Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Bayesian Function-on-Scalars Regression for High-Dimensional Data

Daniel R. Kowal & Daniel C. Bourgeois

To cite this article: Daniel R. Kowal & Daniel C. Bourgeois (2020): Bayesian Function-on-Scalars Regression for High-Dimensional Data, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2019.1710837

To link to this article: https://doi.org/10.1080/10618600.2019.1710837

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2020, VOL. 00, NO. 0, 1–10
https://doi.org/10.1080/10618600.2019.1710837

Taylor & Francis
Taylor & Francis Group

Bayesian Function-on-Scalars Regression for High-Dimensional Data

Daniel R. Kowal and Daniel C. Bourgeois

Department of Statistics, Rice University, Houston, TX

ABSTRACT

We develop a fully Bayesian framework for function-on-scalars regression with many predictors. The functional data response is modeled nonparametrically using unknown basis functions, which produces a flexible and data-adaptive functional basis. We incorporate shrinkage priors that effectively remove unimportant scalar covariates from the model and reduce sensitivity to the number of (unknown) basis functions. For variable selection in functional regression, we propose a decision theoretical posterior summarization technique, which identifies a subset of covariates that retains nearly the predictive accuracy of the full model. Our approach is broadly applicable for Bayesian functional regression models, and unlike existing methods provides joint rather than marginal selection of important predictor variables. Computationally scalable posterior inference is achieved using a Gibbs sampler with linear time complexity in the number of predictors. The resulting algorithm is empirically faster than existing frequentist and Bayesian techniques, and provides joint estimation of model parameters, prediction and imputation of functional trajectories, and uncertainty quantification via the posterior distribution. A simulation study demonstrates improvements in estimation accuracy, uncertainty quantification, and variable selection relative to existing alternatives. The methodology is applied to actigraphy data to investigate the association between intraday physical activity and responses to a sleep questionnaire. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2018
Revised November 2019

KEYWORDS
Actigraphy data; Factor model; MCMC; Shrinkage; Variable selection

1. Introduction

Modern scientific measuring systems commonly record data over a continuous domain, often at high resolutions, and referred to as *functional data*. Functional data are typically high dimensional and highly correlated, and may be measured concurrently with other variables of interest. We consider the problem of *function-on-scalars regression* (FOSR), in which a functional data response is modeled using (possibly many) *scalar* predictors. Applications of FOSR are broad and impactful: examples include blood pressure profiles during pregnancy (Montagna et al. 2012), human motor control following a stroke (Chen, Goldsmith, and Ogden 2016), age-specific fertility rates (Kowal 2018), microarray time course gene expression data (Wang, Chen, and Li 2007), and longitudinal genome-wide association studies (Barber, Reimherr, and Schill 2017; Fan and Reimherr 2017), among others (Ramsay and Silverman 2005; Morris 2015).

FOSR shares the same fundamental goals as multiple regression—estimation and inference for the regression coefficients, selection of important predictor variables, and prediction of new responses—with additional modeling challenges. Within-curve dependence of functional data requires careful modeling of the covariance function, which may be complex, with implications for computational scalability. In addition, the regression coefficients in FOSR are *functions*, which complicates estimation, inference, and selection. Naturally, these challenges are compounded in high-dimensional settings. Finally, prediction in FOSR requires prediction of functional trajectories, which may be partially observed.

To address these challenges, we develop a fully Bayesian framework for FOSR, where the number of predictors p may be greater than the number of observed curves n. Within-curve dependence is modeled nonparametrically using *unknown* basis functions, which produces a data-adaptive functional basis with uncertainty quantification via the posterior distribution. We introduce shrinkage priors to mitigate the impact of unimportant predictor variables and reduce sensitivity to the number of basis functions. For computationally scalable posterior inference, we develop a Gibbs sampling algorithm with linear time complexity in either n or p. The methodology is applicable for densely or sparsely observed functional data, with automatic prediction and imputation at unobserved points within the Gibbs sampler.

Recent developments in FOSR have focused on variable selection, which requires selecting or thresholding entire regression coefficient *functions*. Accordingly, various group penalties have been proposed for variable selection in FOSR, including a group smoothly clipped absolute deviation (Wang, Chen, and Li 2007) and a group-minimax concave penalty (Chen, Goldsmith, and Ogden 2016) for a moderate number of predictors, and a group lasso (Barber, Reimherr, and Schill 2017) and an adaptive group lasso (Fan and Reimherr 2017) for high-dimensional predictors. Although these frequentist methods provide both estimation and selection, finite-sample inference is unavailable and

CONTACT Daniel R. Kowal daniel.kowal@rice.edu Department of Statistics, Rice University, Houston, TX 77251-1892.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JCGS.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

important tuning parameters must be selected, which is often computationally intensive. By comparison, Bayesian approaches provide exact finite-sample inference (up to MCMC error) and do not require selection of tuning parameters. Bayesian FOSR models may identify important predictors marginally using *global Bayesian p-values* (GBPVs; Meyer et al. 2015). However, there is currently no coherent Bayesian framework for simultaneous estimation, inference, and variable selection in FOSR, particularly for moderate to large p.

We propose a decision theoretical approach for Bayesian variable selection in FOSR. Using a loss function that balances sparsity with predictive accuracy for functional responses, we obtain sparse posterior summaries by optimizing the expected loss under the posterior predictive distribution. As a result, we identify a subset of predictors that maintains nearly the predictive accuracy of the full model. Notably, our approach selects variables *jointly* rather than *marginally*, is applicable under a variety of priors, and can be tractably solved using existing software for penalized regression. The methodology may be viewed as an extension of Hahn and Carvalho (2015) to the functional data setting.

We apply our methods to study the association between intraday physical activity and responses to a sleep questionnaire among elderly adults. It is a priori unclear which questionnaire items, if any, are associated with physical activity, as measured by actigraphy data, and whether or not such associations vary throughout the day. The proposed methodology provides the framework to (i) model and impute intraday physical activity trajectories for each individual, (ii) estimate regression coefficient functions and accompanying posterior credible bands, and (iii) select a small subset of the questionnaire items and demographic variables ($p = 74$) that maintain nearly the predictive accuracy of the full model.

The remainder of the article is organized as follows: the model is in Section 2; the MCMC algorithm is in Section 3; the variable selection method is in Section 4; a simulation analysis is in Section 5; the application is in Section 6; we conclude in Section 7. MCMC diagnostics, proofs of theoretical results, and details on the application are provided in the supplementary materials. An `R` package is available on GitHub and example code is available as a supplementary materials.

2. A Bayesian Function-on-Scalars Regression Model

2.1. Model Specification and Assumptions

Let $Y_1(\tau), \ldots, Y_n(\tau)$ be a functional data response with $\tau \in \mathcal{T}$, where $\mathcal{T} \subset \mathbb{R}^D$ is a compact index set and $D \in \mathbb{Z}^+$. Suppose we have p scalar predictors $\{x_{i,j}\}_{j=1}^p$ for $i = 1, \ldots, n$, possibly with $p > n$. We are interested in modeling the association between the scalar predictors $x_{i,j}$ and the functional response Y_i. We propose the following Bayesian FOSR model:

$$Y_i(\tau) = \sum_{k=1}^{K} f_k(\tau)\beta_{k,i} + \epsilon_i(\tau), \tag{1}$$

$$\epsilon_i(\tau) \overset{\text{indep}}{\sim} N(0, \sigma_\epsilon^2), \quad \tau \in \mathcal{T},$$

$$\beta_{k,i} = \mu_k + \sum_{j=1}^{p} x_{i,j}\alpha_{j,k} + \gamma_{k,i}, \tag{2}$$

$$\mu_k \overset{\text{indep}}{\sim} N(0, \sigma_{\mu_k}^2), \quad \alpha_{j,k} \overset{\text{indep}}{\sim} N(0, \sigma_{\alpha_{j,k}}^2),$$

$$\gamma_{k,i} \overset{\text{indep}}{\sim} N(0, \sigma_{\gamma_{k,i}}^2). \tag{3}$$

Model (1) expands the functional data $Y_i(\cdot)$ in the basis $\{f_k(\cdot)\}$ with a regression model for the corresponding basis coefficients $\{\beta_{k,i}\}$. The basis functions $\{f_k(\tau)\}$, which may be known (e.g., splines or wavelets) or unknown (see Section 2.2), capture within-curve dependence of the functional data Y_i, while the coefficients $\{\beta_{k,i}\}$ model between-curve dependence induced by the predictor variables $x_{i,j}$. The intercepts $\{\mu_k\}$, the regression coefficients $\{\alpha_{j,k}\}$, and the subject-specific errors $\{\gamma_{k,i}\}$ are given conditionally Gaussian priors (see Section 2.3), and we assume a Jeffreys' prior for the observation error variance, $[\sigma_\epsilon^2] \propto 1/\sigma_\epsilon^2$.

Models (1)–(3) may be expressed as a more conventional FOSR model. Let $\mathcal{GP}(c, C)$ denote a Gaussian process with mean function c and covariance function C.

