM [25].

To obtain degrees of freedom for individual smooths $j = 1, \ldots, J$, let E_j be an $(|\mathcal{A}^*| + q) \times (|\mathcal{A}^*| + q)$ matrix with 1s on the diagonal positions indexed by $\mathcal{A}_j^* + \sum_{l=0}^{j-1} |\mathcal{A}_l^*|$ and zero elsewhere, where $|\mathcal{A}_j^*|$ is the cardinality of $\mathcal{A}_j^*$ and $\mathcal{A}_0^* = \emptyset$. Also, let $\hat{f}_j = V_j \hat{\boldsymbol{\alpha}}_j$ be the estimate of the j^{th} smooth. Then as Ruppert, Wand and Carroll [30] note, $\hat{f}_j = AE_j(A^TA + \Omega)^{-1}A^T\boldsymbol{y}$. Therefore,

$$\hat{\mathrm{df}}_j = \mathrm{tr}\left(AE_j(A^TA + \Omega)^{-1}A^T\right) = \mathrm{tr}\left(E_j(A^TA + \Omega)^{-1}A^TA\right). \tag{20}$$

In other words, the degrees of freedom for smooth j is the sum of the diagonal elements of $(A^TA + \Omega)^{-1}A^TA$ indexed by $\mathcal{A}_j^* + \sum_{l=0}^{j-1} |\mathcal{A}_l^*|$.

We note that when using the ADMM algorithm, or most likely any proximal algorithm, the fitted $D_j\hat{\boldsymbol{\beta}}_j$, or equivalently $\hat{\boldsymbol{\alpha}}_j$, will typically have several very small non-zero values, but will not typically be sparse. However, the vector $\hat{\boldsymbol{w}}_j$ is sparse, where in the ADMM algorithm we constrain $\boldsymbol{w}_j = D_j\boldsymbol{\beta}_j$. Therefore, in practice we use $\boldsymbol{w}_j$ to obtain the active set $\mathcal{A}_j$.

6.2. Stable and fast approximations

In some cases, such as the application in Section 9, the estimates based on Stein's method (19) and (20) cannot be computed due to numerical instability. In this section, we propose alternatives that are more numerically stable and which are also more computationally efficient.

6.2.1. Based on restricted derivatives

In this approach, we take derivatives of the fitted values restricted to individual smooths. In particular, from Section 6.1, we see that

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \hat{\boldsymbol{\alpha}}_{\mathcal{A}_j^*}} \frac{\partial \hat{\boldsymbol{\alpha}}_{\mathcal{A}_j^*}}{\partial \boldsymbol{y}} = V_{\mathcal{A}_j^*}(V_{\mathcal{A}_j^*}^T V_{\mathcal{A}_j^*})^{-1} V_{\mathcal{A}_j^*}^T$$

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \hat{\boldsymbol{b}}} \frac{\partial \hat{\boldsymbol{b}}}{\partial \boldsymbol{y}} = Z(Z^TZ + \tau S)^{-1}Z^T.$$

We can then approximate the degrees of freedom for each individual smooth and the random effects by

$$\tilde{\mathrm{df}}_j = \begin{cases} \mathrm{tr}\left((V_{\mathcal{A}_j^*}^T V_{\mathcal{A}_j^*})^{-1} V_{\mathcal{A}_j^*}^T V_{\mathcal{A}_j^*}\right) & j = 1, \ldots, J \\ \mathrm{tr}\left((Z^TZ + \tau S)^{-1}Z^TZ\right) & j = J+1 \end{cases} \tag{21}$$

We estimate the overall degrees of freedom as

$$\tilde{\mathrm{df}} = 1 + \sum_{j=1}^{J+1} \tilde{\mathrm{df}}_j \tag{22}$$

where we add 1 for the intercept.

This approach is similar to one described by Ruppert, Wand and Carroll [30, p. 176], though in a different context and for a different purpose. In particular, whereas we use this approach to approximate the degrees of freedom after fitting the model, Ruppert, Wand and Carroll [30] use it to set the degrees of freedom before fitting the model in the context of ℓ_2 penalized loss functions.

6.2.2. Based on ADMM constraint parameters

In this section, we propose estimates of degrees of freedom specific to the ADMM algorithm. As in the previous section, this approach is based on estimates for the individual smooths. Consider the model with $J = 1$ smooth, no random effects, and centered $\boldsymbol{y}$:

$$\|\boldsymbol{y} - F\boldsymbol{\beta}\|_2^2 + \lambda\|D\boldsymbol{\beta}\|_1.$$

Suppose we make the centering constraints described Section 3, i.e. we set $F = \tilde{F}Q$ and $D = \tilde{D}^{(k+1)}Q$ for an $n \times p$ design matrix $\tilde{F}$, a $k+1$ order finite difference matrix $D^{(k+1)}$, and an orthonormal $p \times (p-1)$ matrix Q. Let $\mathcal{A} = \{l \in \{1, \ldots, p-k-1\} : (D\hat{\boldsymbol{\beta}})_l \neq 0\}$ be the active set, and let $|\mathcal{A}|$ be its cardinality. In our context, we expect the design matrices F to be full rank, in which case Theorem 3 of Tibshirani and Taylor [41] (see the first Remark) states that the degrees of freedom is given by $\mathrm{df} = \mathbb{E}[\mathrm{nullity}(D_{-\mathcal{A}})]$. Here, $\mathrm{nullity}(D)$ is the dimension of the null space of matrix D, and $D_{-\mathcal{A}}$ is matrix D with rows indexed by $\mathcal{A}$ removed. Now, D has dimensions $(p-k-1) \times (p-1)$, and we can see by inspection that for all $k < p-1$ the columns of D are linearly independent. Therefore, the rank of $D_{-\mathcal{A}}$ is equal to the number of rows $p-k-1-|\mathcal{A}|$, and the nullity is equal to the number of columns $p-1$ minus the number of rows. This gives $\hat{\mathrm{df}} = \mathrm{nullity}(D_{-\mathcal{A}}) = k + |\mathcal{A}|$ for centered smooths, i.e. the number of non-zero elements of $D\hat{\boldsymbol{\beta}}$ plus one less than the order of the difference penalty. This is similar to the result for ℓ_1 trend filtering, but we have lost one degree of freedom due to the constraint that $\mathbf{1}^T\tilde{F}\tilde{\boldsymbol{\beta}} = \mathbf{0}$. For uncentered smooths, D has dimensions $(p-k-1) \times p$, which gives $\hat{\mathrm{df}} = \mathrm{nullity}(D_{-\mathcal{A}})) = k + 1 + |\mathcal{A}|$.

As before, we note that in the ADMM algorithm, $D\hat{\boldsymbol{\beta}}$ will not generally be sparse, as ADMM is a proximal algorithm. However, the corresponding $\boldsymbol{w}$ is sparse, where in the optimization problem we constrain $D\boldsymbol{\beta} = \boldsymbol{w}$. Now considering a model with smooths $j = 1, \ldots, J$, a numerically stable and fast alternative to (20) is given by

$$\tilde{\mathrm{df}}_j^{\mathrm{ADMM}} = \mathbb{1}[j \in \bar{\mathcal{E}}] + k_j + \sum_{l=1}^{p-k-1} \mathbb{1}\left[w_{jl} \neq 0\right]. \tag{23}$$

where $\bar{\mathcal{E}}$ indexes the un-centered smooths and $\mathbb{1}$ is an indicator variable. We then combine (23) with the restricted derivative approximation for the degrees of freedom of the random effects given above to obtain the overall degrees of freedom

$$\tilde{\mathrm{df}}^{\mathrm{ADMM}} = 1 + \sum_{j=1}^{J} \tilde{\mathrm{df}}_j^{\mathrm{ADMM}} + \mathrm{tr}\left((Z^TZ + \tau S)^{-1}Z^TZ\right), \tag{24}$$

where we add 1 for the intercept.

6.3. Ridge approximation

Let $U = [F_1, \ldots, F_J, Z]$ be the concatenated design matrix of fixed and random effects and

$$\Omega^{\text{ridge}} = \begin{bmatrix} \lambda_1 D_1^T D_1 & & & \\ & \ddots & & \\ & & \lambda_J D_J^T D_J & \\ & & & \tau S \end{bmatrix}$$

be the penalty matrix. Then the hat matrix from the linear smoother approximation (see Section 7) is given by $H = U(U^TU + \Omega^{\text{ridge}})^{-1}U^T$. Similar to before, we can get the overall degrees of freedom as

$$\hat{\text{df}}^{\text{ridge}} = 1 + \text{tr}\left((U^TU + \Omega^{\text{ridge}})^{-1}U^TU\right), \tag{25}$$

where we add 1 for the intercept. To obtain degrees of freedom for individual smooths $j = 1, \ldots, J$, let E_j be a $(p+q)\times(p+q)$ matrix with 1s on the diagonal positions indexed by the columns of F_j and zero elsewhere. Also, let $\hat{f}_j = F_j\hat{\boldsymbol{\beta}}_j$ be the estimate of the j^{th} smooth. Then the ridge approximation for smooth j is given by $\hat{f}_j \approx UE_j(U^TU + \Omega^{\text{ridge}})^{-1}U^T\boldsymbol{y}$. Therefore,

$$\hat{\text{df}}_j^{\text{ridge}} = \text{tr}\left(E_j(U^TU + \Omega^{\text{ridge}})^{-1}U^TU\right) \tag{26}$$

Similar to before, we also propose stable and fast approximations to the ridge estimate of degrees of freedom based on restricted derivatives. In particular, let

$$\tilde{\text{df}}_j^{\text{ridge}} = \begin{cases} \text{tr}\left({F_j}^T F_j + \lambda_j D_j^T D_j)^{-1}F_j^T F\right) & j = 1, \ldots, J \\ \text{tr}\left((Z^TZ + \tau S)^{-1}Z^TZ\right) & j = J+1 \end{cases} \tag{27}$$

Then we can estimate the overall degrees of freedom as

$$\tilde{\text{df}}^{\text{ridge}} = 1 + \sum_{j=1}^{J+1} \tilde{\text{df}}_j^{\text{ridge}} \tag{28}$$

where we add 1 for the intercept.

As noted above, this approach is similar to one described by Ruppert, Wand and Carroll [30, p. 176], though for a different purpose. Whereas we use this approach to obtain the degrees of freedom after fitting the model, Ruppert, Wand and Carroll [30] use it to set the degrees of freedom before fitting the model.

7. Approximate inference

In this section, we discuss approximate inferential methods based on ridge approximations to the ℓ_1 penalized fit and conditional on the smoothing parameters $\lambda_j, j = 1, \ldots, J$ and τ. We use the ADMM algorithm to analyze the approximation. In particular, we note that we can write the ADMM update for

$\boldsymbol{\beta}_j$ as

$$\boldsymbol{\beta}_j^{m+1} = \left(F_j^T F_j + \rho D_j^T D_j\right)^{-1} F_j^T \boldsymbol{y}^{(j,m)} + \boldsymbol{\delta}_j^m \tag{29}$$

where $\boldsymbol{\delta}_j^m = \rho(F_j^T F_j + \rho D_j^T D_j)^{-1} F_j^T D_j^T (\boldsymbol{w}_j^m - \boldsymbol{u}_j^m)$ and $\boldsymbol{y}^{(j,m)} = \boldsymbol{y} - \beta_0^{m+1} - \sum_{l<j} F_l \boldsymbol{\beta}_l^{m+1} - \sum_{l>j} F_l \boldsymbol{\beta}_l^m - Z\boldsymbol{b}^m$. As we note in Observation 2, $\boldsymbol{\delta}_j$ loosely represents the difference in the estimate of $\boldsymbol{\beta}_j$ obtained with the ℓ_1 and ℓ_2 penalties.

