Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Hierarchical Adaptive Regression Kernels for Regression With Functional Predictors

Dawn B. WOODARD, Ciprian CRAINICEANU, and David RUPPERT

We propose a new method for regression using a parsimonious and scientifically interpretable representation of functional predictors. Our approach is designed for data that exhibit features such as spikes, dips, and plateaus whose frequency, location, size, and shape varies stochastically across subjects. We propose Bayesian inference of the joint functional and exposure models, and give a method for efficient computation. We contrast our approach with existing state-of-the-art methods for regression with functional predictors, and show that our method is more effective and efficient for data that include features occurring at varying locations. We apply our methodology to a large and complex dataset from the Sleep Heart Health Study, to quantify the association between sleep characteristics and health outcomes. Software and technical appendices are provided in the online supplementary materials.

Key Words: Electroencephalogram; Functional data analysis; Functional linear model; Kernel mixture; Lévy adaptive regression kernels; Nonparametric Bayes.

1. INTRODUCTION

Due to technological advancements, an increasing number of studies involve functional data such as images or time series. The Sleep Heart Health Study (SHHS) (Quan et al. 1997; Di et al. 2009) relates sleep patterns, as measured using electroencephalogram (EEG) data, to health outcomes, such as cardiovascular health indicators. As in this example, the functional datum is often the predictor in a regression problem. Other examples include estimating chemical variables from spectroscopic data (Osbourne et al. 1984), predicting annual precipitation from daily temperature data (Ramsay and Silverman 2005), and relating magnetic resonance imaging data or diffusion tensor imaging data to health outcomes (Goldsmith et al. 2011).

We introduce a new approach to regression with functional predictors, based on a kernel mixture functional representation. This representation is designed for predictors that

Dawn B. Woodard is Assistant Professor, School of Operations Research and Information Engineering (ORIE), Cornell University, Ithaca, NY 14853 (E-mail: *woodard@cornell.edu*). Ciprian Crainiceanu is Associate Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205 (E-mail: *ccrainic@hsph.edu*). David Ruppert is Andrew Schultz, Jr., Professor of Engineering, Department of Statistical Science and School of Operations Research and Information Engineering (ORIE), Cornell University, Ithaca, NY 14853 (E-mail: *dr24@cornell.edu*).

Journal of Computational and Graphical Statistics, Volume 00, Number 0, Pages 1–24
DOI: 10.1080/10618600.2012.694765

exhibit features such as dips, bumps, and plateaus whose frequency, location, size, and shape varies stochastically across subjects. We use this representation of the functional covariates for predicting a scalar outcome. Summaries of the representation, such as the frequency of bumps, or the average height or width of the bumps, can have intuitive scientific interpretation; we regress the outcome on these functional summaries. This approach does not require alignment or even a common domain for the subject-specific functions. It also does not require the function observation locations to be equally spaced, and naturally handles missing or co-located data. Mathematically, nearly all existing methods for regression with a functional predictor represent the predictor using a set of linearly independent basis functions; we instead represent the predictor using an uncountable and linearly dependent dictionary of generating elements. Regularization is induced through the prior distribution, and consistency properties still hold. This more flexible representation allows for the above scientific interpretation of events occurring at random locations and the ability to relate those events to the outcome variable.

Our methodology is motivated by data from the SHHS, and by interest in using these data to understand the relationship between sleep characteristics and health outcomes. Some of the challenges associated with these data are the large numbers of patients, large number of observations per patient, missing data, and complex variability patterns. Most importantly, the sleep EEG data are desynchronized across patients; each patient goes through sleep cycles whose length, number, and features vary across subjects and may be related to the health outcomes. This desynchronization is clearly visible in Figure 1, which shows the EEG signals of four subjects. Standard functional regression approaches incorrectly treat the time series as being aligned; registration of the signals (Ramsay and Silverman 2005) to achieve alignment is inappropriate since the number and features of the sleep cycles vary widely across subjects. In contrast, our predictor representation naturally captures the presence of sleep cycles occurring at random times.