Proposition 1. Models (1)–(3) implies the functional regression model

$$Y_i(\tau) = \tilde{\mu}(\tau) + \sum_{j=1}^{p} x_{i,j}\tilde{\alpha}_j(\tau) + \tilde{\gamma}_i(\tau) + \epsilon_i(\tau), \tag{4}$$

$$\epsilon_i(\tau) \overset{\text{indep}}{\sim} N(0, \sigma_\epsilon^2), \quad \tau \in \mathcal{T}$$

with expansions $\tilde{\mu}(\tau) \equiv \sum_{k=1}^K f_k(\tau)\mu_k \sim \mathcal{GP}(0, C_\mu)$ for $C_\mu(\tau, u) = \sum_{k=1}^K f_k(\tau)f_k(u)\sigma_{\mu_k}^2$, $\tilde{\alpha}_j(\tau) \equiv \sum_{k=1}^K f_k(\tau)\alpha_{j,k} \overset{\text{indep}}{\sim} \mathcal{GP}(0, C_{\alpha_j})$ for $C_{\alpha_j}(\tau, u) = \sum_{k=1}^K f_k(\tau)f_k(u)\sigma_{\alpha_{j,k}}^2$, and $\tilde{\gamma}_i(\tau) \equiv \sum_{k=1}^K f_k(\tau)\gamma_{k,i} \overset{\text{indep}}{\sim} \mathcal{GP}(0, C_{\gamma_i})$ for $C_{\gamma_i}(\tau, u) = \sum_{k=1}^K f_k(\tau)f_k(u)\sigma_{\gamma_{k,i}}^2$.

The predictors $x_{i,j}$ are directly associated with the functional data $Y_i(\tau)$ via the regression coefficient functions $\tilde{\alpha}_j(\tau) = \sum_k f_k(\tau)\alpha_{j,k}$. The subject-specific error term $\tilde{\gamma}_i(\tau)$ in (4) captures within-curve variability in $Y_i(\tau)$ unexplained by $\{x_{i,j}\}$, which marginally produces a model for within-curve correlations of the FOSR error. Accounting for within-curve error correlations is important for statistically efficient estimation and valid inference in FOSR (Reiss, Huang, and Mennes 2010), especially for high-dimensional predictors (Chen, Goldsmith, and Ogden 2016). Implicit in Model (4) is the assumption that each function $\tilde{\mu}$, $\tilde{\alpha}_j$, and $\tilde{\gamma}_i$ may be expanded in the same basis $\{f_k\}_{k=1}^K$. This assumption may be relaxed, analogous to Di et al. (2009), with appropriate modifications of the computations in Section 3.

2.2. Modeling the Basis Functions

For broad applicability, the basis functions $\{f_k\}$ in (1) should be flexible, efficient for computations, and well-defined for $\mathcal{T} \subset \mathbb{R}^D$ with $D \in \mathbb{Z}^+$. The dimension K is important: each predictor variable $j = 1, \ldots, p$ is accompanied by K coefficients, which may be correlated. FOSR methods that use full basis expansions,

such as splines or wavelets, have been successfully applied for small p (Ramsay and Silverman 2005), including extensions for mixed effects models (Guo 2002; Morris and Carroll 2006; Zhu, Brown, and Morris 2011; Goldsmith and Kitago 2016), but are neither parsimonious nor computationally scalable for moderate to large p. One remedy is to precompute a lower-dimensional basis, such as a functional principal components (FPC) basis. However, this approach implicitly conditions on the estimates FPCs and fails to account for the accompanying uncertainty. Goldsmith, Greven, and Crainiceanu (2013) find that FPC-based methods can substantially underestimate total variability, even for densely sampled functional data.

We instead model each f_k as an unknown, smooth function. By modeling $\{f_k\}$ as unknown, we simultaneously (i) produce a data-adaptive functional basis, which minimizes the number of basis functions K needed, and (ii) incorporate uncertainty quantification of $\{f_k\}$ via the posterior distribution. Smoothness encourages information sharing among nearby points, which reduces variability to produce more stable prediction at unobserved points. Model identifiability is enforced by coupling a matrix orthonormality constraint on $\{f_k\}$ with an ordering constraint on the variance components in (3) (see Section 3).

We adopt the model for $\{f_k\}$ in Kowal (2018), which is closely related to Bayesian (probabilistic) FPCA (Suarez and Ghosal 2017) and reduced-rank functional data models (Montagna et al. 2012; Goldsmith, Zipunnikov, and Schrack 2015), yet offers substantial computational improvements. Let $f_k(\tau) = \boldsymbol{b}'(\tau)\boldsymbol{\psi}_k$ for known basis functions $\boldsymbol{b}'(\tau) = (b_1(\tau), \ldots, b_{L_m}(\tau))$ and unknown basis coefficients $\boldsymbol{\psi}_k$. We use low rank thin plate splines (LR-TPS) for $\boldsymbol{b}(\cdot)$, which are well-defined for $\mathcal{T} \subset \mathbb{R}^D$ with $D \in \mathbb{Z}^+$ and efficient in MCMC samplers (Crainiceanu, Ruppert, and Wand 2005). Smoothness is encouraged via the prior $\boldsymbol{\psi}_k \sim N(\mathbf{0}, \lambda_{f_k}^{-1}\boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\Omega}$ is a $L_m \times L_m$ known roughness penalty matrix. The smoothing parameter λ_{f_k} appears as a prior precision, so we assign a uniform prior distribution on the corresponding standard deviation, $\lambda_{f_k}^{-1/2} \overset{\text{iid}}{\sim} \text{Uniform}(0, 10^4)$ (Kowal, Matteson, and Ruppert 2017). Details on the construction of $\boldsymbol{b}(\cdot)$ and $\boldsymbol{\Omega}$ are given in Kowal (2018) and the relevant full conditional distributions are in Section 3.

2.3. Shrinkage Priors

The regression model (2) may include unimportant predictors, especially for moderate to large p. Without regularization or shrinkage, such irrelevant predictors can reduce estimation accuracy and statistical efficiency. Models (1)–(3) also requires a choice of K. While K may be treated as unknown and estimated in the model, this approach typically requires computationally intensive procedures, such as reversible jump MCMC (Suarez and Ghosal 2017). Instead, we impose ordered shrinkage with respect to $k = 1, \ldots, K$, so that larger number factors are a priori less important. Ordered shrinkage is computationally scalable and empirically reduces sensitivity to the choice of K, provided K is chosen sufficiently large.

We include a groupwise horseshoe prior for the regression coefficients $\alpha_{j,k} \overset{\text{indep}}{\sim} N(0, \sigma^2_{\alpha_{j,k}})$, which extends Carvalho, Polson, and Scott (2010). Shrinkage is applied at both the factor-within-predictor level as well as the predictor-level using a hierarchy of half-Cauchy distributions:

$$\sigma_{\alpha_{j,k}} \overset{\text{ind}}{\sim} C^+(0, \lambda_j), \quad \lambda_j \overset{\text{ind}}{\sim} C^+(0, \lambda_0), \quad \lambda_0 \overset{\text{ind}}{\sim} C^+(0, p^{-1/2}). \tag{5}$$

The scale parameter $\sigma_{\alpha_{j,k}}$ controls the factor k shrinkage for predictor j, λ_j determines the shrinkage for all regression coefficients $\{\alpha_{j,k}\}_{k=1}^K$ for predictor j, and λ_0 corresponds to the global level of sparsity for all predictors $j = 1, \ldots, p$, and is scaled by $p^{-1/2}$ following Piironen and Vehtari (2016). The horseshoe prior and its variants have been successful in a variety of models and applications, including functional regression (Kowal 2018; Kowal, Matteson, and Ruppert 2019).

We apply the multiplicative gamma process (MGP) for ordered shrinkage (Bhattacharya and Dunson 2011). MGP priors for the intercepts $\{\mu_k\}$ and the subject-specific errors $\{\gamma_{k,i}\}$ are represented via priors on the respective variance components in (3). The intercept prior precisions are $\sigma_{\mu_k}^{-2} = \prod_{\ell \le k} \delta_{\mu_\ell}$, where $\delta_{\mu_1} \sim \text{Gamma}(a_{\mu_1}, 1)$ and $\delta_{\mu_\ell} \sim \text{Gamma}(a_{\mu_2}, 1)$ for $\ell > 1$, which implies a stochastic ordering for $\sigma^2_{\mu_k}$ when $a_{\mu_1} > 0$ and $a_{\mu_2} \ge 2$ (Bhattacharya and Dunson 2011). For the subject-specific errors, we let $\sigma^2_{\gamma_{k,i}} = \sigma^2_{\gamma_k}/\xi_{\gamma_{k,i}}$ with $\sigma_{\gamma_k}^{-2} = \prod_{\ell \le k} \delta_{\gamma_\ell}$, $\delta_{\gamma_1} \sim \text{Gamma}(a_{\gamma_1}, 1)$, $\delta_{\gamma_\ell} \sim \text{Gamma}(a_{\gamma_2}, 1)$ for $\ell > 1$, and $\xi_{\gamma_{k,i}} \overset{\text{iid}}{\sim} \text{Gamma}(\nu_\gamma/2, \nu_\gamma/2)$, as in Bhattacharya and Dunson (2011) and Montagna et al. (2012). The hyperpriors $a_{\mu_1}, a_{\mu_2}, a_{\gamma_1}, a_{\gamma_2} \overset{\text{iid}}{\sim} \text{Gamma}(2, 1)$ allow the data to determine the rate of ordered shrinkage separately for $\{\mu_k\}$ and $\{\gamma_{k,i}\}$. Lastly, the hyperprior $\nu_\gamma \sim \text{Uniform}(2, 128)$ induces heavy tails in the marginal distribution of $\gamma_{k,i}$ for additional model robustness.