Observation 2. *With the ℓ_1 penalty, i.e. $\|D_j\boldsymbol{\beta}_j\|_1$, in general $\boldsymbol{\delta}_j^m \neq \mathbf{0}$. However, with the ℓ_2 penalty, i.e. $\|D_j\boldsymbol{\beta}_j\|_2^2$, and $\lambda_j = \rho$, we have $\boldsymbol{\delta}_j^m = \mathbf{0}$.*

Proof of Observation 2. Similar to the ridge update for $\boldsymbol{b}$, if we changed $\lambda_j\|D_j\boldsymbol{\beta}_j\|_1$ to $(\lambda_j/2)\|D_j\boldsymbol{\beta}_j\|_2^2$ in (8) we could remove the $\boldsymbol{w}_j$ term and the constraint that $D_j\boldsymbol{\beta}_j^m = \boldsymbol{w}_j$ from (12) to obtain the ridge update $\boldsymbol{\beta}_j^{m+1} = \left(F_j^T F_j + \lambda_j D_j^T D_j\right)^{-1} F_j^T \boldsymbol{y}^{(j,m)}$. Then since we assumed $\lambda_j = \rho$, we have $\boldsymbol{\beta}_j^{m+1} = \left(F_j^T F_j + \rho D_j^T D_j\right)^{-1} F_j^T \boldsymbol{y}^{(j,m)}$. By comparison with (29), we see that $\boldsymbol{\delta}_j^m = \mathbf{0}$. □

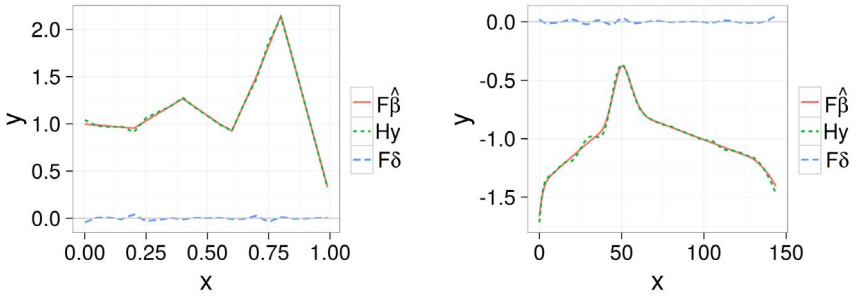
Observation 2 motivates our approximate inferential strategy. Letting $\hat{\boldsymbol{f}}_j$ be the j^{th} fitted smooth, and letting $\boldsymbol{y}^{(j)} = \boldsymbol{y} - \hat{\beta}_0 - \sum_{l\neq j} F_l\hat{\boldsymbol{\beta}}_l - Z\hat{\boldsymbol{b}}$, we have

$$\begin{aligned}
\hat{\boldsymbol{f}}_j = F_j\hat{\boldsymbol{\beta}}_j &= F_j(F_j^T F_j + \rho D_j^T D_j)^{-1} F_j^T \boldsymbol{y}^{(j)} + F_j\hat{\boldsymbol{\delta}}_j && (30)\\
&\approx F_j(F_j^T F_j + \rho D_j^T D_j)^{-1} F_j^T \boldsymbol{y}^{(j)} && (\text{assuming } F_j\hat{\boldsymbol{\delta}}_j \approx \mathbf{0})\\
&\approx F_j(F_j^T F_j + \lambda_j D_j^T D_j)^{-1} F_j^T \boldsymbol{y}^{(j)} && (\text{assuming } \lambda_j \approx \rho)\\
&= H_j\boldsymbol{y}^{(j)} && (31)
\end{aligned}$$

where $H_j = F_j(F_j^T F_j + \lambda_j D_j^T D_j)^{-1} F_j^T$. We obtain confidence intervals for the linear smoother (31) centered around the estimated fit (30), ignore $F_j\boldsymbol{\delta}_j$ when estimating variance, and assume $\lambda_j \approx \rho$. We also condition on the smoothing parameters $\lambda_1, \ldots, \lambda_J$ and τ.

Figure 2 gives a visual demonstration of the approximation for the simulation presented in Section 8 and the application shown in Section 9. As seen in Figure 2, in these examples the ℓ_1 fit and ridge approximation are very similar. If this holds in general, then this would suggest that 1) the approximate inferential procedures we propose might have reliable coverage probabilities, and 2) there may be minimal practical advantage to using an ℓ_1 penalty instead of the standard ℓ_2 penalty. However, as shown in Section 8.3, the ℓ_1 penalty appears to perform noticeably better in certain situations, including the detection of change points.

Before presenting the confidence bands in greater detail, we discuss our approach for estimating variance in Section 7.1, which we then use to form confidence bands in Section 7.2.



(a) Simulation (Section 8) (b) Application (Section 9)

FIG 2. *Linear smoother approximation to the ℓ_1 penalized fit in the simulation (see Section 8) and application (see Section 9). The solid red line is the ℓ_1 penalized fit, the dotted green line is the linear smoother approximation, and the dashed blue line is the difference between the two.*

7.1. Variance

Let $\boldsymbol{r} = \boldsymbol{y} - \sum_{j=1}^{J} F_j \hat{\boldsymbol{\beta}}_j - Z\hat{\boldsymbol{b}}$ be an $n \times 1$ vector of residuals. We estimate the overall variance as $\hat{\sigma}_\epsilon^2 = \|\boldsymbol{r}\|_2^2 / \hat{\mathrm{df}}_{\mathrm{resid}}$, where $\hat{\mathrm{df}}_{\mathrm{resid}}$ is the residual degrees of freedom. When possible, we use the estimate based on Stein's method (19) and set $\hat{\mathrm{df}}_{\mathrm{resid}} = n - \hat{\mathrm{df}}$. If Stein's method is not numerically stable, then we use the restricted derivatives approximation (22) and set $\hat{\mathrm{df}}_{\mathrm{resid}} = n - \tilde{\mathrm{df}}$. As another alternative, we could also use the ADMM approximation and set $\hat{\mathrm{df}}_{\mathrm{resid}} = n - \tilde{\mathrm{df}}^{\mathrm{ADMM}}$.

7.2. Confidence bands

In this section, we obtain confidence bands for typical subjects, i.e. for subjects for whom $\boldsymbol{b}_i = \boldsymbol{0}$. Since we assume a normal outcome, this is equivalent to the marginal population level response.

7.2.1. Frequentist confidence bands

Ignoring the distribution on $D_j\boldsymbol{\beta}_j$ and treating $\boldsymbol{\beta}_l$, $l \neq j$ as fixed, $\boldsymbol{y}^{(j)}$ is normal with variance $\mathrm{Var}(\boldsymbol{y}^{(j)}) = \sigma_\epsilon^2 I_n + \sigma_b^2 Z S^+ Z^T$, where S^+ is the Moore-Penrose generalized inverse of matrix S (as noted in Section 3, S may not be positive definite). Therefore, $\widehat{\mathrm{Var}}(\hat{\boldsymbol{f}}_j) \approx H_j \widehat{\mathrm{Var}}(\boldsymbol{y}^{(j)}) H_j^T$ where $\widehat{\mathrm{Var}}(\boldsymbol{y}^{(j)})$ is an $n \times n$ estimate of $\mathrm{Var}(\boldsymbol{y}^{(j)})$ with $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_b^2$ plugged in for σ_ϵ^2 and σ_b^2 respectively, and $\hat{\boldsymbol{f}}_j \dot{\sim} N(\hat{\boldsymbol{f}}_j, H_j \widehat{\mathrm{Var}}(\boldsymbol{y}^{(j)}) H_j^T)$. The estimated variance of the fit at a single point x, which we denote as $\widehat{\mathrm{Var}}(\hat{f}_j(x))$, is the corresponding diagonal element of $H_j \widehat{\mathrm{Var}}(\boldsymbol{y}^{(j)}) H_j^T$. Therefore, asymptotic pointwise $1-\alpha$ confidence bands take

the form $\hat{f}_j(x) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{f}_j(x))}$ where $\Phi(z_a) = a$ and Φ is the standard normal CDF, e.g. $z_{1-\alpha/2} = 1.96$ for $\alpha = 0.05$.

For the purposes of interpretation, we include the intercept term in the confidence band for the $j = 1$ smooth, but not for the remaining smooths.

7.2.2. Bayesian credible bands

Many authors, including Wood [47], recommend using Bayesian confidence bands for nonparametric and semiparametric models, because the point estimates are themselves biased. While Bayesian credible bands do not remedy the bias, they are self consistent.

To this end, we replace the element-wise Laplace prior with the (generally improper) joint normal prior that is equivalent to the standard ℓ_2 penalty: $\boldsymbol{\beta}_j \sim N\left(\mathbf{0}, (\lambda_j D_j^T D_j)^{-1}\right)$. This leads to the posterior

$$\boldsymbol{\beta}_j|\boldsymbol{y} \dot{\sim} N\left(\hat{\boldsymbol{\beta}}_j, \underbrace{(F_j^T \widehat{\mathrm{Var}}(\boldsymbol{y}^{(j)})^{-1} F_j + \lambda_j D_j^T D_j)}_{W_j}{}^{-1}\right). \tag{32}$$

We can then form simultaneous Bayesian credible bands for $\boldsymbol{f}_j|\boldsymbol{y}$ by simulating from the posterior (32) and taking quantiles from $F_j\boldsymbol{\beta}_j^b, b = 1, \ldots, B$. Alternatively, for a faster approximation we use frequentist confidence bands with $F_j W_j^{-1} F_j^T$ in place of $H_j \widehat{\mathrm{Var}}(\boldsymbol{y}^{(j)}) H_j^T$. In practice, we have found the simultaneous credible bands and the faster approximation to be nearly indistinguishable.[4]

As before, for the purposes of interpretation, we include the intercept term in the credible band for the $j = 1$ smooth, but not for the remaining smooths.

8. Simulation

We simulated data from a piecewise linear mean curve as shown in Figure 3. Each subject had a random intercept and is observed over only a portion of the domain. There are 50 subjects, each with between 4 and 14 measurements (450 total observations). The random intercepts were normally distributed with variance $\sigma_b^2 = 1$, and the overall noise was normally distributed with variance $\sigma_\epsilon^2 = 0.01$.

In all models, we used order 2 (degree 1) B-splines with $p = 21$ basis functions.

[4] It appears that the latter (faster) method is the default in the `mgcv` package [47]. As in `mgcv`, we only need to compute the diagonal elements of $F_j W_j^{-1} F_j^T$ as $\texttt{rowSums}((F_j W_j^{-1}) \circ F_j)$, where $\circ$ is the Hadamard (element-wise) product.