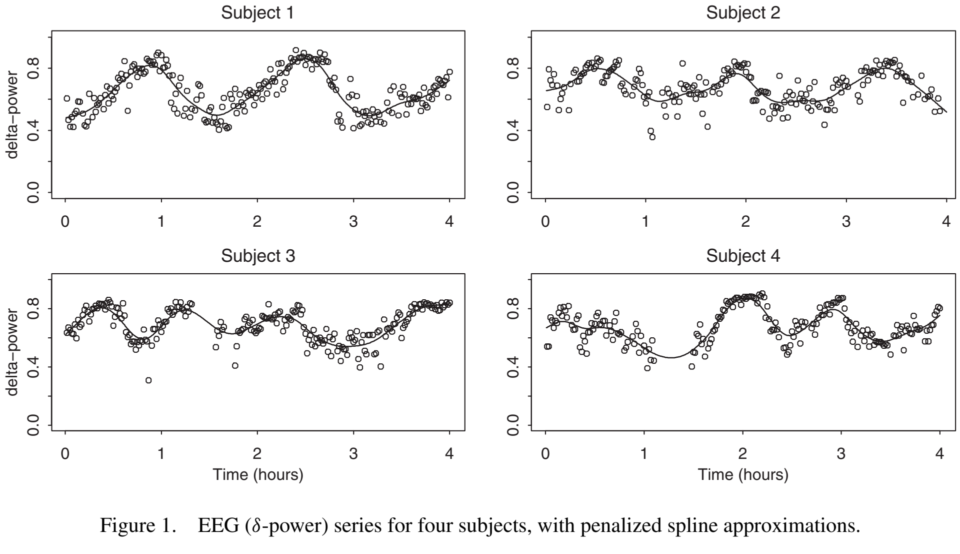

Figure 1. EEG (δ-power) series for four subjects, with penalized spline approximations.

To fully account for uncertainty in the functions when obtaining inferences for the regression coefficients, we introduce a joint hierarchical model. Failure to account for this uncertainty can lead to biased estimation and incorrect standard errors (Carroll et al. 2006; Crainiceanu, Staicu, and Di 2009). We call our approach hierarchical adaptive regression kernels (HARK). We show that this method is computationally feasible using an approximation to the posterior distribution obtained via a technique called "modularization" or "cutting feedback" (Liu, Bayarri, and Berger 2009; Lunn et al. 2009; McCandless et al. 2010). With this approximation, the slower part of the computation can be done in parallel across subjects, so even a very large number of subjects can be handled easily. In the process, we correct a problem in the modularization method as used in the above citations: while the modularization approximation yields a well-defined joint distribution on the parameters, the Markov chain methods previously used were not guaranteed to converge to that distribution.

The functional representation we use has been employed previously for consistent nonparametric estimation of a *single function* under the name Lévy adaptive regression kernels (LARK; Clyde, House, and Wolpert 2006). Our approach is different from LARK because we: (1) model a *population of functions*, where the frequency, location, and shape of the features vary across subjects; (2) predict an outcome on that population; and (3) introduce a method for efficient posterior computation for the joint functional and exposure models, over the entire population.

We contrast our approach with a state-of-the-art method for regression with a functional predictor, penalized functional regression (PFR; Goldsmith et al. 2011; Crainiceanu and Reiss 2011). Like some other existing methods (Cardot, Ferraty, and Sarda 2003; Reiss and Ogden 2007), PFR assumes that the functional predictor has a common domain across subjects (perhaps after registration), and that the outcome is linearly related to the function value $f_i(x)$ at each location x. The latter implicitly assumes alignment of the predictor across subjects.