3. MCMC Sampling Algorithm

We construct an MCMC sampling algorithm that consists of efficient closed form sampling steps (with the exception of the shrinkage prior hyperparameters $a_{\mu_1}, a_{\mu_2}, a_{\gamma_1}, a_{\gamma_2}, \nu_\gamma$, which alternatively may be fixed in advanced). The main blocks of the sampling algorithm are (i) the basis functions $\{f_k\}$ in (1), (ii) the regression coefficients $\{\mu_k, \alpha_{j,k}, \gamma_{k,i}\}$ in (2), and (iii) the variance components in (1) and (3). An overview of the algorithm is presented here, with details provided in the supplement.

Suppose we observe the functional data Y_i at observation points $\{\tau_\ell\}_{\ell=1}^m$. For notational convenience, we assume the observation points are identical for all subjects i, but later relax that assumption (see Section 6). The likelihood in (1) becomes

$$\boldsymbol{Y}_i = \sum_{k=1}^K \boldsymbol{f}_k \beta_{k,i} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2_\epsilon \boldsymbol{I}_m), \tag{6}$$

where $\boldsymbol{Y}_i = (Y_i(\tau_1), \ldots, Y_i(\tau_m))'$ and $\boldsymbol{f}_k = (f_k(\tau_1), \ldots, f_k(\tau_m))'$. For functional data, often m is large and the components of $\boldsymbol{Y}_i$ are highly correlated. Therefore, MCMC sampling algorithms must be constructed carefully to ensure both computational and MCMC efficiency.

First, we sample the unknown basis functions $f_k(\tau) = \boldsymbol{b}'(\tau)\boldsymbol{\psi}_k$ by iteratively drawing from the full conditional distributions $[\boldsymbol{\psi}_k | \cdots] \sim N\left(\boldsymbol{Q}_{\psi_k}^{-1}\boldsymbol{\ell}_{\psi_k}, \boldsymbol{Q}_{\psi_k}^{-1}\right)$ for $k = 1, \ldots, K$,

where $\boldsymbol{Q}_{\psi_k} = \sigma_\epsilon^{-2}(\boldsymbol{B}'\boldsymbol{B})\sum_{i=1}^n \beta_{k,i}^2 + \lambda_{f_k}\boldsymbol{\Omega}$ and $\boldsymbol{\ell}_{\psi_k} = \sigma_\epsilon^{-2}\boldsymbol{B}'\sum_{i=1}^n \left[\beta_{k,i}\left(\boldsymbol{Y}_i - \sum_{k'\neq k}\boldsymbol{f}_{k'}\beta_{k',i}\right)\right]$, and set $\boldsymbol{f}_k = \boldsymbol{B}\boldsymbol{\psi}_k$ for $\boldsymbol{B} = (\boldsymbol{b}(\tau_1),\ldots,\boldsymbol{b}(\tau_m))'$. In addition, we enforce the matrix orthonormality constraint $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$ on the basis matrix $\boldsymbol{F} = (\boldsymbol{f}_1,\ldots,\boldsymbol{f}_K)$ during sampling: for each k, we (i) condition on the (linear) orthogonality constraints $\boldsymbol{f}_k'\boldsymbol{f}_\ell = 0$ for $\ell \neq k$ and (ii) rescale $\boldsymbol{f}_k$ to unit norm. Notably, $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$ is satisfied for *every* MCMC iteration. This constraint, coupled with the ordered MGP prior in Section 2.3, provides identifiability (up to sign changes, which in our experience are not problematic in the MCMC sampler).

The matrix orthonormality constraint $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$ also provides important simplifications of the challenging likelihood in (2), which we leverage to achieve substantial improvements in computational efficiency for sampling the parameters in (2) and (3). These computational gains are essential for moderate to large p, and produce algorithms for fully Bayesian inference that are empirically faster than existing alternatives. Consider the following:

Lemma 1. Under the constraint $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$, the joint likelihood in (6) for $\{\beta_{k,i}\}$ is equivalent to the *working likelihood* implied by

$$y_{k,i} = \beta_{k,i} + e_{k,i}, \quad e_{k,i} \overset{\text{indep}}{\sim} N(0,\sigma_\epsilon^2) \tag{7}$$

up to a constant that does not depend on $\beta_{k,i}$, where $y_{k,i} = \boldsymbol{f}_k'\boldsymbol{Y}_i$ and $e_{k,i} = \boldsymbol{f}_k'\boldsymbol{\epsilon}_i$.

The utility of Lemma 1 is that, for sampling the parameters in (2) and (3) which depend on $\beta_{k,i}$, we may replace the likelihood (6) with the simpler (7). Notably, (7) depends on m only via the projection $y_{k,i} = \boldsymbol{f}_k'\boldsymbol{Y}_i$ and eliminates correlations among the components of $\boldsymbol{Y}_i$.

We sample the regression parameters $\{\mu_k, \alpha_{j,k}, \gamma_{k,i}\}_{j,k,i}$ jointly, which improves MCMC efficiency, and is a common strategy in Bayesian mixed effects models. This is accomplished in two steps: (i) sample the regression coefficients $\{\mu_k, \alpha_{j,k}\}$ after marginalizing over the subject-specific effects $\{\gamma_{k,i}\}$, and (ii) sample $\{\gamma_{k,i}\}$ conditional on $\{\mu_k, \alpha_{j,k}\}$. The full conditional distributions are (i) $[\boldsymbol{\alpha}_k|\boldsymbol{Y},\ldots] \sim N\left(\boldsymbol{Q}_{\alpha_k}^{-1}\boldsymbol{\ell}_{\alpha_k}, \boldsymbol{Q}_{\alpha_k}^{-1}\right)$ for $\boldsymbol{\alpha}_k = (\mu_k, \alpha_{1,k},\ldots,\alpha_{p,k})'$, where $\boldsymbol{Q}_{\alpha_k} = \boldsymbol{X}'\boldsymbol{\Sigma}_{y_k}^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_{\alpha_k}^{-1}$ and $\boldsymbol{\ell}_{\alpha_k} = \boldsymbol{X}'\boldsymbol{\Sigma}_{y_k}^{-1}\boldsymbol{y}_k$ for $n \times (p+1)$ design matrix $\boldsymbol{X}$, marginal variance $\boldsymbol{\Sigma}_{y_k} = \text{diag}\left(\{\sigma_{\gamma_{k,i}}^2 + \sigma_\epsilon^2\}_{i=1}^n\right)$, prior variance $\boldsymbol{\Sigma}_{\alpha_k} = \text{diag}\left(\sigma_{\mu_k}^2, \sigma_{\alpha_{1,k}}^2, \ldots, \sigma_{\alpha_{p,k}}^2\right)$, and projected data $\boldsymbol{y}_k = (y_{k,1},\ldots,y_{k,n})'$; and (ii) $\left[\gamma_{k,i}|\boldsymbol{Y}, \{\mu_k, \alpha_{j,k}\}, \ldots\right] \overset{\text{indep}}{\sim} N\left(Q_{\gamma_{k,i}}^{-1}\ell_{\gamma_{k,i}}, Q_{\gamma_{k,i}}^{-1}\right)$, where $Q_{\gamma_{k,i}} = \sigma_\epsilon^{-2} + \sigma_{\gamma_{k,i}}^{-2}$ and $\ell_{\gamma_{k,i}} = \sigma_\epsilon^{-2}\left(y_{k,i} - \mu_k - \sum_{j=1}^p x_{i,j}\alpha_{j,k}\right)$. A key feature of the proposed sampling strategy is that both component samplers (i) and (ii) are efficient. For the regression sampler (i), we apply Rue (2001) when $p < n$ with computational complexity $\mathcal{O}(p^3)$ and Bhattacharya, Chakraborty, and Mallick (2016) when $p > n$ with computational complexity $\mathcal{O}(n^2 p)$. The Bhattacharya, Chakraborty, and Mallick (2016) sampler is designed for high-dimensional Bayesian regression, but to the best of our knowledge has not been incorporated into Bayesian functional regression. The subject-specific effects sampler (ii) has computational complexity $\mathcal{O}(nK)$ for all $\{\gamma_{k,i}\}$, which allows us to incorporate subject-specific functional effects, modeled nonparametrically in τ via $\{f_k\}$, with minimal additional computational cost.

The remaining sampling steps for the variance components in (1) and (3) consist of standard conjugate updates and parameter expansions, and are provided in the supplement. Note that although we treat K as fixed, the MCMC algorithm may be modified to select the number of basis functions adaptively as in Bhattacharya and Dunson (2011).