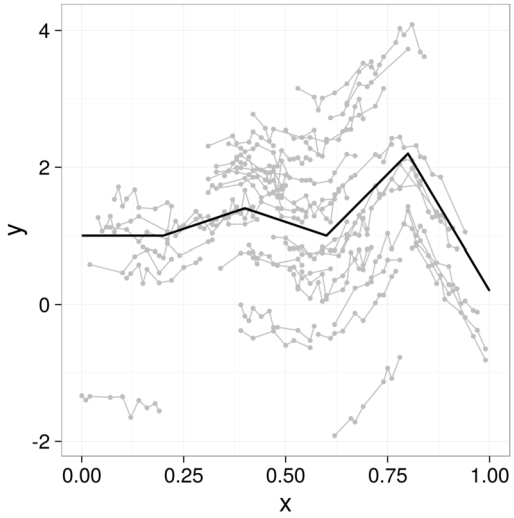


FIG 3. *Simulated data: true marginal curve in black, observed (simulated) data in gray.*

8.1. Frequentist estimation

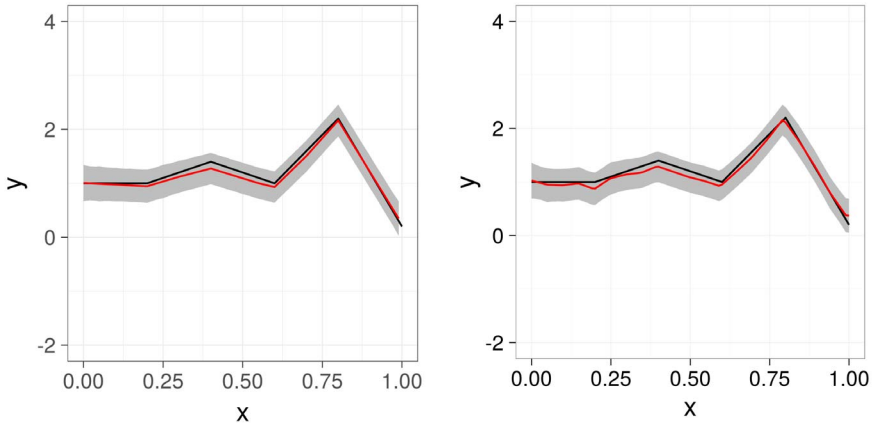
We fit models with $J = 1$ smooth term and random intercepts. To obtain estimates for the ℓ_1 penalized model, we used ADMM and 5-fold CV to minimize

$$\underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^{p-1}, \boldsymbol{b} \in \mathbb{R}^N}{\text{minimize}} \frac{1}{2} \|\boldsymbol{y} - \beta_0 \mathbf{1} - F\boldsymbol{\beta} - Z\boldsymbol{b}\|_2^2 + \lambda \|D^{(2)}\boldsymbol{\beta}\|_1 + \tau \boldsymbol{b}^T \boldsymbol{b}. \tag{33}$$

where $Z_{il} = 1$ if observation i belongs to subject l and zero otherwise. As noted above, we used order 2 (degree 1) B-splines with $p = 21$ basis functions, i.e. $F \in \mathbb{R}^{n \times (p-1)}$ where $n = 450$ and $p = 21$. After estimating λ and τ via CV, we used LME updates to estimate σ_b^2 and $\boldsymbol{b}$ in the final model. We also fit an equivalent model with an ℓ_2 penalty using the `mgcv` package [47], i.e. with $(\lambda/2)\|D^{(2)}\boldsymbol{\beta}\|_2^2$ in place of $\lambda\|D^{(2)}\boldsymbol{\beta}\|_1$ in (33). Figure 4 shows the marginal mean with 95% credible intervals, and Figure 5 shows the subject-specific predicted curves.

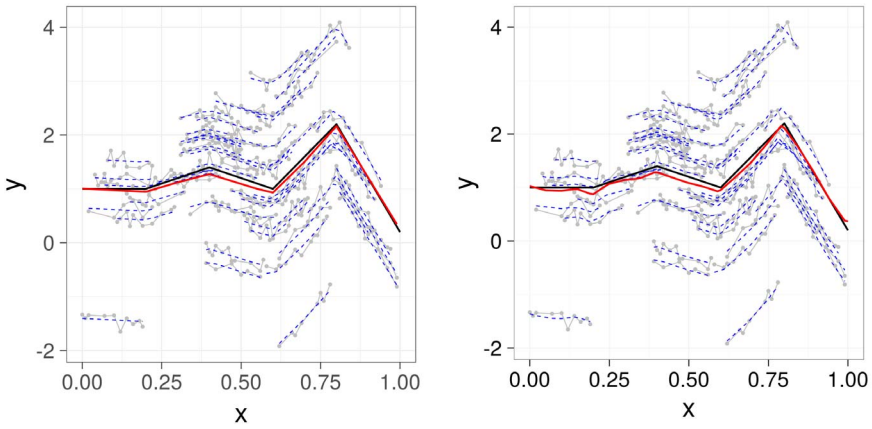
As seen in Figures 4 and 5, the results from the ℓ_1 and ℓ_2 penalized models are very similar. However, the ℓ_1 penalized model does slightly better at identifying the change points and the line segments. We explore this further in Section 8.3.

Table 1 compares the degrees of freedom and variance estimates from the ℓ_1 penalized fit against those from the ℓ_2 penalized fit. From Table 1, we see that the ridge degrees of freedom $\hat{\text{df}}^{\text{ridge}}$ appears reasonable, as it is near the estimate for the ℓ_2 penalized model. The true degrees of freedom $\hat{\text{df}}$ also seems reasonable. Ideally, the degrees of freedom for the ℓ_1 penalized fit should equal six, as there are four change points and we are using a second order difference penalty (see Section 6.2).



(a) ℓ_1 fit with ADMM and CV (b) ℓ_2 fit with mgcv [46]

FIG 4. *Marginal mean and 95% credible intervals from frequentist estimation: black is true marginal mean, red is estimated marginal mean*



(a) ℓ_1 fit with ADMM and CV (b) ℓ_2 fit with mgcv [46]

FIG 5. *Subject-specific predicted curves from frequentist estimation: black is true marginal mean, red is estimated marginal mean, blue is subject-specific curves*

TABLE 1
Estimated degrees of freedom for smooth F and variance in ℓ_1 and ℓ_2 penalized models

	Penalty		
Estimator	ℓ_1	ℓ_2	Truth
$\hat{\text{df}}^{\text{ridge}}$	17.7	19.0	–
$\hat{\text{df}}$	10	–	–
$\hat{\sigma}^2_\epsilon$	0.0093	0.0106	0.01
$\hat{\sigma}^2_b$	1.06	1.05	1

TABLE 2
Comparison of degrees of freedom estimates for the ℓ_1 penalized model

Estimator	Description	Smooth		
		Overall	F	Z
$\hat{\text{df}}$	Stein (19) and (20)	14.3	10.0	3.29
$\tilde{\text{df}}$	Restricted (21) and (22)	14.6	10.0	3.63
$\tilde{\text{df}}^{\text{ADMM}}$	ADMM (23) and (24)	13.6	9.0	3.63
$\hat{\text{df}}^{\text{ridge}}$	Ridge (25) and (26)	22.1	17.7	3.31
$\tilde{\text{df}}^{\text{ridge}}$	Ridge restricted (27) and (28)	22.4	17.8	3.63

Table 2 compares the different estimates of degrees of freedom. In this simulation, the degrees of freedom based on the ridge approximation is larger than that from Stein's formula, and the approximations based on restricted derivatives are equal or near the estimate with Stein's formula.

8.2. Bayesian estimation

We modeled the data as $\boldsymbol{y}|\boldsymbol{b} = \beta_0\mathbf{1} + F\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\epsilon}$ where

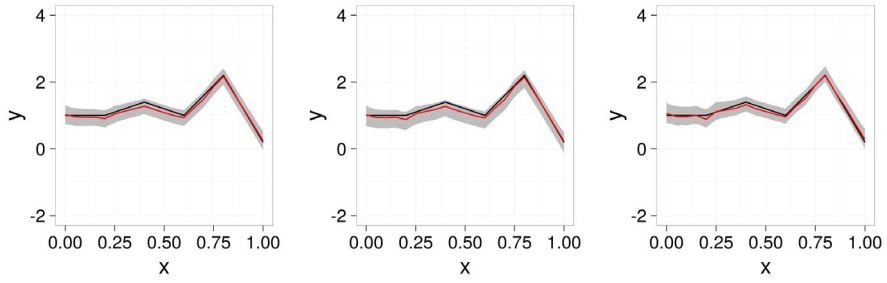
$$\begin{aligned}
\boldsymbol{\epsilon} &\sim N(\mathbf{0}, \sigma_\epsilon^2 I) \\
\boldsymbol{b} &\sim N(0, \sigma_b^2 I) \\
D^{(2)}\boldsymbol{\beta} &\sim \text{Laplace}(\mathbf{0}, \sigma_\lambda^2 I) \\
p(\sigma_\epsilon) &\propto 1 \\
p(\sigma_b) &\propto 1 \\
p(\log(\sigma_\lambda)) &\propto 1.
\end{aligned}$$

We also fit models with normal and diffuse priors for $D^{(2)}\boldsymbol{\beta}$.

We fit all models with `rstan` [36], each with four chains of 2,000 iterations with the first 1,000 iterations of each chain used as warmup. The MCMC chains, not shown, appeared to be reasonably well mixing and stationary, and had $\hat{R}$ values under 1.1 [see 12].[5] Figure 6 shows the marginal mean with 95% credible intervals, and Figure 7 shows point estimates.

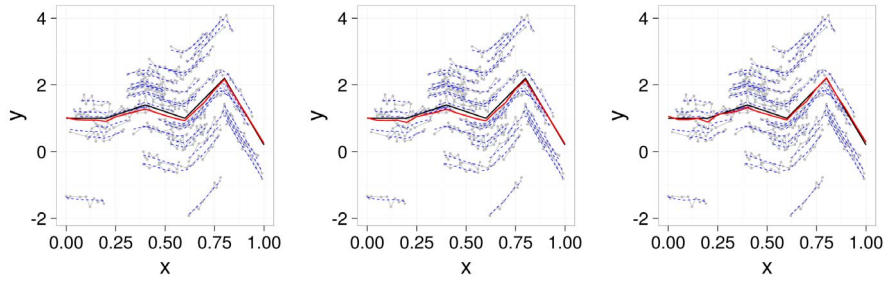
As seen in Figures 6 and 7, all models performed well and gave similar fits as above. Similar to before, the Laplace prior appears to better enforce a piece-wise linear fit, particularly around $x = 0.2$.

[5]As described by Gelman et al. [12, pp. 284–285], for each scalar parameter, $\hat{R}$ is the square root of the ratio of the marginal posterior variance (a weighed average of between- and within-chain variances) to the mean within-chain variance. As the number of iterations in the MCMC chains goes to infinity, $\hat{R}$ converges to 1 from above. Consequently, $\hat{R}$ can be interpreted as a scale reduction factor, and Gelman et al. [12] recommend ensuring that $\hat{R} < 1$ for all parameters.