We show that for simulated data that include features that occur at random locations, PFR requires a large amount of data to detect any relationship between the predictor and the outcome, and is unable to represent that relationship accurately for any sample size. In contrast, HARK can capture the relationship effectively even for small sample sizes (see Section 5). Conversely, if the data are generated from the functional linear model (defined in Section 2), PFR works well but HARK does not effectively capture the relationship between predictor and outcome. Applying both HARK and PFR to the sleep data, HARK detects several important relationships that PFR does not—for instance, that the frequency and magnitude of fluctuations in the EEG series are negatively associated with the respiratory disturbance index (RDI) of the subject.

Alternative methods for joint modeling of a functional predictor and a scalar outcome are given in Bigelow and Dunson (2009) and Dunson (2010). These approaches represent the subject-specific functions $f_i(x)$ using a basis expansion and estimate the distribution of the resulting coefficient vectors, either jointly with the outcome or conditional on the outcome. Since these methods relate the outcome to the coefficients of a basis expansion, they still assume alignment of the subject-specific functions. In relating the coefficients to the outcome, these methods are extremely flexible. Since they must estimate a distribution with dimension greater than or equal to the number of bases, for statistical efficiency reasons, they either use a small number of bases (the choice of which can be learned from the data; Bigelow and Dunson 2009) or make substantial assumptions about the structure of this distribution (Dunson 2010).

Section 2 introduces our model for the subject-specific functions, and Section 3 links the subject-level models hierarchically to a regression model. Section 4 describes our computational procedure. Section 5 gives a simulation study, and HARK and PFR are applied to the sleep data in Section 6.

2. FUNCTIONAL DATA MODEL

First, we describe the nonparametric functional representation we use. For each subject i, we have noisy observations $\{W_i(x_{ik})\}_{k=1}^{K_i}$ of a functional predictor $f_i(x)$ at locations x_{ik} in the (potentially subject-specific) domain $\mathcal{X}_i$. To represent the functional predictor, we use the mixture form:

$$f_i(x) = \beta_{0i}(x) + \sum_{m=1}^{M_i} \gamma_{im} \mathcal{K}(x, s_{im}). \tag{1}$$

Here, $\mathcal{K}(x, s)$ is a specified kernel function on $\mathcal{X}_i \times \mathcal{S}$, where the parameters of the kernel are defined on a space $\mathcal{S}$. Also, $M_i < \infty$ is the number of mixture components, and $\gamma_{im} \in \mathbb{R}$ and $s_{im} \in \mathcal{S}$ are the magnitudes and parameter vectors of those mixture components, respectively. All of these quantities, except the kernel function $\mathcal{K}$, are taken to be unknown. The scaling and other parameters are allowed to vary between the components, "adapting" to the local features of the function. The background signal $\beta_{0i}(x)$ typically has a parametric form, such as a polynomial function of x; for simplicity, we take $\beta_{0i}(x) = \beta_{0i}$ to be an unknown constant, but extensions to more general forms are straightforward (cf. Best, Ickstadt, and Wolpert 2000).

Contrast this with standard functional data analysis approaches that do not model the functional predictor directly; instead, they typically assume that the outcome is linearly related to $\int f_i(x)\beta(x)dx$ for some function $\beta(x)$. This framework is called the *functional linear model with scalar response* (Cardot, Ferraty, and Sarda 2003; Müller and Stadtmüller 2005). It is implemented by representing the functional predictor as a weighted sum of linearly independent basis functions and using the coefficients from this representation as predictors in a standard regression model. Examples of bases include principal component (PC) functions (cf. Müller and Stadtmüller 2005; Goldsmith et al. 2011), spline bases (cf. James 2002), the Fourier basis (cf. Ramsay and Silverman 2005), or partial least-squares factors (cf. Goutis and Fearn 1996; Reiss and Ogden 2007). Estimation in the regression model proceeds via traditional methods (e.g., least squares) or by incorporating a roughness penalty (Marx and Eilers 1999; Cardot, Ferraty, and Sarda 2003).