In Figure 1, we present empirical computing times for the proposed algorithm (*FOSR). For comparison with existing frequentist and Bayesian methods for FOSR, we include Reiss, Huang, and Mennes (2010), which uses a generalized least squares estimation procedure (refund:GLS), and Goldsmith and Kitago (2016), which provides a Gibbs sampler (refund:Gibbs) and a variational Bayes algorithm (refund:VB) for a Bayesian FOSR model, all implemented in the `refund` package in R (Goldsmith et al. 2016). The `refund` implementations require $p < n$. Figure 1 demonstrates that the proposed algorithm is fast, scalable, and superior to Bayesian and frequentist alternatives, and empirically validates linear time complexity in either p or n. Note that the linear time complexity is purely computational, and refers to computing time per MCMC iteration, and

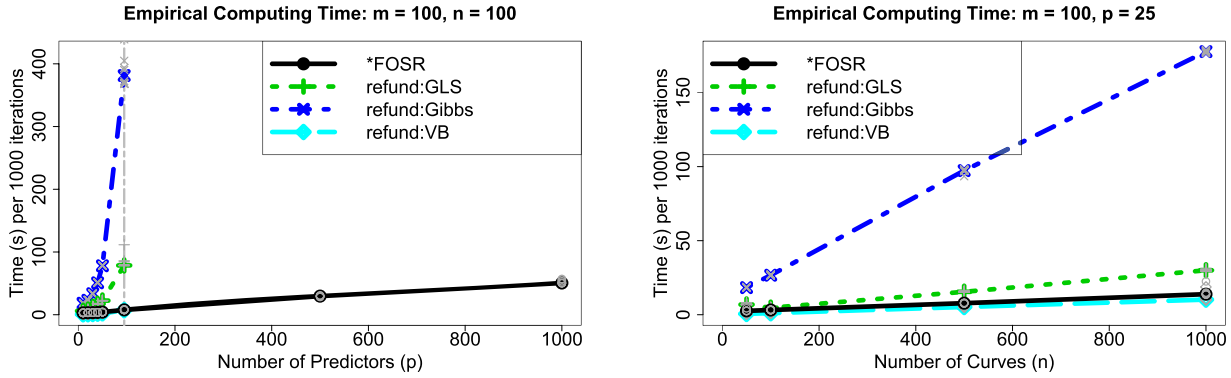

Figure 1. Empirical computing time for several FOSR algorithms with varying p (left) and varying n (right) reported in seconds per 1000 iterations for MCMC methods and total computing time for optimization methods (using R on a MacBook Pro, 2.7 GHz Intel Core i5). The computing time for each of 10 replicates is plotted (dark gray), with the largest point (color) indicating the median. The vertical gray dashed line (left) indicates $p = 95$, for which *FOSR runs in about 7 sec. The proposed algorithm is empirically faster than all methods except refund:VB, which performs similarly, yet requires $p < n$.

therefore does not imply a faster convergence rate to the stationary distribution. A group lasso alternative was also considered, but omitted from Figure 1 since it was noncompetitive.

Collectively, the identifiability constraint $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$ and the likelihood simplification of Lemma 1, together with the joint sampler for $\{\mu_k, \alpha_{j,k}, \gamma_{k,i}\}_{j,k,i}$ and the application of Bhattacharya, Chakraborty, and Mallick (2016), provide the computational infrastructure to simultaneously (i) model the basis functions $\{f_k\}$ as unknown to be learned from the data, (ii) include many predictors, even $p \gg n$, and (iii) incorporate subject-specific random effects, which is essential for functional response regression. To the best of our knowledge, there are no existing methods for Bayesian functional regression that offer these features.

4. Variable Selection in Functional Regression

Variable selection in FOSR is important for obtaining parsimonious and interpretable model summaries and reducing storage costs. Fundamentally, we are interested in identifying a subset of predictors that maintains nearly the predictive ability of the full model. The tradeoff between accuracy and sparsity may be defined precisely by a loss function; estimation and variable selection are achieved by minimizing the expected loss under the posterior predictive distribution. This approach, recently pioneered by Hahn and Carvalho (2015) for multiple linear regression and building upon Laud and Ibrahim (1995) and Gelfand and Ghosh (1998), effectively *decouples shrinkage and selection* (DSS): posterior summarization, which may incorporate sparsity constraints for variable selection, is distinct from the choice of the prior. Importantly, unlike *marginal* selection approaches, such as the median probability model under sparsity priors (Barbieri and Berger 2004), hard-thresholding under shrinkage priors (Carvalho, Polson, and Scott 2010), or GBPVs (Meyer et al. 2015), DSS selects variables *jointly*, while accounting for collinearity among predictors.

Consider predicting new observations $\tilde{Y}_i(\tau_\ell)$ for each subject $i = 1, \ldots, n$ for a predefined set of points $\{\tau_\ell\}_{\ell=1}^m$ and design points $\tilde{\boldsymbol{x}}_i = (\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,p})'$, which may differ from $\{x_{i,j}\}_{j=1}^p$. Variable selection is provided by group sparsity: removing the jth predictor is equivalent to setting the target regression function $\tilde{\delta}_j(\tau) = 0$ for *all* $\tau \in \mathcal{T}$. We write $\tilde{\delta}_j(\cdot)$ evaluated at $\{\tau_\ell\}_{\ell=1}^m$ as follows: $\tilde{\boldsymbol{\delta}}_j = (\tilde{\delta}_j(\tau_1), \ldots, \tilde{\delta}_j(\tau_m))' = \boldsymbol{F}\boldsymbol{\delta}_j$, where $\boldsymbol{F}$ is an $m \times K$ basis matrix with $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$ and $\boldsymbol{\delta}_j = (\delta_{j,1}, \ldots, \delta_{j,K})'$ are the basis coefficients. For now, suppose that $\boldsymbol{F}$ is known: important examples include the identity matrix $\boldsymbol{F} = \boldsymbol{I}$, so that $\tilde{\boldsymbol{\delta}}_j = \boldsymbol{\delta}_j$, and the (orthonormal) spline matrix $\boldsymbol{F} = \boldsymbol{B}$; the case of unknown $\boldsymbol{F}$ from Section 2.2 is considered subsequently. We propose the following loss function:

$$\mathcal{L}(\tilde{\boldsymbol{Y}}, \boldsymbol{\Delta}) = \frac{1}{nm}\sum_{i=1}^{n} \left|\left|\tilde{\boldsymbol{Y}}_i - \boldsymbol{F}\boldsymbol{\delta}_0 - \boldsymbol{F}\boldsymbol{\Delta}\tilde{\boldsymbol{x}}_i\right|\right|_2^2 + \lambda\left|\left|\boldsymbol{\Delta}\right|\right|_0, \quad (8)$$

where $||\cdot||_2$ is the Euclidean norm, $\tilde{\boldsymbol{Y}}_i = (\tilde{Y}_i(\tau_1), \ldots, \tilde{Y}_i(\tau_m))'$, $\boldsymbol{\delta}_0 = (\delta_{0,1}, \ldots, \delta_{0,K})$ is the vector of basis coefficients for the functional intercept, $\boldsymbol{\Delta}$ is the $K \times p$ matrix with entries $(\boldsymbol{\Delta})_{k,j} = \delta_{j,k}$ the kth basis coefficient for predictor j, and $||\boldsymbol{\Delta}||_0 = \sum_{j=1}^p \mathbb{I}\{\delta_{j,k} \neq 0 \text{ for some } k = 1, \ldots, K\}$. The predictive ability of $\{\tilde{\delta}_j(\cdot)\}$ is measured by the squared prediction error over all points $\tau_1, \ldots, \tau_m$ and subjects $i = 1, \ldots, n$, while the second term in (8) encourages column-wise sparsity of the basis coefficients $\boldsymbol{\Delta}$ with the tradeoff determined by $\lambda > 0$. Since $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$, sparsity of the basis coefficients $\boldsymbol{\delta}_j = \boldsymbol{0}_K$ is equivalent to sparsity of the regression functions $\tilde{\boldsymbol{\delta}}_j = \boldsymbol{0}_M$ evaluated at $\{\tau_\ell\}_{\ell=1}^m$. When $K \ll m$, the sparsity penalty in (8) operates in a lower-dimensional space, which produces more stable and efficient computations.

To select predictors, we minimize the expectation of the loss function (8) under the posterior predictive distribution, $[\tilde{\boldsymbol{Y}}|\boldsymbol{Y}]$, which marginalizes over model parameters, say $\boldsymbol{\theta}$. Although $[\tilde{\boldsymbol{Y}}|\boldsymbol{Y}]$ is unavailable in closed form, the conditional predictive distribution $[\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}]$ and the posterior distribution of the model parameters $[\boldsymbol{\theta}|\boldsymbol{Y}]$ are sufficient for obtaining a useful representation of the posterior predictive expected loss. The *conditional* predictive distribution for $\tilde{\boldsymbol{Y}}_i$, after marginalizing over the subject-specific effects $\tilde{\gamma}_i(\cdot)$, is $[\tilde{\boldsymbol{Y}}_i|\boldsymbol{\theta}] \overset{\text{indep}}{\sim} N(\boldsymbol{F}\boldsymbol{\mu} + \boldsymbol{F}\boldsymbol{A}\tilde{\boldsymbol{x}}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\theta} = \left(\{\mu_k\}, \{\alpha_{j,k}\}, \{\sigma_{\gamma_{k,i}}\}, \{f_k\}, \sigma_\epsilon\right)$ are the relevant model parameters, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$, $\boldsymbol{A}$ is the $K \times p$ matrix with entries $(\boldsymbol{A})_{k,j} = \alpha_{j,k}$, and $\boldsymbol{\Sigma}_i = \sigma_\epsilon^2\boldsymbol{I}_m + \boldsymbol{F}\boldsymbol{\Sigma}_{\gamma_i}\boldsymbol{F}'$ with $\boldsymbol{\Sigma}_{\gamma_i} = \text{diag}\left(\{\sigma_{\gamma_{k,i}}^2\}_{k=1}^K\right)$. The expected loss may be simplified:

Theorem 1. For a *known* basis matrix $\boldsymbol{F}$ with $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$, the posterior predictive expectation of the loss in (8) is

$$\mathbb{E}_{[\tilde{\boldsymbol{Y}}|\boldsymbol{Y}]}\mathcal{L}(\tilde{\boldsymbol{Y}}, \boldsymbol{\Delta}) = C(\boldsymbol{Y}) + \frac{1}{nm}\sum_{i=1}^{n} \left|\left| \left(\bar{\boldsymbol{\mu}} + \bar{\boldsymbol{A}}\tilde{\boldsymbol{x}}_i\right) - \left(\boldsymbol{\delta}_0 + \boldsymbol{\Delta}\tilde{\boldsymbol{x}}_i\right)\right|\right|_2^2 + \lambda||\boldsymbol{\Delta}||_0, \quad (9)$$

where $C(\boldsymbol{Y})$ is a constant that does not depend on $\boldsymbol{\delta}_0$ and $\boldsymbol{\Delta}$, $\bar{\boldsymbol{\mu}} = \mathbb{E}[\boldsymbol{\mu}|\boldsymbol{Y}]$, and $\bar{\boldsymbol{A}} = \mathbb{E}[\boldsymbol{A}|\boldsymbol{Y}]$.