(a) $D\boldsymbol{\beta} \sim \text{Laplace}(\mathbf{0}, \sigma_\lambda^2 I)$ (b) $D\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\lambda^2 I)$ (c) No prior on $D\boldsymbol{\beta}$

FIG 6. *Credible bands for Bayesian models with order 2 (degree 1) B-splines. Black is true marginal mean, red dashed is estimated marginal mean, gray area is 95% credible interval*



(a) $D\boldsymbol{\beta} \sim \text{Laplace}(\mathbf{0}, \sigma_\lambda^2 I)$ (b) $D\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\lambda^2 I)$ (c) No prior on $D\boldsymbol{\beta}$

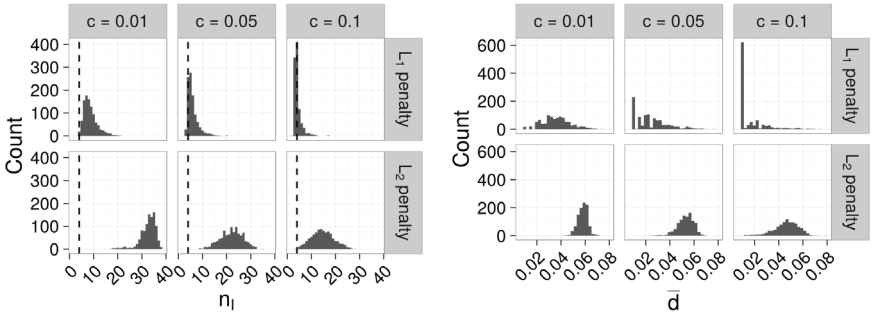
FIG 7. *Subject-specific predicted curves from Bayesian models fit with order 2 (degree 1) B-splines. Gray is observed data, black is true marginal mean, red dashed is estimated marginal mean, and blue dashed is subject-specific predictions*

8.3. Change point detection

We simulated 1,000 datasets with the same generating mechanism used to produce the data shown in Figure 3 and measured the performance of the ℓ_1 and ℓ_2 penalized models on two criteria: 1) the number of inflection points found, and 2) the distance between the estimated inflection points and the closest true inflection point. To that end, let $\mathcal{T} = \{\tau_1, \ldots, \tau_4\}$ be the set of true inflection points, and $M = \max_{x \in \mathcal{X}} |\hat{f}''(x)|$ be the maximum absolute second derivative of the estimated function, where $\mathcal{X} = \{x_1, x_2, \ldots\}$ is the ordered set of unique simulated x values. We approximate $\hat{f}''$ by

$$\hat{f}''(x_i) \approx \frac{(\hat{f}(x_{i+1}) - \hat{f}(x_i))/(x_{i+1} - x_i) - (\hat{f}(x_i) - \hat{f}(x_{i-1}))/(x_i - x_{i-1})}{x_{i+1} - x_i}.$$

Then let $\mathcal{I} = \{x \in \mathcal{X} : |\hat{f}''(x)| \geq cM\}$ be the set of estimated inflection points, where $c \in (0, 1)$ is a cutoff value defining how large the second derivative must



(a) Number of estimated inflection points (b) Mean absolute deviance

FIG 8. *Results from 1,000 simulated datasets measuring ability of the models to detect inflection points*

be to be counted as an inflection point. Also, let $n_{\mathcal{I}} = |\mathcal{I}|$ be the number of estimated inflection points, and $\bar{d} = n_{\mathcal{I}}^{-1} \sum_{x \in \mathcal{I}} \min_{\tau \in \mathcal{T}} |x - \tau|$ be the mean absolute deviance of the estimated inflection points.

Figure 8 shows the results from 1,000 simulated datasets. The ℓ_1 penalized model was better able to 1) find the correct number of inflection points, and 2) determine the location of the inflection points.

8.4. Coverage probability

We simulated 1,000 datasets with the same generating mechanism used to produce the data shown in Figure 3 and measured the coverage probability of the approximate Bayesian credible bands described in Section 7.2.2 for the ℓ_1 penalized model, and simultaneous Bayesian credible bands for the ℓ_2 penalized model [47]. Figure 9 shows the coverage probabilities for both approaches. As seen in Figure 9, the confidence bands perform similarly and are near the nominal rate over most of the x domain. Both approaches have difficulty maintaining nominal coverage at the edges of the x domain.

9. Application

9.1. Data description and preparation

In this section, we analyze electrodermal activity (EDA) data collected as part of a stress study. In brief, all subjects completed a written questionnaire prior to the study, which categorized the subjects as having either low vigilance or high vigilance personality types. During the study, all participants wore wristbands that measured EDA while undergoing stress-inducing activities, including giving a public speech and performing mental arithmetic in front of an audience. The scientific questions were: 1) Is EDA higher among high vigilance subjects, and

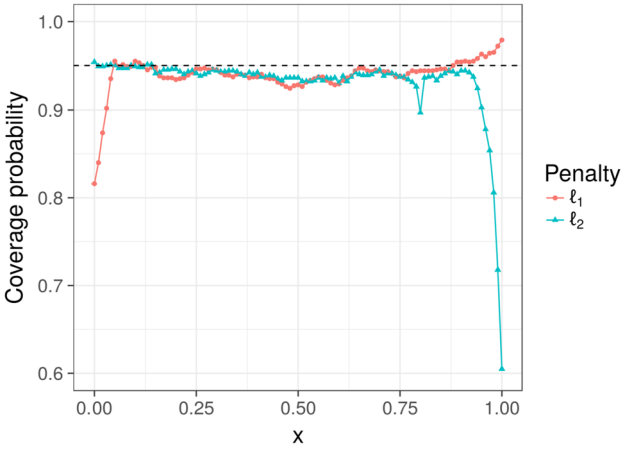


FIG 9. *Coverage probability from 1,000 simulated datasets using approximate Bayesian credible bands for the ℓ_1 penalized model and simultaneous Bayesian credible bands for the ℓ_2 penalized model.*

2) when did trends in stress levels change? In this section, we demonstrate how P-splines with an ℓ_1 penalty can address both questions.

The raw EDA data are shown in Figure 10. After excluding subjects who had EDA measurements of essentially zero throughout the entire study, we were left with ten high vigilance subjects and seven low vigilance subjects.

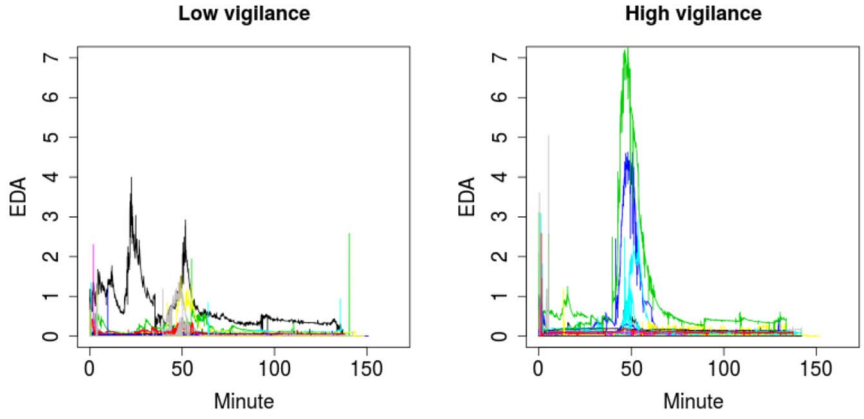


FIG 10. *Raw electrodermal activity (EDA) data by experimental group*

To remove the extreme second-by-second fluctuations in EDA, which we believe are artifacts of the measurement device as opposed to real biological signals, we smoothed each curve separately with a Nadaraya–Watson kernel estimator using the `ksmooth` function in R. We then thinned the data to reduce computational burden, taking 100 evenly spaced measurements from each subject.

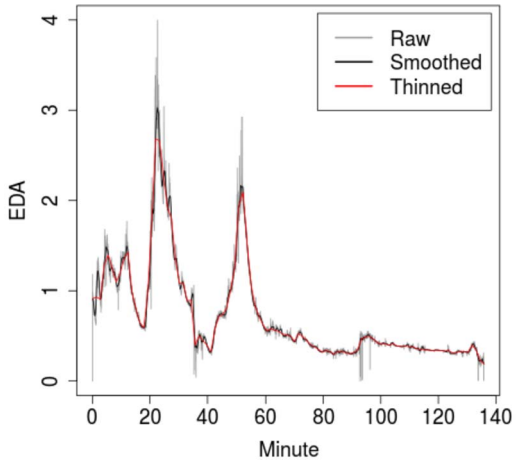


FIG 11. *Raw, smoothed, and thinned electrodermal activity data for a single subject*

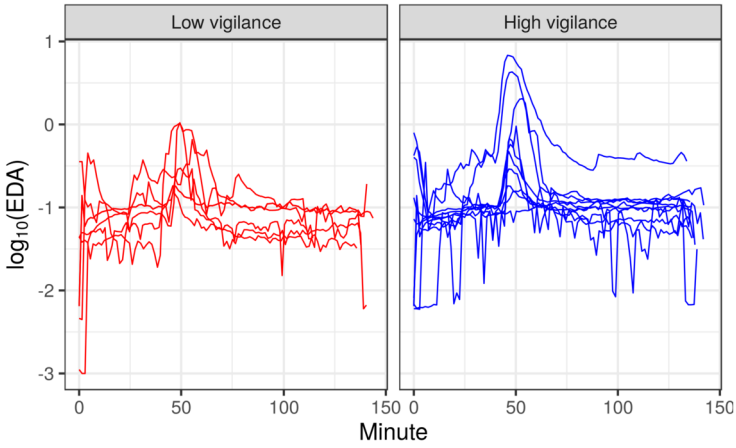


FIG 12. *Electrodermal activity (EDA) data used in the analysis (seven low vigilance and ten high vigilance subjects). Note: subjects not aligned in time (x-axis).*

Figure 11 shows the results of this process for a single subject, and Figure 12 shows the prepared data for all subjects. Because of the limited number of subjects, as well as issues of misalignment in the time series across individuals, the results presented here should be considered as illustrative rather than of full scientific validity.

9.2. Models

In all models, we fit the structure

$$y_i(x) = \beta_0 + \beta_1(x) + \mathbb{1}_{\text{high}}[i]\beta_2(x) + b_i(x) + \epsilon_i(x)$$

where x represents time in minutes, $\mathbb{1}_{\text{high}}[i] = 1$ if subject i has high vigilance and $\mathbb{1}_{\text{high}}[i] = 0$ if subject i has low vigilance, $b_i(x)$ are random subject-specific curves, and $\epsilon_i(x) \sim N(0, \sigma_\epsilon^2)$. For $\beta_1(x)$, $\beta_2(x)$, and $b_i(x)$, we used a fourth order B-spline basis with 31 basis functions each and a second order difference penalty ($k = 1$).

Written in matrix notation, the ℓ_1 penalized model is

$$\min \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \sum_{j=1}^{2} F_j\boldsymbol{\beta}_j - Z\boldsymbol{b}\|_2^2 + \sum_{j=1}^{2}\lambda_j\|D^{(2)}\boldsymbol{\beta}_j\|_1 + \boldsymbol{b}^T S\boldsymbol{b} \tag{34}$$

where $\boldsymbol{y}$ is a stacked vector for subjects $i = 1, \ldots, 17$, F_1 is an $n \times p$ design matrix where $n = 1,700$ and $p = 31$, and $F_2 = \text{diag}(\mathbb{1}_{\text{high}}[\boldsymbol{i}])F_1$ where $\boldsymbol{i}$ is an $n \times 1$ vector of subject IDs. In other words, F_2 is equal to F_1, but with rows corresponding to low vigilance subjects zeroed out. We set

$$Z = \begin{bmatrix} Z_1 & & \\ & \ddots & \\ & & Z_{17} \end{bmatrix}$$

where each Z_i is an $n_i \times 31$ random effects design matrix of order 4 B-splines evaluated at the input points for subject i, and

$$S = \begin{bmatrix} S_1 & & \\ & \ddots & \\ & & S_{17} \end{bmatrix}$$

where $S_{i,jl} = \int \phi''_{ij}(t)\phi''_{il}(t)dt$ are smoothing spline penalty matrices. We also mean-centered F_1 as described in Section 3, with the corresponding changes in dimensions.

To fit a comparable ℓ_2 penalized model, in which $\lambda_j\|D^{(2)}\boldsymbol{\beta}_j\|_1$ in (34) is replaced with $(\lambda_j/2)\|D^{(2)}\boldsymbol{\beta}_j\|_2^2$, we rotated the random effect design and penalty matrices Z and S as described in Section 3. To facilitate the use of existing software, we used a normal prior for the "unpenalized" random effect coefficients, i.e. $\check{\boldsymbol{b}}_f \sim N(\mathbf{0}, \sigma_f^2 I)$.