The kernel mixture (1) can also be viewed as representing $f_i(\cdot)$ via a linear expansion, but instead of linearly independent basis functions, an uncountable and linearly dependent dictionary of generating elements is used. Although the coefficients are no longer unique, one can obtain a more parsimonious representation, by using fewer nonzero coefficients to attain the same accuracy. Regularization is induced through the prior distribution; this effect was described in detail by Clyde and Wolpert (2007). Sufficient conditions for consistency of function estimation using such kernel mixture models are given in Pillai (2008).

The kernel mixture representation (1) has been used previously for Bayesian estimation of a single function by applying a Lévy process prior for the mixture components; this approach is called Lévy adaptive regression kernels (LARK; Clyde, House, and Wolpert 2006; Wolpert, Clyde, and Tu 2011). LARK models have been applied to one-dimensional curve fitting and spatial and spatiotemporal modeling in Wolpert, Clyde, and Tu (2011) and Woodard, Wolpert, and O'Connell (2010). They have also been used for peak identification in mass spectroscopy (House 2006; Clyde, House, and Wolpert 2006; House et al. 2010).

Instead of estimating a single function as in these citations, we estimate a population of functions where the number, magnitude, and parameters of the mixture components vary across the population; furthermore, we use the functions to predict outcomes. Our approach is distinct, for instance, from the approach of MacLehose and Dunson (2009), which represents the functions using the kernel mixture form (1) but takes a common set of mixture components across subjects. It is also different from the hierarchical model of House (2006), which uses (1) but takes a common $M_i = M$ and specifies $\{(\gamma_{im}, s_{im})\}_{m=1}^{M_i}$ as random effects centered at a common set of components $\{(\gamma_m, s_m)\}_{m=1}^{M}$.

The most natural way to model a population of functions is via a hierarchical specification, allowing borrowing of information across the population (Bigelow and Dunson 2007; MacLehose and Dunson 2009). However, this approach is very computationally intensive when the number of subjects is large, since all functions must be simultaneously estimated. To facilitate computation, we will assign independent prior distributions to the subject-specific functional representations. However, in the spirit of hierarchical modeling, we estimate the hyperparameters of those prior distributions from the population of functions in an empirical Bayes fashion. This approach can be viewed as an empirical version of hierarchical modeling of the population of functions that scales more effectively to large and complex applications such as the sleep study (SHHS).

We do not use a Lévy process prior for the functional representations. The prior we use is more flexible and facilitates interpretation of the mixture components by minimizing the occurrence of redundant or spurious mixture components.

In the sleep application and our simulations, the functions are defined on a time domain $\mathcal{X}_i \subset \mathbb{R}$, and we use the unnormalized Gaussian kernel

$$\mathcal{K}(x, s) = \exp\{-(x - \mu)^2/(2\sigma^2)\} \tag{2}$$

for $s = (\mu, \sigma^2)$ so that $\mathcal{S} = \mathcal{X}_i \times \mathbb{R}^+$. This kernel effectively captures many of the features seen in the sleep data (see Section 6.2). One should choose the kernel form appropriately in accordance with the context; in the air pollution application of Wolpert, Clyde, and Tu (2011), for instance, a double-exponential kernel is used. Unlike support vector machines and related approaches, symmetry or even continuity of $\mathcal{K}$ is not required, so there is a great deal of flexibility in this choice. It is even possible to use multiple types of kernels so that s includes an indicator of the type and $\mathcal{K}(x, s)$ has a mixture form. The need to select one or more appropriate kernel forms in HARK is analogous to the need to select an appropriate set of basis functions when using standard functional data analysis approaches.

In the rest of this section, we provide an example, then complete the statistical model for the subject-specific functions by specifying a likelihood based on $f_i(x)$ and prior distributions for the unknown quantities. We link the models for the subject-specific functions hierarchically to a regression model for the outcome in Section 3.