The posterior predictive expected loss (9) is a penalized regression with response $\bar{\boldsymbol{\mu}} + \bar{\boldsymbol{A}}\tilde{\boldsymbol{x}}_i$ constructed from posterior expectations and a sparsity penalty on $\boldsymbol{\Delta}$. Notably, (9) is valid under *any* prior for $\boldsymbol{\mu}$ and $\boldsymbol{A}$ as long as the posterior means $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{A}}$ exist, and therefore is broadly applicable for Bayesian FOSR models. For computational tractability, we replace the penalty $||\boldsymbol{\Delta}||_0$ with a convex relaxation: $||\boldsymbol{\Delta}||_1 = \sum_{j=1}^p \sqrt{K}||\boldsymbol{\Delta}_j||_2$ for $\boldsymbol{\Delta}_j = (\delta_{j,1}, \ldots, \delta_{j,K})'$, which is the group lasso penalty (Yuan and Lin 2006). To avoid overshrinkage, we use an adaptive group lasso penalty:

$$||\boldsymbol{\Delta}||_{1^*} = \sum_{j=1}^{p} w_j||\boldsymbol{\Delta}_j||_2, \quad w_j = 1/||\bar{\boldsymbol{A}}_j||_2 \quad (10)$$

for $\bar{\boldsymbol{A}}_j$ the posterior mean of $(\alpha_{j,1}, \ldots, \alpha_{j,K})'$. The use of the adaptive group lasso is analogous to the use of the adaptive lasso in Hahn and Carvalho (2015). The solution path for minimization of (9) with modified penalty (10) may be computed using existing software, such as the `gglasso` package in `R` (Yang and Zou 2017).

When $\boldsymbol{F}$ is unknown, as in Section 2.2, the loss function (8) depends on $\tilde{\boldsymbol{\delta}}_j = \boldsymbol{F}\boldsymbol{\delta}_j$, and therefore depends on unknown model parameters $\{f_k\}$. To account for the uncertainty in $\{f_k\}$,

we construct an *expected* loss function directly by integrating over $\boldsymbol{F}$:

$$\mathcal{EL}(\boldsymbol{\Delta}) = \mathbb{E}_{[\boldsymbol{F}|\boldsymbol{Y}]}\left\{\mathbb{E}_{[\tilde{\boldsymbol{Y}}|\boldsymbol{F},\boldsymbol{Y}]}\left[\frac{1}{nm}\sum_{i=1}^{n}||\tilde{\boldsymbol{Y}}_i - \boldsymbol{F}\boldsymbol{\delta}_0 - \boldsymbol{F}\boldsymbol{\Delta}\tilde{\boldsymbol{x}}_i||_2^2 + \lambda||\boldsymbol{\Delta}||_{0.}\right]\right\} \tag{11}$$

The additional complexity in (11) is due to the uncertainty of $\{f_k\}$, which requires successive averaging over the distributions $[\tilde{\boldsymbol{Y}}|\boldsymbol{F},\boldsymbol{Y}]$ and $[\boldsymbol{F}|\boldsymbol{Y}]$ to obtain the posterior predictive expectation under $[\tilde{\boldsymbol{Y}}|\boldsymbol{Y}]$. Note that in general, $[\tilde{\boldsymbol{Y}}|\boldsymbol{F},\boldsymbol{Y}]$ is not equal to $[\tilde{\boldsymbol{Y}}|\boldsymbol{F}]$: while it is often assumed that $\tilde{\boldsymbol{Y}}$ is conditionally independent of $\boldsymbol{Y}$ given *all* model parameters, this distribution is only conditional on $\boldsymbol{F}$. We again obtain a convenient simplification:

Theorem 2. For an *unknown* basis matrix $\boldsymbol{F}$ with $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$, the expected loss $\mathcal{EL}(\boldsymbol{\Delta})$ in (11) is equivalent to (9) up to constants that do not depend on $\boldsymbol{\delta}_0$ and $\boldsymbol{\Delta}$.

Despite the additional complexity introduced by modeling $\boldsymbol{F}$ as unknown in Section 2.2, the orthonormality constraint $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_K$ provides a mechanism by which the expected loss (11) collapses into the penalized regression (9). As a result, optimization may proceed as before. Given a solution $\hat{\boldsymbol{\Delta}}$ to (9) using the convex relaxation (10), the proposed DSS estimator of the regression functions is $\mathbb{E}[\boldsymbol{F}\hat{\boldsymbol{\Delta}}|\boldsymbol{Y}] = \bar{\boldsymbol{F}}\hat{\boldsymbol{\Delta}}$ where $\bar{\boldsymbol{F}} = \mathbb{E}[\boldsymbol{F}|\boldsymbol{Y}]$.

We select the tuning parameter $\lambda > 0$ in (9) by identifying a sparsified model for which the predictive ability is within a range of uncertainty of that of the full model. Consider the following notion of proportion of variability explained, analogous to R^2 in linear regression:

$$\rho^2 = \frac{||\boldsymbol{A}\tilde{\boldsymbol{X}}'||_F^2}{\mathbb{E}_{[\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}]}||\tilde{\boldsymbol{Y}}||_F^2} = \frac{||\boldsymbol{A}\tilde{\boldsymbol{X}}'||_F^2}{||\boldsymbol{A}\tilde{\boldsymbol{X}}'||_F^2 + \sum_{i=1}^{n}\text{tr}(\boldsymbol{\Sigma}_i)}, \tag{12}$$

where $\tilde{\boldsymbol{X}}$ is the $n \times p$ matrix of predictors $[\tilde{x}_{i,j}]_{i,j}$, $\tilde{\boldsymbol{Y}}$ is the $n \times m$ matrix of predictive values $[\tilde{Y}_i(\tau_\ell)]_{i,\ell}$, $||\cdot||_F$ denotes the Frobenius norm, and $\text{tr}(\cdot)$ is the trace operator. Since $\sum_{i=1}^{n}\text{tr}(\boldsymbol{\Sigma}_i) = nm\sigma_\epsilon^2 + \sum_{i=1}^{n}\sum_{k=1}^{K}\sigma_{\gamma_{k,i}}^2$, the total variance in the denominator of (12) includes the regression model term $\boldsymbol{A}\tilde{\boldsymbol{X}}'$, the observation error variance σ_ϵ^2, and the subject-specific variances $\sigma_{\gamma_{k,i}}^2$. We exclude the intercept from (12), but this is not strictly necessary.

For a given λ, let $\hat{\boldsymbol{\Delta}}_\lambda$ denote the solution to (9) with the adaptive group lasso penalty. Model discrepancy, due to re-estimation under sparsity, contributes to total variance:

$$\rho_\lambda^2 = \frac{||\boldsymbol{A}\tilde{\boldsymbol{X}}'||_F^2}{\mathbb{E}_{[\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}]}||\tilde{\boldsymbol{Y}}||_F^2 + ||\boldsymbol{A}\tilde{\boldsymbol{X}}' - \hat{\boldsymbol{\Delta}}_\lambda\tilde{\boldsymbol{X}}'||_F^2}, \tag{13}$$

which we may simplify as above using $\mathbb{E}_{[\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}]}||\tilde{\boldsymbol{Y}}||_F^2 = ||\boldsymbol{A}\tilde{\boldsymbol{X}}'||_F^2 + nm\sigma_\epsilon^2 + \sum_{i=1}^{n}\sum_{k=1}^{K}\sigma_{\gamma_{k,i}}^2$.

We compare the posterior distributions of ρ^2 and ρ_λ^2 to assess disparities in predictive ability among sparsified models. We construct a posterior selection summary plot as follows: (i) using the `gglasso` package (Yang and Zou 2017), minimize (9) with the penalty (10) for $\hat{\boldsymbol{\Delta}}_\lambda$ on a grid of λ values; (ii) for each $\hat{\boldsymbol{\Delta}}_\lambda$, compute the posterior distribution of ρ_λ^2 by substituting the posterior draws of $\boldsymbol{A}$, σ_ϵ, and $\sigma_{\gamma_{k,i}}$ into (13); and (iii) plot the expected value and 90% credible intervals of ρ_λ^2 against model size $||\hat{\boldsymbol{\Delta}}_\lambda||_0$. The resulting plot shows how predictive ability declines with sparsity, but importantly includes the accompanying uncertainty (see Figure 4). As a general guideline for model selection, we choose the smallest model for which the 90% posterior credible interval for ρ_λ^2 contains $\mathbb{E}[\rho^2|\boldsymbol{Y}]$.

5. Simulations

5.1. Simulation Design

A simulation study was conducted to (i) compare the estimation accuracy of the proposed method against existing alternatives, (ii) evaluate uncertainty quantification among Bayesian FOSR models, and (iii) study the variable selection properties of the proposed DSS procedure. We are primarily interested in how these properties vary in the number of predictors, so we consider $p \in \{20, 50, 500\}$ with fixed $p_1 = 10$ nonnull predictors. All simulations use $n = 100$ curves with $m = 30$ equally spaced points in $[0, 1]$.