We also fit a Bayesian model using the same rotations and equivalent penalties as above. In particular, we modeled the data as $\boldsymbol{y}|\boldsymbol{b} = \beta_0\mathbf{1} + \sum_{j=1}^{J} F_j\boldsymbol{\beta}_j + \check{Z}_r\check{\boldsymbol{b}}_r + \breve{Z}_f\breve{\boldsymbol{b}}_f + \boldsymbol{\epsilon}$ where

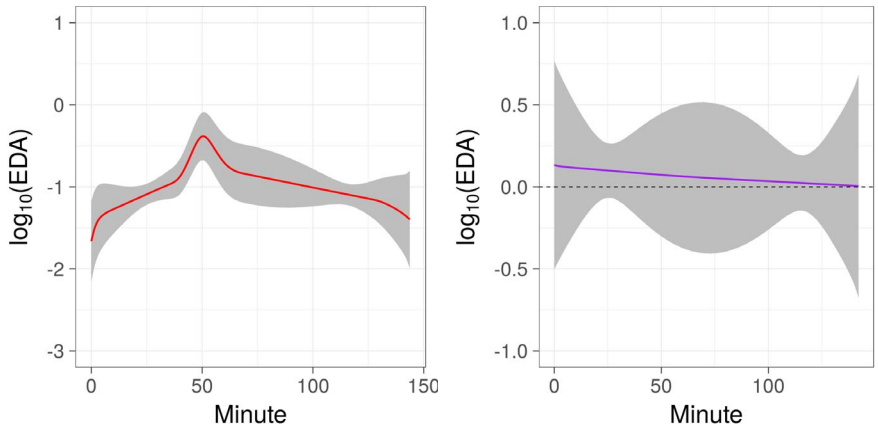
$$\begin{aligned} \check{\boldsymbol{b}}_r &\sim N(\mathbf{0}, \sigma_r^2 I) \\ \breve{\boldsymbol{b}}_f &\sim N(\mathbf{0}, \sigma_f^2 I) \\ (D_j\boldsymbol{\beta}_j)_l &\sim \text{Laplace}(0, a_j) \text{ for } a_j = \sigma_\epsilon^2/(2\lambda_j), l = 1, \ldots, p_j - k_j - 1, j = 1, \ldots, J \\ \boldsymbol{\epsilon} &\sim N\left(\mathbf{0}, \sigma_\epsilon^2 I_n\right). \end{aligned} \tag{35}$$

9.3. Results

9.3.1. Frequentist estimation

We tried to use CV to estimate the smoothing parameters for the ℓ_1 penalized model. However, with only 17 subjects split between two groups, we only did 3-fold CV. CV did not find a visually reasonable fit so we set the tuning parameters by hand.

Figure 13 shows the estimated marginal mean and 95% credible bands for the ℓ_1 penalized model, and Figure 14 shows the subject-specific predicted curves for the ℓ_1 penalized model. As seen in Figure 13a, our model identified a few inflection points, particularly near minutes 40, 50, and 60. From Figure 13b it appears that the difference in EDA between the low and high vigilance subjects was not statistically significant. Also, as seen in Figure 14, the subject-specific predicted curves are shrunk towards the mean, which is expected, because the predicted curves are analogous to BLUPs, although they are not linear smoothers.



(a) $\hat{\beta}_1(x)$ (low vigilance) (b) $\hat{\beta}_2(x)$ (high − low vigilance)

FIG 13. *ℓ_1 penalized model: parameter estimates with 95% confidence bands*

Figure 15 shows the estimated marginal mean and 95% credible bands for the ℓ_2 penalized model, and Figure 16 shows the subject-specific predicted curves for the ℓ_2 penalized model. The estimate shown in Figure 15a is similar to that shown in Figure 13a, though the inflection points are slightly less pronounced in Figure 15a. The results in Figure 15b are for the most part substantively the same as those in Figure 13b; the ℓ_2 penalized model does not show a statistically significant difference between the low and high vigilance subjects, with the possible exception of minutes 45 to 66. As seen in Figure 16, the predicted subject-specific curves from the ℓ_2 penalized model are also shrunk towards the mean.

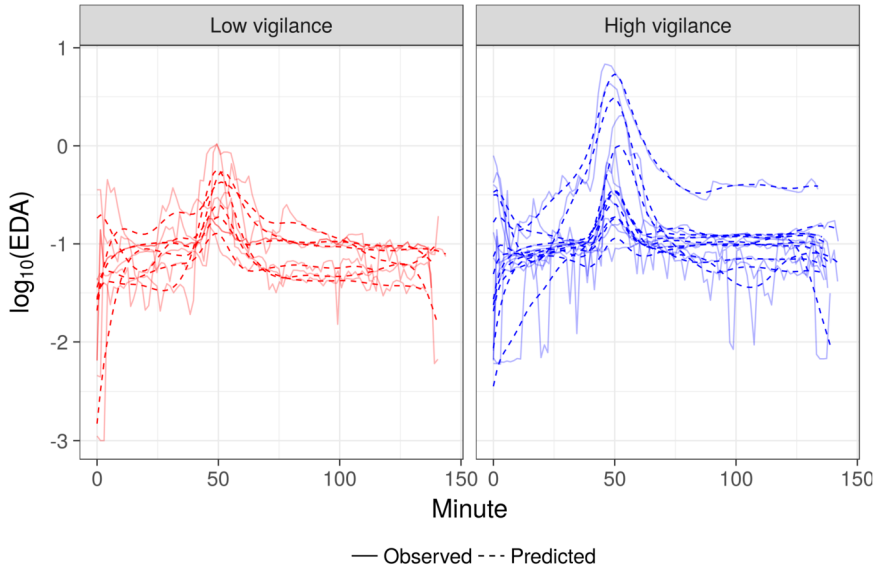
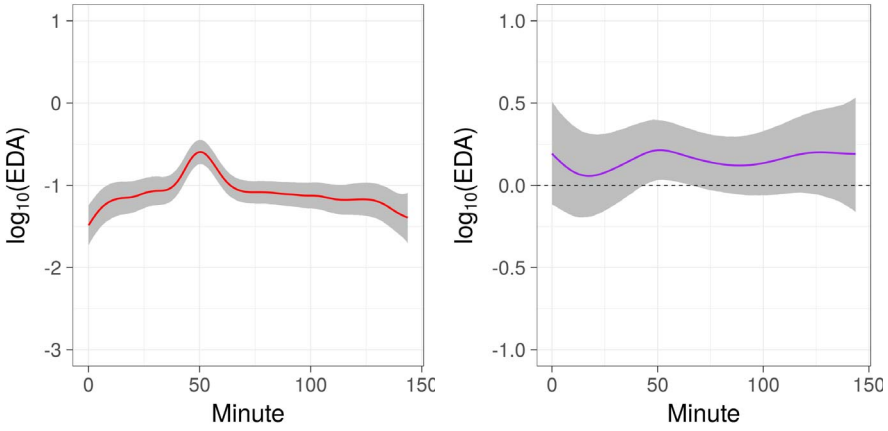


FIG 14. ℓ_1 *penalized model: subject-specific predicted curves*



(a) $\hat{\beta}_1(x)$ (low vigilance) (b) $\hat{\beta}_2(x)$ (high − low vigilance)

FIG 15. ℓ_2 *penalized model: parameter estimates with 95% confidence bands*

Table 3 shows the estimated degrees of freedom for the ℓ_1 penalized model. Stein's method $\hat{\mathrm{df}}$ ((19) and (20)) and the ridge approximation $\hat{\mathrm{df}}^{\mathrm{ridge}}$ ((25) and (26)) were numerically instable ($A^TA+\Omega$ and $U^TU+\Omega^{\mathrm{ridge}}$ were computationally singular). Therefore we used the restricted derivative approximation $\tilde{\mathrm{df}}$ to estimate the variance, as described in Section 7.1. In the ℓ_2 penalized model,

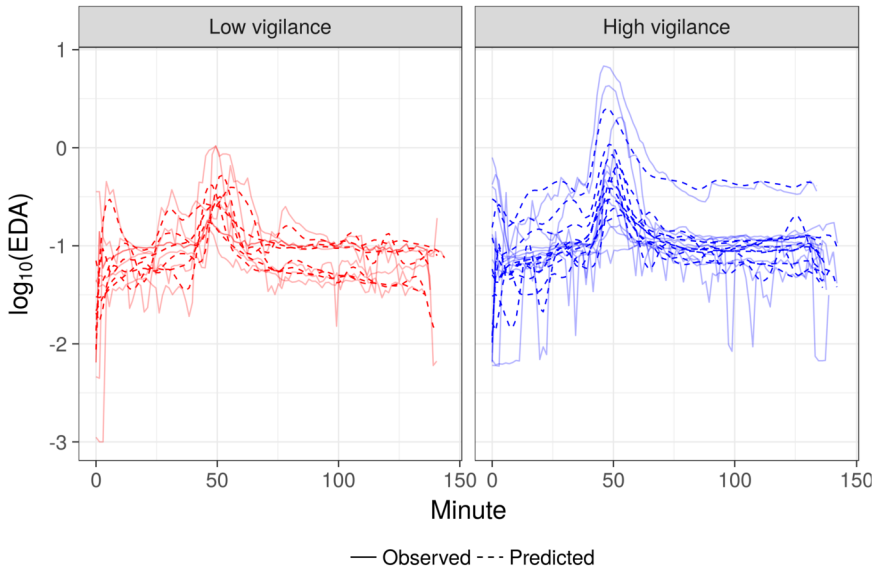


FIG 16. *ℓ_2 penalized model: subject-specific predicted curves*

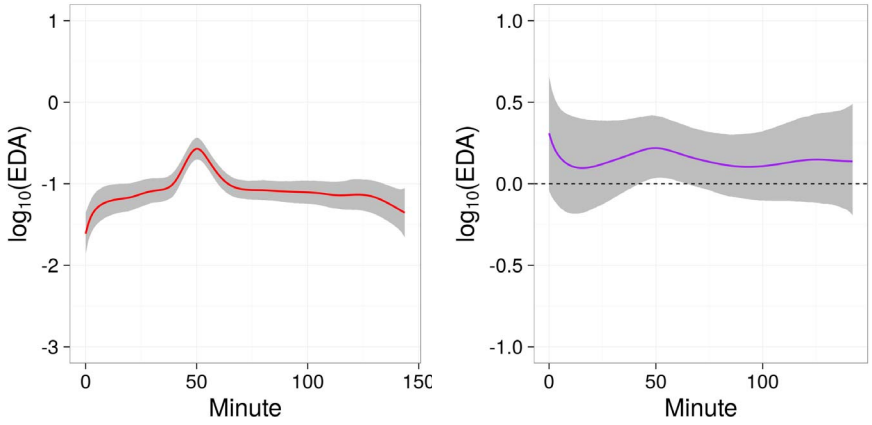
smooth F_1 had 14.2 degrees of freedom, and smooth F_2 had 6.96 degrees of freedom.

TABLE 3
Comparison of degrees of freedom estimates for the ℓ_1 penalized model

Estimator	Description	Smooth			
		Overall	F_1	F_2	Z
$\hat{\mathrm{df}}$	Stein (19) and (20)	–	–	–	–
$\tilde{\mathrm{df}}$	Restricted (21) and (22)	194	10.0	2.00	181
$\tilde{\mathrm{df}}^{\mathrm{ADMM}}$	ADMM (23) and (24)	193	9.00	2.00	181
$\hat{\mathrm{df}}^{\mathrm{ridge}}$	Ridge (25) and (26)	–	–	–	–
$\tilde{\mathrm{df}}^{\mathrm{ridge}}$	Ridge restricted (27) and (28)	216	21.1	13.50	181

9.3.2. Bayesian estimation

We fit the model described in Section 9.2 with an element-wise Laplace prior on $D\boldsymbol{\beta}$ given by (35). To fit the model, we used `rstan` [36] with four chains of 5,000 iterations each, with the first 2,500 iterations of each chain used as warmup. The MCMC chains, not shown, appeared to be reasonably well mixing and stationary with $\hat{R}$ values under 1.1 [see 12]. Figure 17 shows the marginal means with 95% credible bands, and Figure 18 shows the subject-specific curves. Similar to the ℓ_2 penalized model, the Bayesian model found a slightly statistically significant difference between low and high vigilance between minutes 42 and 65.