2.1 EXAMPLE

We illustrate Bayesian function estimation using the representation (1) by applying LARK to the "Bumps" test function given in Donoho and Johnstone (1994), which is a mixture of kernels of the form $\mathcal{K}(x, s) = (1 + |(x - \mu)/\sigma|)^{-4}$ for $s = (\mu, \sigma)$. We simulated a single time series (so that $i = 1$), plotted in Figure 2, by adding $N(0, 0.01)$ noise to the test function at 2048 equally spaced locations in the domain. To illustrate that the representation (1) can be accurately recovered, we apply LARK using kernels of the form $\mathcal{K}(x, s) = (1 + |(x - \mu)/\sigma|)^{-4}$.

The function estimate is shown in Figure 2, superimposed on the true function; the two are indistinguishable. The LARK representation of the function, as given by the mixture components from a single posterior sample, is also shown. It is clear that the test function has been recovered accurately and represented parsimoniously. There are 11 mixture components in the test function, and 12 in the posterior sample. One of these is redundant; our prior specification in HARK will be designed to minimize the occurrence of such redundant components.

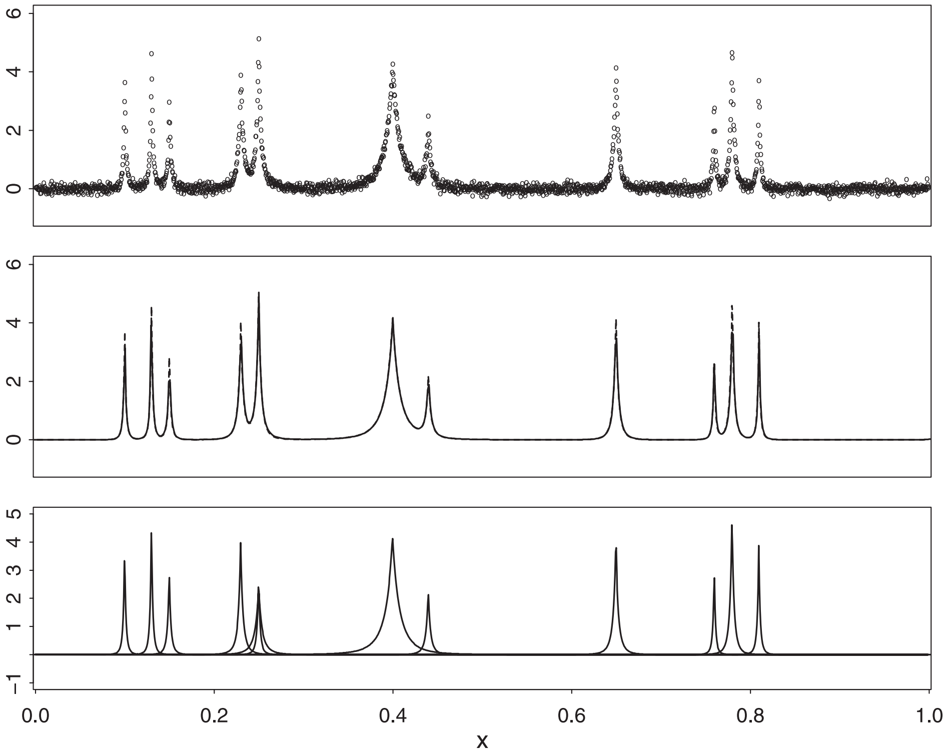

Figure 2. Estimation of a test function. Top: simulated data. Middle: test function (dashed curve) and LARK estimate (solid curve) are indistinguishable. Bottom: LARK representation, given by the mixture components from a single posterior sample.

2.2 LIKELIHOOD

Next, we specify the likelihood function, that is, the probability density for the observations $\{W_i(x_{ik}) \approx f_i(x_{ik})\}_{k=1}^{K_i}$ for each subject i. We use a normal error model $W_i(x_{ik}) \sim N(f_i(x_{ik}), \tau_i^2)$ for some variance parameter τ_i^2, leading to the likelihood

$$\left(2\pi\tau_i^2\right)^{-K_i/2} \exp\left\{-\frac{1}{2\tau_i^2}\sum_{k=1}^{K_i}\left[W_i(x_{ik}) - f_i(x_{ik})\right]^2\right\}, \tag{3}$$

which is a function of the parameter vector $\omega_i = (\beta_{0i}, \tau_i^2, \{(\gamma_{im}, s_{im})\}_{m=1}^{M_i})$, which includes τ_i^2, the intercept β_{0i}, and the set of mixture component magnitudes and parameters.