For each subject i, we simulate correlated predictors $\{x_{i,j}\}_{j=1}^{p}$ from a normal distribution with mean zero and covariance $\text{cov}(x_{i,j}, x_{i,j'}) = 0.75^{|j-j'|}$, with $p_1 = 10$ nonnull predictors evenly spaced from $1, \ldots, p$. Functional observations are simulated based on Models (1) and (2). For the true basis functions, we set $f_1^*(\tau) = 1/\sqrt{m}$ and for $k = 2, \ldots, K^* = 4$, we let f_k^* be an orthogonal polynomial of degree k. For each nonnull predictor j, we uniformly sample K_j^* factors to be nonzero, where K_j^* follows a Poisson(1) distribution truncated to $[1, K^*]$, and draw the nonzero factor coefficients $\alpha_{j,k}^* \overset{\text{indep}}{\sim} N(0, 1/k^2)$. Each nonnull predictor j may be associated with the functional response via a subset of $\{f_k^*\}_{k=1}^{K^*}$. We simulate the true factors $\beta_{k,i}^* = \mu_k^* + \sum_{j=1}^{p}x_{j,i}\alpha_{j,k}^* + \gamma_{k,i}^*$, where $\mu_k^* = 1/k$ and $\gamma_{k,i}^* \overset{\text{indep}}{\sim} N(0, 1/k^2)$, which incorporates subject-specific random effects. Based on the true curves $Y_i^*(\tau) = \sum_{k=1}^{K^*}f_k^*(\tau)\beta_{k,i}^*$, the functional observations are $Y_i(\tau) = Y_i^*(\tau) + \sigma^*\epsilon_i^*(\tau)$, where $\epsilon_i^*(\tau) \overset{\text{iid}}{\sim} N(0, 1)$. The observation error standard deviation σ^* is determined by the *root-signal-to-noise ratio* (RSNR): $\sigma^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(Y_i^*(\tau_j) - \bar{Y}^*)^2}{nm-1}} \Big/ \text{RSNR}$ where $\bar{Y}^*$ is the sample mean of $\{Y_i^*(\tau_j)\}_{j,i}$. We select RNSR = 5 for moderately noisy functional data.

5.2. Methods for Comparison

We implement Models (1)–(3) with $K = 6 > K^* = 4$ (*FOSR). Using this posterior distribution, we compute the sparse DSS estimates from Section 4 (*FOSR-DSS). For comparison with *FOSR-DSS, we also select variables marginally using GBPVs, which retain a variable j if the simultaneous credible bands for $\{\tilde{\alpha}_j(\tau_\ell)\}_{\ell=1}^{m}$ exclude zero for some τ_ℓ.

We include two variations of (1)–(3) based on alternative models for $\{f_k\}$: Basis-FPCA, which estimates $\{f_k\}$ as FPCs using Xiao, Li, and Ruppert (2013) with the number of FPCs selected

to explain 99% of the variability in $\{Y_i(\tau_\ell) - \bar{Y}(\tau_\ell)\}_{\ell,i}$, and Basis-Spline, which uses an (orthonormalized) LR-TPS basis for $\{f_k\}$. Both Basis-FPCA and Basis-Spline use normal-inverse-gamma priors for (3). Importantly, Basis-FPCA and Basis-Spline provide baselines for assessing the potential gains in point estimation and uncertainty quantification associated with the proposed shrinkage priors and the model for $\{f_k\}$. Basis-FPCA and Basis-Spline are implemented using the proposed Gibbs sampler by omitting the basis function sampling step, and rely on the computational results of Section 3 for scalability.

Finally, we include three FOSR methods from the `refund` package in R (see Figure 1): estimation with a group lasso for variable selection (refund:Lasso), estimation using generalized least squares (refund:GLS; Reiss, Huang, and Mennes 2010), and a Bayesian model using FPCs to estimate the residual covariance (refund:Gibbs; Goldsmith and Kitago 2016). For $p = 500 > n = 100$, the `refund` methods are not computationally feasible.

We compare methods using three metrics. For point estimation, we use the root mean square error of the regression coefficient functions, $\text{RMSE} = \sqrt{\frac{1}{pm}\sum_{j=1}^{p}\sum_{l=1}^{m}\left[\tilde{\alpha}_j(\tau_\ell) - \tilde{\alpha}_j^*(\tau_\ell)\right]^2}$, where $\tilde{\alpha}_j(\tau_\ell)$ is the estimated regression coefficient for predictor j and observation point τ_ℓ and $\tilde{\alpha}_j^*(\tau_\ell) = \sum_{k=1}^{K^*} f_k^*(\tau_\ell)\alpha_{j,k}^*$ is the true regression coefficient. The Bayesian methods use the posterior expectation of $\tilde{\alpha}_j(\tau_\ell)$ as the estimator. For uncertainty quantification among the Bayesian methods, we compute the mean credible interval width for all regression coefficient functions, $\text{MCIW} = \frac{1}{pm}\sum_{j=1}^{p}\sum_{\ell=1}^{m}\left[\tilde{\alpha}_j^{(97.5)}(\tau_\ell) - \tilde{\alpha}_j^{(2.5)}(\tau_\ell)\right]$, where $\tilde{\alpha}_j^{(q)}(\tau_\ell)$ is the q% quantile of the posterior distribution for $\tilde{\alpha}_j(\tau_\ell)$, along with the empirical coverage probability, $\frac{1}{pm}\sum_{j=1}^{p}\sum_{\ell=1}^{m}\mathbb{I}\{\tilde{\alpha}_j^{(2.5)}(\tau_\ell) \le \tilde{\alpha}_j^*(\tau_\ell) \le \tilde{\alpha}_j^{(97.5)}(\tau_\ell)\}$. The goal is to achieve the smallest MCIW with the correct nominal coverage (0.95). For comparing variable selection techniques, we compute receiver-operating characteristic (ROC) curves. ROC curves plot the true positive rate, or sensitivity, against the false positive rate, or 1 − specificity, as the decision threshold varies, where a positive predictor corresponds to a nonzero regression function $\tilde{\alpha}_j(\tau) \neq 0$ for some τ. ROC curves further toward the upper left corner indicate superiority of the method.

5.3. Simulation Results

Figure 2 displays RMSEs and MCIWs for 100 simulated datasets. The RMSEs show that *FOSR provides the best point estimation for all p, with the most substantial improvements for moderate to large $p \ge 50$. The sparse estimates from *FOSR-DSS perform at least as well as refund:Lasso for $p = 20$, and are outperformed only by *FOSR for $p \ge 50$. The case of $p = 500$ demonstrates the utility of the shrinkage priors of Section 2.3, as Basis-FPCA and Basis-Spline are clearly dominated by *FOSR and *FOSR-DSS. Estimation of $\{f_k\}$ is also important: there is a sizable gap for all p between Basis-Spline, which does *not* estimate $\{f_k\}$, and *FOSR, *FOSR-DSS, and Basis-FPCA, which do estimate $\{f_k\}$.

The MCIWs in Figure 2 demonstrate that the proposed *FOSR provides significantly narrower credible intervals than competing Bayesian methods, particularly for $p \ge 50$, while maintaining approximately the correct nominal coverage. Note that although Basis-Spline produces narrow credible intervals, it suffers from severe undercoverage. The large improvements of *FOSR relative to Basis-FPCA suggest that the proposed shrinkage priors and model for the basis functions—which, unlike Basis-FPCA, accounts for the uncertainty of the unknown $\{f_k\}$—are both important for more precise uncertainty quantification.

In Figure 3, we show ROC curves for the competing variable selection techniques: *FOSR-DSS, FOSR-GBPV, and refund:Lasso. For all p, the proposed *FOSR-DSS is at least as good as the other methods, with greater improvements relative to FOSR-GBPV as p increases. Notably, refund:Lasso is inferior to both Bayesian methods for variable selection.

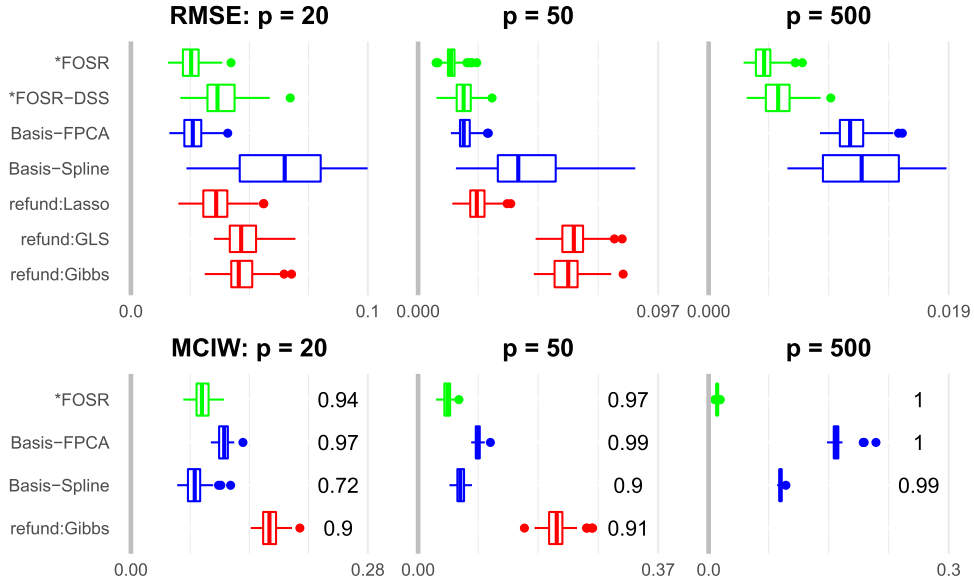

Figure 2. Root mean squared errors (top row) and mean 95% credible interval widths with empirical coverage probabilities (bottom row) for $\tilde{\alpha}_j(\tau)$. The proposed methods (*FOSR and *FOSR-DSS) are in green, the $\{f_k\}$-modified methods (Basis-FPCA and Basis-Spline) are in blue, and existing methods (refund:Lasso, refund:GLS, and refund:Gibbs) are in red. Existing methods were not feasible for $p = 500 > n = 100$. For display purposes, outliers of Basis-Spline are not shown.