(a) $\hat{\beta}_1(x)$ (low vigilance subjects) (b) $\hat{\beta}_2(x)$ (high − low vigilance subjects)

FIG 17. *Bayesian model: parameter estimates with 95% confidence bands*

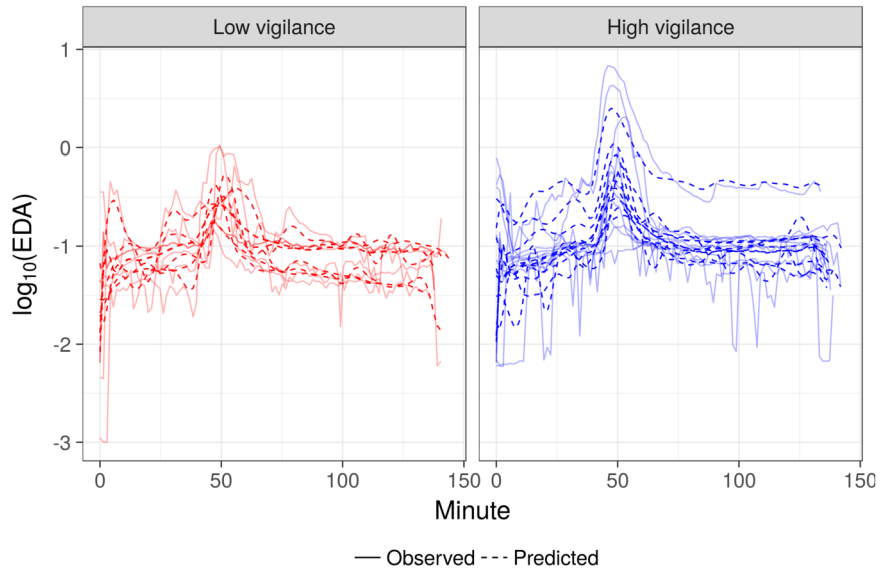


FIG 18. *Bayesian model: subject-specific predicted curves*

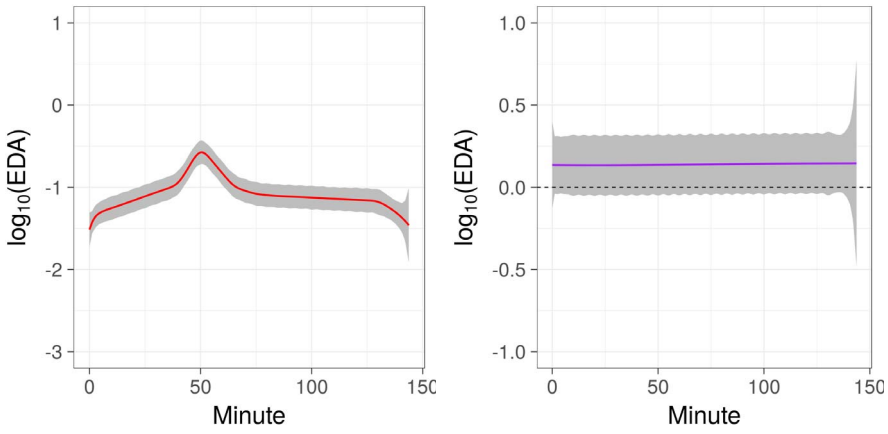
9.4. Alternative correlation structure

For comparison, we also fit ℓ_1 and ℓ_2 penalized models with alternative correlation structures similar to that recommended by Ruppert, Wand and Carroll [30, p. 192].

For the ℓ_1 penalized model, in place of the correlation structure implied by the penalty matrix S described above, we set the penalty matrix to $S := I_q$. While

this is a simplification of the correlation structure recommended by Ruppert, Wand and Carroll [30, p. 192], we think it offers a similar amount of flexibility.

Figure 19 shows the estimated marginal mean and 95% credible bands, and Figure 20 shows the subject-specific predicted curves. The point estimates shown in Figure 19 are similar to that shown in Figure 13. However, the confidence intervals in Figure 19 appear more reasonable. The subject-specific predicted curves shown in 20 are not shrunk towards the mean as much as in Figure 14.



(a) $\hat{\beta}_1(x)$ (low vigilance) (b) $\hat{\beta}_2(x)$ (high − low vigilance)

FIG 19. *ℓ_1 penalized model with alternative correlation structure: parameter estimates with 95% confidence bands*

For the ℓ_2 Penalized model, in place of the correlation structure implied by the penalty matrix S described above, we augmented each Z_i matrix on the left with the columns $[\mathbf{1}, \boldsymbol{x}_i]$, where $\boldsymbol{x}_i$ is an $n_i \times 1$ vector of measurement times for subject i. We then replaced $Z_i\boldsymbol{b}_i$ with $[\mathbf{1}, \boldsymbol{x}_i, Z_i](\boldsymbol{u}_i^T, \boldsymbol{b}_i^T)^T$, and assumed $(\boldsymbol{u}_i^T, \boldsymbol{b}_i^T)^T \sim N(\mathbf{0}, \Sigma_i)$ where

$$\Sigma_i = \begin{bmatrix} \Sigma' & \\ & \sigma_b^2 I_{q_i} \end{bmatrix}$$

and Σ' is a common 2×2 unstructured positive definite matrix. To model the within-subject correlations, we used a continuous autoregressive process of order 1. In particular, $\text{Cor}(y_i(x_{ij}), y_i(x_{ij'})) = \zeta^{|x_{ij} - x_{ij'}|}$ for a common parameter $\zeta > 0$.

Figure 21 shows the estimated marginal mean and 95% credible bands, and Figure 16 shows the subject-specific predicted curves. The estimates shown in Figure 21 are similar to that shown in Figure 15. While estimates of the difference between low and high vigilance subjects differs between this model and the ℓ_2 penalized model in Section 9.3, the more notable difference is in the subject-specific predicted curves. As seen in Figure 22, the predicted subject-specific curves are not shrunk towards the mean as much as in Figure 16.

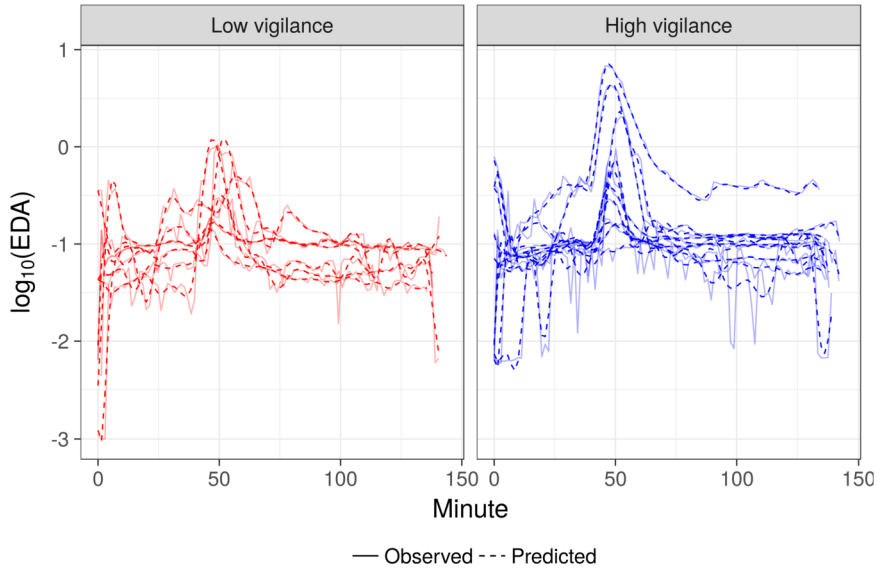
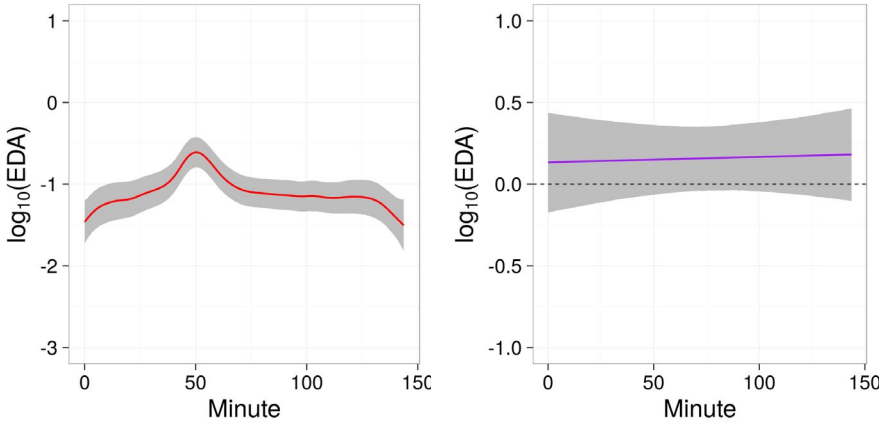


FIG 20. ℓ_1 *penalized model with alternative correlation structure: subject-specific predicted curves*



(a) $\hat{\beta}_1(x)$ (low vigilance subjects) (b) $\hat{\beta}_2(x)$ (high − low vigilance subjects)

FIG 21. ℓ_2 *penalized model with alternative correlation structure: parameter estimates with 95% confidence bands*

Table 4 shows the mean squared error (MSE) and computing time for the ℓ_1 penalized and ℓ_2 penalized models. In Table 4, computing time for the ℓ_1 penalized model does not include cross-validation, because the parameters were hand-tuned (with only 17 subjects and a complex random effects structure,

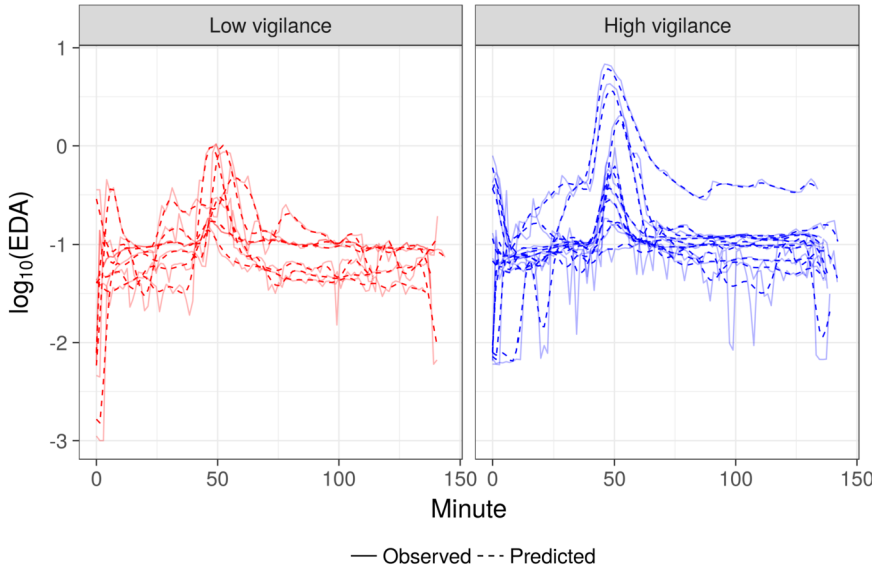


FIG 22. *ℓ_2 penalized model with alternative correlation structure: subject-specific predicted curves*

cross-validation did not find reasonable parameter values). As can be seen in Table 4, the alternative correlation structure led to smaller MSE for both the ℓ_1 and ℓ_2 penalized models, and less computing time for the ℓ_2 penalized model.

TABLE 4

MSE and computing time for different random curve correlation structures using the ℓ_1 penalized and ℓ_2 penalized models. "Smoothing" refers to smoothing splines used in Section 9.3 and "Alternative" refers to the correlation structures described in Section 9.4.