2.3 PRIOR DISTRIBUTION

For a Bayesian model, we must specify a prior distribution for each of the elements of the parameter vector ω_i, as defined in Section 2.2. One can obtain an empirical estimate $\hat{\omega}_i$ of ω_i for each subject i as described in Appendix A in the online supplementary materials; the distribution of these estimates across subjects tells us what values of the parameters are reasonable, and will guide our prior specification. For instance, a Poisson prior distribution might be an obvious choice for a prior on the number of mixture components M_i. Indeed, this is used in LARK prior specification since only a single functional observation is available, and so, there is not enough information in the data to question this choice. However, in the context of estimating a population of functions, the information in the data may conflict with this choice. For applications where only a small number of mixture components is typical, the Poisson distribution can be overdispersed, putting too much prior mass on values of M_i above and below than what is reasonable in that application. For instance, in the sleep application, the empirical estimates $\hat{M}_i$ nearly all fall in the range 3–8 and have a mean of 4.2. A Poisson distribution with mean 4.2 places almost 24% of its probability outside of this range; such a prior can, for instance, lead to overestimation of the number of mixture components by inclusion of spurious mixture components (redundant components or components with small magnitude). When we use the mixture representation of the function to predict outcomes, it is important that the features of the functional data are captured without redundancy. For this reason, we instead use a discrete prior for M_i, with the probability vector equal to the empirical frequencies in $\{\hat{M}_i\}_{i=1}^n$, where n is the number of subjects.

With this choice, the function f_i is C^∞ smooth so long as the kernel $\mathcal{K}$ is C^∞ smooth [e.g., for (2)], since $M_i < \infty$ almost surely. Conditional on M_i, the γ_{im} values are assumed to be independently distributed according to a symmetric gamma distribution,

$$\pi(\gamma) = \frac{\rho^\alpha}{2\Gamma(\alpha)}|\gamma|^{\alpha-1}e^{-\rho|\gamma|}, \tag{4}$$

that is, a gamma distribution for $|\gamma_{im}|$. Since this prior is symmetric about $\gamma = 0$, we have $\mathsf{E}(f_i(x)) = \beta_{0i}$ for all x. The prior (4) is closely related to the symmetric gamma random field, which is the most common Lévy prior distribution for cases where γ can take either positive or negative values (Clyde, House, and Wolpert 2006; Wolpert, Clyde, and Tu 2011). Specifically, the prior for γ, conditional on M, in the symmetric gamma random field is equal to Equation (4) with $\alpha = 0$; we have chosen the more flexible form (4) because the prior mean and variance of $|\gamma|$ can be separately controlled. Selection of α and ρ is described in Appendix B in the supplementary materials. One could also consider putting a random effect prior on the γ_{im} within subject i, but this is nonstandard in LARK models, and would be unlikely to yield improvement for our application, where we will see that the estimated number of mixture components per subject is relatively small.

In models that use linearly independent basis expansions, the coefficients are typically assigned shrinkage prior distributions, such as normal or double-exponential priors, to yield a smooth estimated function (Lang and Brezger 2004; Bigelow and Dunson 2007). In the kernel mixture model (1), the summation is over a set of subject-specific dictionary elements, and parsimony is induced by the prior on M_i rather than by the prior on γ_{im}. It is not necessary to shrink the coefficients γ_{im} toward zero, and in fact, since our goal is interpretation of the mixture components, we do not want the γ_{im} to be estimated to be very close or equal to zero. For this reason, the prior (4) is more appropriate: for $\alpha > 1$, it has one strictly positive and one strictly negative mode and places little probability near zero since $\pi(0) = 0$. Our method of selecting α typically yields $\alpha > 1$, as explained in Appendix B.