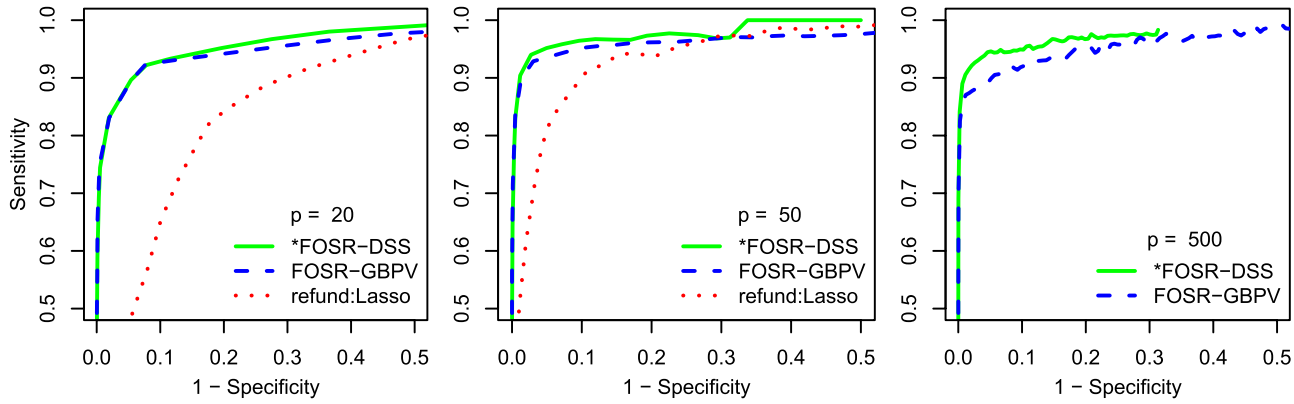

Figure 3. ROC curves to compare variable selection techniques. Each point along the ROC curve is the average sensitivity and 1 − specificity of a given model size. For $p = 500$, *FOSR-DSS selects only a few false positives, so the corresponding ROC curve does not extend to lower (i.e., worse) specificity.

6. Time-of-Day Physical Activity Levels for Elderly Adults

We analyze time-of-day physical activity levels obtained from the National Sleep Research Resource (Dean et al. 2016; Zhang et al. 2018) in conjunction with the Multi-Ethnic Study of Atherosclerosis. The dataset consists of actigraphy measurements and an accompanying sleep questionnaire (see the supplementary materials) for elderly men and women. Undoubtedly, physical activity is an important component in many health studies. The simplest measurement of physical activity is a questionnaire: subjects may report their activities and behaviors during the day. However, it is unclear how accurate these responses are, and whether they truly correspond to real physical activity patterns. In contrast, wearable devices provide high-resolution and direct measurements of physical activity. However, wearable devices introduce additional costs and complexity into a study, and may not be feasible in all cases. The aims of our investigation are (i) to assess the predictive ability of the questionnaire for time-of-day physical activity levels and (ii) to identify a minimal set of questionnaire items that maintain the predictive ability of the full questionnaire.

To focus primarily on waking hours, we considered activity levels from 6 a.m. to 10 p.m. on Wednesdays and Saturdays. The raw activity counts were aggregated into 20-min time increments and modeled as a function of time-of-day with $m = 48$ observation points. Days with more than 10% missingness typically had long periods of no activity recorded and were removed from the analysis. Six percent of the remaining days had missing observation points, which were imputed automatically within the Gibbs sampler. Covariate information included age, gender, race/ethnicity, a weekday/weekend indicator, and questionnaire responses. Due to substantial missingness among questionnaire responses, we modeled questionnaire items as categorical variables, grouped into high, low, and missing responses where appropriate (see the supplementary materials for details). In total, 2059 people ages 54–95 were considered over a cumulative $n = 3568$ days, with 34 items from the questionnaire and $p = 74$. In the supplementary materials, we repeat our analysis for all available day-of-the-week observations, and include both day-within-subject and subject-specific random effects. Despite the larger sample size and more complex random effects structure—both of which are straightforward to accommodate in the proposed modeling framework—the conclusions presented below are unchanged.

Posterior samples from Models (1)–(3) were obtained using the MCMC algorithm in the supplementary materials. We ran the MCMC algorithm for 20,000 iterations and discarded the initial 2000 iterations as a burn-in, resulting in 18,000 MCMC samples for posterior inference; results were not sensitive to these choices, and longer simulations produced identical conclusions. Results are reported for $K = 6$; larger values of K produced similar results. Traceplots demonstrate good mixing and suggest convergence (see the supplementary materials).

We applied the DSS procedure from Section 4 to identify demographic variables and questionnaire items that predict intraday physical activity. We perform selection on the *category levels* rather than the questionnaire items: selected levels imply inclusion of the corresponding questionnaire item (along with the baseline category), but do not require other levels be included. The selection summary plot in Figure 4 displays the tradeoff between predictive accuracy and sparsity: the proportion of variability explained increases quickly between models of size six and ten, but does not notably increase for larger models. The most reasonable model size is between eight and ten, while larger models do not offer additional predictive ability. Guided by Figure 4, we select 10 predictors: `Weekend`, `Gender`, `age`, and the seven categories of the four questionnaire items in Table 1. Impressively, these four questionnaire items (along with

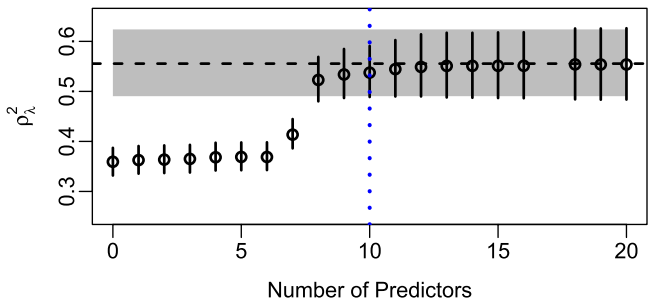

Figure 4. The selection summary plot for the time-of-day physical activity data. Shown is the proportion of variability explained (with 95% credible intervals) for increasing model sizes. The horizontal line denotes the proportion of variability explained by the full model with the gray band denoting the 95% credible interval. As model size increases, the explanatory power of the model increases. The vertical dotted line denotes the selected model with ten predictors.

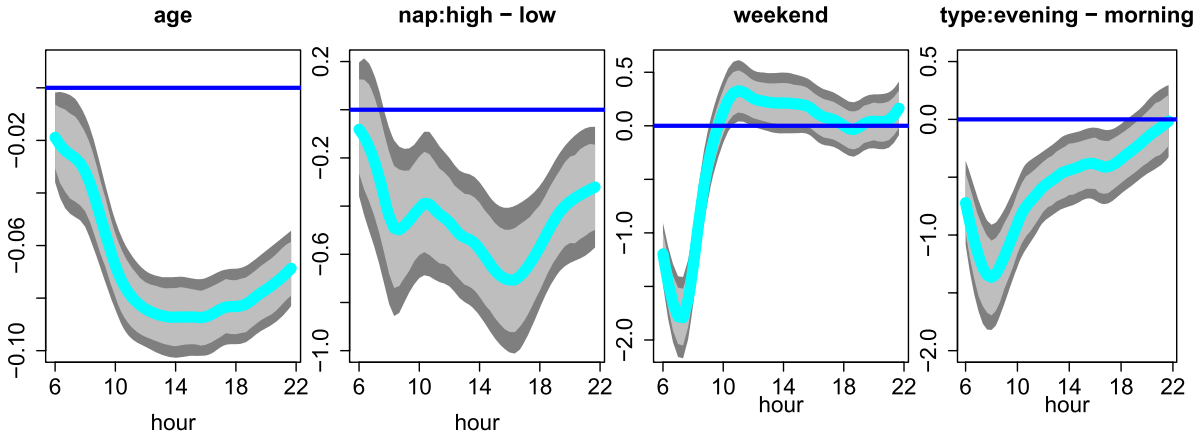

Figure 5. Estimated regression coefficient functions for DSS-selected variables, 95% pointwise credible intervals (light gray), and 95% simultaneous credible bands (dark gray). A horizontal blue line denotes zero change in activity.

Table 1. The questionnaire item name, categories, and question content of the items selected.

Item name	Categories	Question content
`bedtmwkday`	Missing/**5**–**7**/**8**–**9**/10–11/**later**	Bedtime weekday
`nap`	Missing/**low**/high	Usual week: number of naps
`types`	Missing/**evening**/morning/neither	Type of person: morning or evening
`wkdaysleepdur`	Missing/<7/**7–9**/>**9**	Sleep duration weekday

NOTE: The first category in each row is the contrasting variable and the selected categories are given in bold.