	ℓ_1 penalty		ℓ_2 penalty	
	Smoothing	Alternative	Smoothing	Alternative
MSE	0.0195	0.00649	0.0348	0.00764
Computing time (seconds)	10.9*	11.2*	166	56.9

*Does not include cross-validation (parameters hand-tuned)

Note: Models fit on a laptop with Intel i7 quad CPUs at 2.67 GHz and 8 GB memory

10. Discussion and potential extensions

As demonstrated in this article, P-splines with an ℓ_1 penalty can be useful for analyzing repeated measures data. Compared to related work with ℓ_1 penalties, our model is ambitious in that we allow for multiple smoothing parameters and propose approximate inferential procedures that do not require Bayesian estimation. However, these are also the two aspects of our proposed approach that require additional future work.

Regarding estimation, our current approach of using ADMM and CV appears to work reasonably well for random intercepts, but is not yet reliable for random

curves. In the future, we plan to develop more robust estimation techniques, particularly for smoothing parameters. As one possibility, we have done preliminary work to minimize quantities similar to GCV and AIC instead of the more computationally intensive CV, though these approaches do not seem as promising as their ℓ_2 counterparts. It may also be helpful to set the degrees of freedom prior to fitting the model. When possible, Bayesian estimation may be the most reliable way to currently fit these models. Bayesian estimation also opens the possibility of using other sparsity inducing priors, such as spike and slab models [18].

Regarding inference, in future work it may be possible to use the $\boldsymbol{\delta}$ quantity to bound difference between ℓ_1 and ℓ_2 penalized fits under certain assumptions on the data. It may also be helpful to investigate the use of post-selection inference methods to develop confidence bands for linear combinations of the active set, and to further investigate through simulations the performance of our proposed approximations of degrees of freedom. However, we note that our primary use of the degrees of freedom estimate $\hat{\mathrm{df}}$ is to obtain the residual degrees of freedom $\hat{\mathrm{df}}_{\mathrm{resid}} = n - \hat{\mathrm{df}}$, which we then use to estimate the variances $\hat{\sigma}^2_\epsilon = \|\boldsymbol{r}\|_2^2/\hat{\mathrm{df}}_{\mathrm{resid}}$. Therefore, when $n \gg \hat{\mathrm{df}}$, $\hat{\sigma}^2_\epsilon$ is not very sensitive to $\hat{\mathrm{df}}$, in which case it is not critical for our purposes to obtain an exact estimate of degrees of freedom.

As for P-splines with an ℓ_2 penalty, users must select both the order M of the B-splines and the order $k+1$ of the finite differences. These choices will depend on the scientific problem and analytical goals. Using $k = 1$ (2^{nd} order differences) is likely an appropriate starting point for most applications, and larger k could be used to increase the amount of smoothness.

For P-splines with an ℓ_2 penalty, in most cases the knot placement is not critical so long as the number of knots is large enough [29, 9]. We believe this also holds for P-splines with an ℓ_1 penalty, though further experimentation is needed to support this assumption. In practice, we recommend fitting models with a few different knot placements and widths to determine whether the model is sensitive to those choices for the data at hand.

Regarding the rate of convergence, from Observation 1 and the work of Tibshirani [39], we know that for equally spaced data and $F = I_n$, P-splines with an ℓ_1 penalty achieve the minimax rate of convergence for the class of weakly differentiable functions of bounded variation. When there are less knots than data points, we do not think it is possible to achieve the minimax rate of convergence. However, if the knots are selected well, it may be possible to achieve the same performance in practice.

It could also be useful to extend these results to generalized additive models to allow for non-normal responses, and to extend the approach of Sadhanala and Tibshirani [31] to include random effects and multiple smoothing parameters.

Acknowledgements

We thank the associate editor and two anonymous referees for their very insightful comments and suggestions, which greatly improved the paper.

Appendix A: Simulated demonstration with two smooths

In this appendix, we simulated data similar to that in Section 8, but with an additional varying-coefficient smooth. In particular, we simulated data for two groups with 50 subjects in each group and between 4 and 14 measurements per subject (900 total observations). The data for subject i at time t was generated as $y_{it} = \beta_0 + f_1(x_{it}) + f_2(x_{it})\mathbb{1}[\text{subject } i \text{ in Group 2}] + b_i + \epsilon_{it}$ where $b_i \sim N(0, \sigma_b^2)$ and $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ for $\sigma_b^2 = 1$ and $\sigma_\epsilon^2 = 0.01$. The true group means $f_1(x)$ and $f_2(x)$ are shown in Figure 23 and the simulated data are shown in Figure 24.

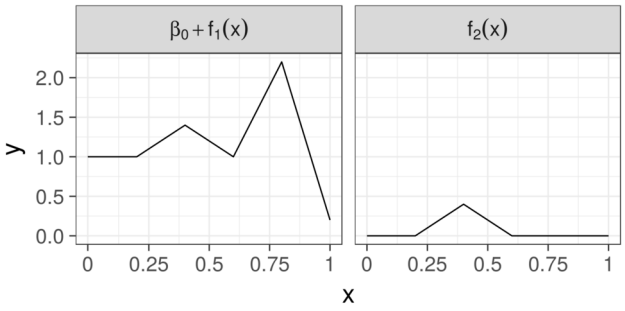


FIG 23. *True means*

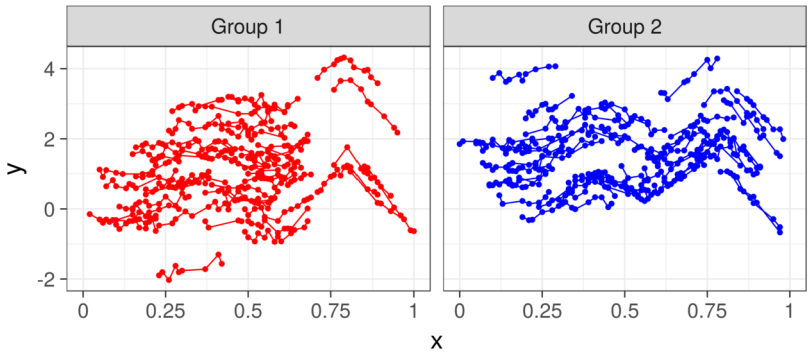


FIG 24. *Simulated data*

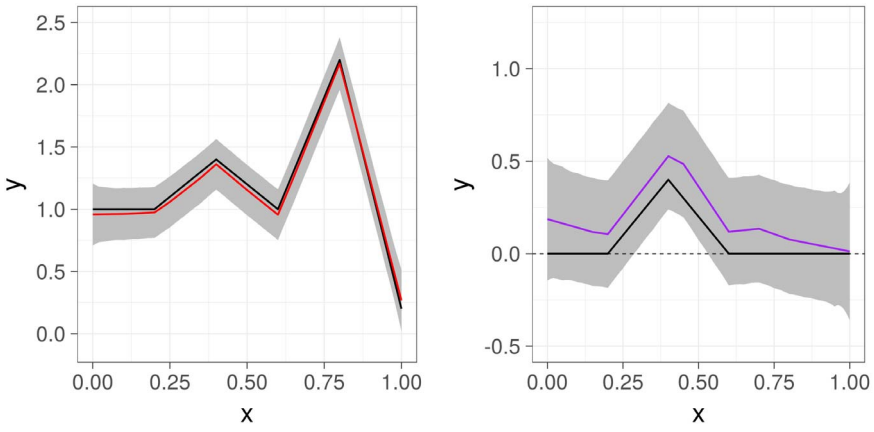
We fit a varying-coefficient model with $J = 2$ smooths to the data. In particular, we used ADMM and 5-fold CV to minimize

$$\underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta}_1 \in \mathbb{R}^{p-1}, \boldsymbol{\beta}_2 \in \mathbb{R}^p, \boldsymbol{b} \in \mathbb{R}^N}{\text{minimize}} \frac{1}{2}\|\boldsymbol{y} - \beta_0 \mathbf{1} - F_1\boldsymbol{\beta}_1 - F_2\boldsymbol{\beta}_2 - Z\boldsymbol{b}\|_2^2 + \lambda_1\|D^{(2)}\boldsymbol{\beta}_1\|_1 + \lambda_2\|D^{(2)}\boldsymbol{\beta}_2\|_1 + \tau\boldsymbol{b}^T\boldsymbol{b}. \tag{36}$$

where $F_1, F_2 \in \mathbb{R}^{n \times p}$ were formed with second order (first degree) B-splines and $p = 21$ basis functions, $F_2 = \text{diag}(\boldsymbol{u})F_1$ where $u_i = \mathbb{1}[\text{subject } i \text{ in Group 2}]$, and $Z_{il} = 1$ if observation i belongs to subject l and zero otherwise. We also fit

an equivalent model with an ℓ_2 penalty using the `mgcv` package [47], i.e. with $(\lambda_j/2)\|D^{(2)}\boldsymbol{\beta}_j\|_2^2$ in place of $\lambda_j\|D^{(2)}\boldsymbol{\beta}_j\|_1$ in (36), $j = 1, 2$.

The estimated curves are shown in Figure 25 for the ℓ_1 penalized model and in Figure 26 for the ℓ_2 penalized model. We used 5-fold CV to estimate the smoothing parameters λ_1, λ_2 and τ in the ℓ_1 penalized model, and LME updates to estimate σ_b^2 and $\boldsymbol{b}$ in the final model. As seen in Figures 25 and 26, the fits are similar, but the results with the ℓ_1 penalized model are slightly closer to the truth.



(a) $\hat{\beta}_0 + \hat{f}_1(x)$ (b) $\hat{f}_2(x)$

FIG 25. *Marginal mean and 95% credible intervals from ℓ_1 penalized model fit with ADMM and CV: black is true marginal mean, red is estimated marginal mean*

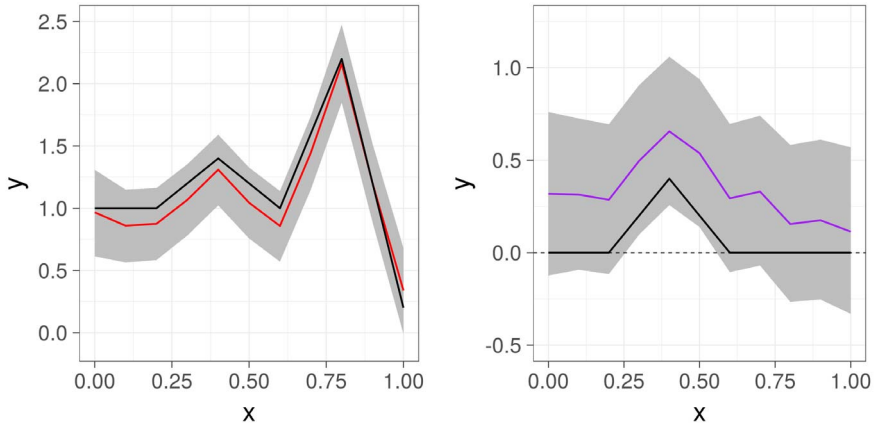
Table 5 shows the degrees of freedom and variance estimates with the ℓ_1 penalized and ℓ_2 penalized models. As seen in Table 5, variance estimates from both the ℓ_1 and ℓ_2 penalized models are very near the true values.