`Weekend`, `Gender`, and `age`) retain nearly the predictive ability of the full questionnaire, and explain almost 60% of the variability in time-of-day physical activity levels. Figure 4 also provides a measure of variable importance: the nearly 2-fold increase in predictive ability from $p = 6$ to $p = 9$ is achieved by including the individual's age and questionnaire responses about the weekly number of naps, maintaining an early bedtime (`bedtmwkday` 5–7), and obtaining 7–9 hr of sleep each weeknight (`wkdaysleepdur` 7–9).

The estimated coefficient functions and 95% simultaneous credible bands for several DSS-selected variables are displayed in Figure 5. Most notably, these effects are time-varying, which confirms the importance of using a *functional* regression model. The simultaneous credible bands for age are below zero for all times of day, indicating that daily activity declines with age. This effect is largest from 10 a.m. to 4 p.m., which we note is the most active time of day in the study. Individuals who reported frequent napping also tended to be less active throughout the day, particularly in the mid-afternoon. Physical activity levels were lower on Saturday mornings compared to Wednesday mornings, but were similar throughout the rest of the day. Lastly, self-described "evening people" were less active in the morning than "morning people," but more surprisingly, were also less active overall.

We confirm the stability of the proposed variable selection procedure in the supplementary materials. Specifically, we repeatedly apply the approach from Section 4 to random subsamples of the data of size $n = 100$ and $n = 500$ days, and identify the variables that were selected consistently across these random subsets of the data. Naturally, the larger sample size produces results more consistent with the full dataset of $n = 3568$ days.

7. Discussion

We developed a fully Bayesian framework for function-on-scalars regression with many predictors. Nonparametric and unknown basis functions were proposed for greater modeling flexibility and proper uncertainty quantification via the posterior distribution. Carefully designed shrinkage priors were employed to minimize the impact of unimportant predictor variables, which is particularly important for moderate to large p. We introduced a novel variable selection technique for functional regression, which identifies a sparse subset of predictors that minimizes loss in predictive performance relative to the full model. A simulation study illustrated the improvements in point estimation, uncertainty quantification, and variable selection offered by the proposed methodology. Full posterior inference was provided by an efficient Gibbs sampler with computational complexity scaling in either n or p. The methodology was applied to an actigraphy and sleep questionnaire dataset from the National Sleep Research Resource. Our analysis identifies a small subset of questionnaire items that are highly predictive of intraday physical activity. While it is of interest to know *which* questions provide the predictive power, it is also interesting to note that *there exists* a small number of questions that can nearly recover the predictive power of the full questionnaire.

The proposed methodology offers several promising extensions. Modifications for binomial and count data are available by coupling the computational approach of Section 3 with well-known Gaussian parameter expansions (Kowal 2019). Alternative models for $\{f_k\}$ may be introduced for other applications, for example wavelets for nonsmooth functional data or Fourier basis functions for periodic functional data. Lastly, the proposed DSS approach offers a general framework for Bayesian variable selection and posterior summarization in other functional, spatial, and time series regression problems.

More broadly, Bayesian functional data models would benefit from further development of MCMC diagnostics customized

for this setting. In particular, adequate measures of *functional effective sample size*, such as for the regression coefficient functions $\{\tilde{\alpha}_j(\tau_\ell)\}_{\ell,j}$, are needed for proper assessment of MCMC efficiency and derivation of stopping rules. As Vats, Flegal, and Jones (2019) demonstrate in the multivariate setting, a functional effective sample size should account for within-curve dependence and provide a global measure over the domain $\mathcal{T}$. Such diagnostic tools for MCMC efficiency are important accompaniments to scalable MCMC algorithms, such as the proposed approach for function-on-scalars regression, and remain a promising area for future research.

Supplementary Materials

Additional results: A document containing MCMC diagnostics, proofs of theoretical results, and additional details and analysis related to the application. (PDF)

R code: Code for model implementation, the simulations of Section 5, and preprocessing the physical activity data. (Zipped file)

Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship (Bourgeois) under grant DGE-1450681 and partially supported by NSF award #1547433 RTG: Cross-training in Statistics and Computer Science.

References

Barber, R. F., Reimherr, M., and Schill, T. (2017), "The Function-on-Scalar LASSO With Applications to Longitudinal GWAS," *Electronic Journal of Statistics*, 11, 1351–1389. [1]

Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897. [5]

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016), "Fast Sampling With Gaussian Scale Mixture Priors in High-Dimensional Regression," *Biometrika*, 103, 985–991. [4,5]

Bhattacharya, A., and Dunson, D. B. (2011), "Sparse Bayesian Infinite Factor Models," *Biometrika*, 98, 291–306. [3,4]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [3,5]

Chen, Y., Goldsmith, J., and Ogden, R. T. (2016), "Variable Selection in Function-on-Scalar Regression," *Stat*, 5, 88–101. [1,2]

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005), "Bayesian Analysis for Penalized Spline Regression Using WinBUGS," *Journal of Statistical Software*, 14, 1–24. [3]

Dean, D. A., Goldberger, A. L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., Sahoo, S. S., Jayapandian, C. P., Cui, L., Morrical, M. G., and Surovec, S. (2016), "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," *Sleep*, 39, 1151–1164. [8]

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009), "Multilevel Functional Principal Component Analysis," *The Annals of Applied Statistics*, 3, 458. [2]

Fan, Z., and Reimherr, M. (2017), "High-Dimensional Adaptive Function-on-Scalar Regression," *Econometrics and Statistics*, 1, 167–183. [1]

Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11. [5]

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013), "Corrected Confidence Bands for Functional Data Using Principal Components," *Biometrics*, 69, 41–51. [3]

Goldsmith, J., and Kitago, T. (2016), "Assessing Systematic Effects of Stroke on Motor Control by Using Hierarchical Function-on-Scalar Regression," *Journal of the Royal Statistical Society*, Series C, 65, 215–236. [3,4,7]

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2016), "`refund`: Regression With Functional Data," R Package Version 0.1-16. [4]

Goldsmith, J., Zipunnikov, V., and Schrack, J. (2015), "Generalized Multilevel Function-on-Scalar Regression and Principal Component Analysis," *Biometrics*, 71, 344–353. [3]

Guo, W. (2002), "Functional Mixed Effects Models," *Biometrics*, 58, 121–128. [3]

Hahn, P. R., and Carvalho, C. M. (2015), "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective," *Journal of the American Statistical Association*, 110, 435–448. [2,5]

Kowal, D. R. (2018), "Dynamic Function-on-Scalars Regression," arXiv no. 1806.01460. [1,3]

——— (2019), "Integer-Valued Functional Data Analysis for Measles Forecasting," arXiv no. 1902.07788. [9]

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2017), "A Bayesian Multivariate Functional Dynamic Linear Model," *Journal of the American Statistical Association*, 112, 733–744. [3]

——— (2019), "Dynamic Shrinkage Processes," *Journal of the Royal Statistical Society*, Series B, 81, 781–804. doi:10.1111/rssb.12325 [3]

Laud, P. W., and Ibrahim, J. G. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society*, Series B, 57, 247–262. [5]

Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P., and Morris, J. S. (2015), "Bayesian Function-on-Function Regression for Multilevel Functional Data," *Biometrics*, 71, 563–574. [2,5]

Montagna, S., Tokdar, S. T., Neelon, B., and Dunson, D. B. (2012), "Bayesian Latent Factor Regression for Functional and Longitudinal Data," *Biometrics*, 68, 1064–1073. [1,3]

Morris, J. S. (2015), "Functional Regression," *Annual Review of Statistics and Its Application*, 2, 321–359. [1]

Morris, J. S., and Carroll, R. J. (2006), "Wavelet-Based Functional Mixed Models," *Journal of the Royal Statistical Society*, Series B, 68, 179–199. [3]

Piironen, J., and Vehtari, A. (2016), "On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior," arXiv no. 1610.05559. [3]

Ramsay, J., and Silverman, B. (2005), *Functional Data Analysis*, New York: Springer. [1,3]

Reiss, P. T., Huang, L., and Mennes, M. (2010), "Fast Function-on-Scalar Regression With Penalized Basis Expansions," *The International Journal of Biostatistics*, 6, 1–30. [2,4,7]

Rue, H. (2001), "Fast Sampling of Gaussian Markov Random Fields," *Journal of the Royal Statistical Society*, Series B, 63, 325–338. [4]

Suarez, A. J., and Ghosal, S. (2017), "Bayesian Estimation of Principal Components for Functional Data," *Bayesian Analysis*, 12, 311–333. [3]

Vats, D., Flegal, J. M., and Jones, G. L. (2019), "Multivariate Output Analysis for Markov Chain Monte Carlo," *Biometrika*, 106, 321–337. [10]

Wang, L., Chen, G., and Li, H. (2007), "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data," *Bioinformatics*, 23, 1486–1494. [1]

Xiao, L., Li, Y., and Ruppert, D. (2013), "Fast Bivariate *P*-Splines: The Sandwich Smoother," *Journal of the Royal Statistical Society*, Series B, 75, 577–599. [6]

Yang, Y., and Zou, H. (2017), "`gglasso`: Group Lasso Penalized Learning Using a Unified BMD Algorithm," R Package Version 1.4. [5,6]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [5]

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. (2018), "The National Sleep Research Resource: Towards a Sleep Data Commons," *Journal of the American Medical Informatics Association*, 25, 1351–1358. [8]

Zhu, H., Brown, P. J., and Morris, J. S. (2011), "Robust, Adaptive Functional Regression in Functional Mixed Model Framework," *Journal of the American Statistical Association*, 106, 1167–1179. [3]