TABLE 5
Degrees of freedom and variance in ℓ_1 and ℓ_2 penalized models

	Penalty				
	ℓ_1		ℓ_2		
	$j=1$	$j=2$	$j=1$	$j=2$	Truth
df (ridge)	17.7	17.8	19.3	13.8	–
df (Stein)	12	9	–	–	–
$\hat{\sigma}_\epsilon^2$	0.0090		0.010		0.01
$\hat{\sigma}_b^2$	1.04		1.02		1

Appendix B: Details for $\boldsymbol{\lambda_j^{\max}}$

Letting $\boldsymbol{r}_j = \boldsymbol{y} - \beta_0\mathbf{1} - \sum_{\ell\neq j} F_\ell\boldsymbol{\beta}_\ell - Z\boldsymbol{b}$ be the j^{th} partial residuals, we can write the terms in (8) that involve $\boldsymbol{\beta}_j$ as $(1/2)\|\boldsymbol{r}_j - F_j\boldsymbol{\beta}_j\|_2^2 + \lambda_j\|D_j\boldsymbol{\beta}_j\|_1$. Then taking the sub-differential of (8) with respect to $\boldsymbol{\beta}_j$, we have



(a) $\hat{\beta}_0 + \hat{f}_1(x)$ (b) $\hat{f}_2(x)$

FIG 26. *Marginal mean and 95% credible intervals from ℓ_2 penalized model fit with mgcv [47]: black is true marginal mean, red is estimated marginal mean*

$$\mathbf{0} = -F_j^T(\boldsymbol{r}_j - F_j\hat{\boldsymbol{\beta}}_j) + D_j^T\lambda_j\boldsymbol{s}_j \tag{37}$$

for some $\boldsymbol{s}_j = (s_{j,1}, \ldots, s_{j,p_j-k_j-1})^T$ where

$$s_{j,\ell} \in \begin{cases} \{1\} & \text{if } (D\hat{\boldsymbol{\beta}}_j)_\ell > 1 \\ \{-1\} & \text{if } (D\hat{\boldsymbol{\beta}}_j)_\ell < 1 \\ [-1,1] & \text{if } (D\hat{\boldsymbol{\beta}}_j)_\ell = 0. \end{cases}$$

Solving (37) for $\hat{\boldsymbol{\beta}}_j$, we have $\hat{\boldsymbol{\beta}}_j = (F_j^TF_j)^{-1}F_j^T\boldsymbol{r}_j - D_j^T\lambda_j\boldsymbol{s}_j$. Multiplying through by D_j and noting that $D_jD_j^T$ is full rank and thus invertible, we have

$$(D_jD_j^T)^{-1}D_j\hat{\boldsymbol{\beta}}_j = (D_jD_j^T)^{-1}D_j(F_j^TF_j)^{-1}F_j^T\boldsymbol{r}_j - \lambda_j\boldsymbol{s}_j. \tag{38}$$

Setting $D_j\hat{\boldsymbol{\beta}}_j = \mathbf{0}$ in (38), we get that $(D_jD_j^T)^{-1}D_j(F_j^TF_j)^{-1}F_j^T\boldsymbol{r}_j = \lambda_j\boldsymbol{s}_j$ where $s_{j,\ell} \in [-1,1]$ for all ℓ. This can only hold if $\lambda_j = \|(D_jD_j^T)^{-1}D_j(F_j^TF_j)^{-1}F_j^T\boldsymbol{r}_j\|_\infty$, which gives us $\lambda_j^{\max}$.

Appendix C: Controlling total variation with the ℓ_1 penalty

Let $f(x) = \sum_{j=1}^p \beta_j\phi_j^M(x)$. Suppose the knots are equally spaced, and let $h_{M-k-1} = (M-k-1)/(t_{j+M-k-1} - t_j)$ for all j and $0 \le k < M-1$. Then on the interval $[t_M = x_{\min}, t_{p+1} = x_{\max}]$, from [4, p. 117] we have

$$\frac{d^{k+1}}{dx^{k+1}}f(x) = h_{M-1}\cdots h_{M-k-1}\sum_{j=k+2}^{p}\nabla^{k+1}\beta_j\phi_j^{M-k-1}(x) \tag{39}$$

where ∇^{k+1} is the $(k+1)^{th}$ order backwards difference.

Let $a_{k+1}^M = \max_{j\in\{k+2,\dots p\}} \int_{x_{\min}}^{x_{\max}} \phi_j^{M-k-1}(x)dx$. We note that a_{k+1}^M is finite and positive for all $0 \le k < M-1$. Then from (39), we have

$$\begin{aligned}
&\frac{1}{h_{M-1}\cdots h_{M-k-1}} \int_{x_{\min}}^{x_{\max}} \left|\frac{d^{k+1}}{dx^{k+1}} f(x)\right| dx \\
&\qquad = \int_{x_{\min}}^{x_{\max}} \left|\sum_{j=k+2}^{p} \nabla^{k+1}\beta_j \phi_j^{M-k-1}(x)\right| dx \\
&\qquad = \int_{x_{\min}}^{x_{\max}} \left|\sum_{j=k+2}^{p} (D^{(k+1)}\boldsymbol{\beta})_{j-k-1} \phi_j^{M-k-1}(x)\right| dx \\
&\qquad \le \int_{x_{\min}}^{x_{\max}} \sum_{j=k+2}^{p} \left|(D^{(k+1)}\boldsymbol{\beta})_{j-k-1} \phi_j^{M-k-1}(x)\right| dx \\
&\qquad = \sum_{j=k+2}^{p} \int_{x_{\min}}^{x_{\max}} \left|(D^{(k+1)}\boldsymbol{\beta})_{j-k-1} \phi_j^{M-k-1}(x)\right| dx \\
&\qquad = \sum_{j=k+2}^{p} \left|(D^{(k+1)}\boldsymbol{\beta})_{j-k-1}\right| \int_{x_{\min}}^{x_{\max}} \phi_j^{M-k-1}(x)dx \qquad (40) \\
&\qquad \le a_{k+1}^M \sum_{j=k+2}^{p} \left|(D^{(k+1)}\boldsymbol{\beta})_{j-k-1}\right| \\
&\qquad = a_{k+1}^M \|D^{(k+1)}\boldsymbol{\beta}\|_1 \qquad (41)
\end{aligned}$$

where (40) follows because $\phi_j^{M-k-1}(x) \ge 0\ \forall x \in \mathbb{R}$.

Rewriting (41), for $0 \le k < M-1$ we have

$$\int_{x_{\min}}^{x_{\max}} \left|\frac{d^{k+1}}{dx^{k+1}} f(x)\right| dx \le C_{M,k+1} \|D^{(k+1)}\boldsymbol{\beta}\|_1$$

where $C_{M,k+1} = a_{k+1}^M h_{M-1}\cdots h_{M-k-1}$ is a constant. This shows that controlling the ℓ_1 norm of the $(k+1)^{th}$ order finite differences in coefficients also controls the total variation of the k^{th} derivative of the function.

Supplementary Material

Supplementary material for "P-splines with an ℓ_1 penalty for repeated measures"

(doi: 10.1214/18-EJS1487SUPP; .zip). Code and R package for all simulations and analyses. These materials are also available at https://github.com/bdsegal/code-for-psplinesl1-paper (code) and https://github.com/bdsegal/psplinesl1 (R package).

References

[1] Bollaerts, K., Eilers, P. H. C. and Aerts, M. (2006). Quantile regression with monotonicity restrictions using P-splines and the L1-norm. *Statistical Modelling* **6** 189–207. MR2252359

[2] Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* **3** 1–122.

[3] Chen, H. and Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics* **67** 861–870. MR2829260

[4] De Boor, C. (2001). *A practical guide to splines*, Revised ed. Springer, New York, NY. MR1900298

[5] Donoho, D. L. and Johnstone, I. M. (1988). Minimax estimation via wavelet shrinkage. *The Annals of Statistics* **26** 879–921. MR1635414

[6] Efron, B. (1986). How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association* **81** 461–470. MR0845884

[7] Eilers, P. H. C. (2000). Robust and Quantile Smoothing with P-splines and the L1 Norm. In *Proceedings of the 15th International Workshop on Statistical Modelling, Bilbao*.

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

[9] Eilers, P. H. C., Marx, B. D. and Durbán, M. (2015). Twenty years of P-splines. *SORT: statistics and operations research transactions* **39** 149–186. MR3467488

[10] Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008). *Longitudinal data analysis*. Chapman and Hall/CRC, Boca Raton, FL. MR1500110

[11] Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2** 1360–1383. MR2655663

[12] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014). *Bayesian data analysis*, 3rd ed. Chapman and Hall/CRC, Boca Raton, FL. MR3235677

[13] Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review* **55** 245–259. MR0963142

[14] Guo, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050

[15] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1** 297–318. MR0858512

[16] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, 1st ed. *Monographs on Statistics and Applied Probability*. Chapman & Hall, London. MR1082147

[17] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **55** 757–796. MR1229881

[18] Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics* **33** 730–773. MR2163158

[19] Janson, L., Fithian, W. and Hastie, T. J. (2015). Effective degrees of freedom: a flawed metaphor. *Biometrika* 1–8. MR3371017

[20] Kim, S.-J., Koh, K., Boyd, S. and Gorinevsky, D. (2009). ℓ_1 Trend Filtering. *SIAM review* **51** 339–360. MR2505584

[21] Lin, Y., Zhang, H. H. et al. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* **34** 2272–2297. MR2291500

[22] Lou, Y., Bien, J., Caruana, R. and Gehrke, J. (2016). Sparse Partially Linear Additive Models. *Journal of Computational and Graphical Statistics* **25**. MR3572032

[23] Mammen, E., van de Geer, S. et al. (1997). Locally adaptive regression splines. *The Annals of Statistics* **25** 387–413. MR1429931

[24] Meier, L., Van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics* **37** 3779–3821. MR2572443

[25] Petersen, A., Witten, D. and Simon, N. (2016). Fused lasso additive model. *Journal of Computational and Graphical Statistics* **25** 1005–1025. MR3572026

[26] Ramdas, A. and Tibshirani, R. J. (2016). Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics* **25** 839–858. MR3533641

[27] Ravikumar, P., Lafferty, J. D., Liu, H. and Wasserman, L. (2009). Sparse Additive Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**. MR2750255

[28] Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259. MR1833314

[29] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics* **11** 735–757. MR1944261

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

[31] Sadhanala, V. and Tibshirani, R. J. (2017). Additive Models with Trend Filtering. *arXiv preprint arXiv:1702.05037*.

[32] Scheipl, F., Staicu, A.-M. and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics* **24** 447–501. MR3357391

[33] Segal, B. D., Elliott, M. R., Braun, T. and Jiang, H. (2018). Supplementary material for "P-splines with an ℓ_1 penalty for repeated measures".

[34] Speed, T. (1991). Comment on "That BLUP is a Good Thing: The Estimation of Random Effects". *Statistical science* **6** 42–44.

[35] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9** 1135–1151. MR0630098
[36] Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.14.1.
[37] R Core Team (2017). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
[38] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242
[39] Tibshirani, R. J. (2014a). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* **42** 285–323. MR3189487
[40] Tibshirani, R. J. (2014b). Supplement to "Adaptive piecewise polynomial estimation via trend filtering". MR3189487
[41] Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **2** 1198–1232. MR2985948
[42] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108. MR2136641
[43] Wahba, G. (1990). *Spline models for observational data*. Society for industrial and applied mathematics, Philadelphia, PA. MR1045442
[44] Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 159–174. MR1625640
[45] Wang, Y.-X., Smola, A. and Tibshirani, R. (2014). The falling factorial basis and its statistical applications. In *International Conference on Machine Learning* 730–738.
[46] Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* **99** 673–686. MR2090902
[47] Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, Boca Raton, FL. MR2206355
[48] Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 3–36. MR2797734
[49] Wood, S. N., Goude, Y. and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64** 139–155. MR3293922
[50] Wood, S. N., Pya, N. and Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association* **111** 1548–1575. MR3601714
[51] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67. MR2212574

[52] Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93** 710–719. MR1631369

[53] Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37** 3468–3497. MR2549566

[54] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320. MR2137327