The μ_{im} values are assumed to be a priori independently uniformly distributed on the domain $\mathcal{X}_i$, as in Best, Ickstadt, and Wolpert (2000), Woodard, Wolpert, and O'Connell (2010), and Wolpert, Clyde, and Tu (2011). The σ_{im}^2 parameters are assumed to independently have an inverse gamma distribution, with shape and scale parameters $\alpha_\sigma > 0$ and $\rho_\sigma > 0$, that is, $\pi(\sigma^2) = \frac{\rho_\sigma^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)}(\sigma^2)^{-(\alpha_\sigma+1)}e^{-\rho_\sigma/\sigma^2}$, as in Wolpert, Clyde, and Tu (2011).

We assign β_{0i} a normal prior distribution and τ_i^2 an inverse gamma prior distribution in accordance with common practice (Gilks, Richardson, and Spiegelhalter 1996), although these choices are flexible. The selection of the hyperparameters for these priors, as well as the hyperparameters α, ρ, α_σ, and ρ_σ, is via empirical Bayes (cf. Carlin and Louis 2008), as described in Appendix B.

3. REGRESSION USING THE FUNCTIONAL REPRESENTATION

Next, we define HARK by combining the model for the subject-specific functions with a regression model for the outcome. This approach is reasonable when we hypothesize that the functions include features such as spikes occurring at random locations, and that the frequency, average magnitude, average duration, etc., of the features may be related to the outcome.

Consider the case of a single functional predictor; multiple functional predictors can be handled analogously, for instance, by assuming additivity of their effects. Take a vector of summary statistics of the functional representation; for example, for the case of a Gaussian kernel, one can take $\theta_i \equiv \theta(\omega_i) = (\beta_{0i}, \tau_i^2, M_i, \bar{\gamma}_i, \bar{\mu}_i, \bar{\sigma}_i^2)$, where $\bar{\gamma}_i = \mathbf{1}_{\{M_i>0\}} \sum_{m=1}^{M_i} |\gamma_{im}|/M_i$, $\bar{\mu}_i = \mathbf{1}_{\{M_i>0\}} \sum_{m=1}^{M_i} \mu_{im}/M_i$, and $\bar{\sigma}_i^2 = \mathbf{1}_{\{M_i>0\}} \sum_{m=1}^{M_i} \sigma_{im}^2/M_i$.

Taking a continuous outcome variable Y_i, and allowing for additional scalar covariates V_i, our model is

$$Y_i = V_i\psi^T + \sum_{j=1}^{J} g_j(\theta_{ij}) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \phi^2), \tag{5}$$

where ψ is a regression coefficient vector, J is the length of the vector θ_i, g_j are unknown functions, and $\phi^2 > 0$ is the residual variance. Extension to count-valued, binary, and other types of outcome is straightforward using the generalized linear model framework (Dey, Ghosh, and Mallick 2000).

A variety of specifications are possible for the functions g_j (DiMatteo, Genovese, and Kass 2001; Lang and Brezger 2004); we represent them using quadratic B-splines (de Boor 2001). For knot specification, one can place prior distributions on the number and locations of the knots and estimate them from the data (DiMatteo, Genovese, and Kass 2001); however, fixed knots lead to faster convergence of the computational method described in Section 4. For the function g_j on domain $[a_j, b_j]$, we take a fixed number R of interior knots at equally spaced quantiles of some estimates of the predictor values $\{\theta_{ij}\}_{i=1}^n$. For this purpose, we estimate θ_{ij} using its posterior mean, given $\{W_i(x_{ik})\}_{k=1}^{K_i}$. With R interior knots, there are $R + 3$ quadratic B-spline basis functions for each predictor, the last one of which is omitted for identifiability, leaving $P = R + 2$ bases; the intercept is included as a separate term. In the simulations and data analysis, we take $R = 3$.

Denoting the resulting B-spline basis functions by $B_{jp}(\cdot)$, we write the functions g_j as

$$\sum_{j=1}^{J} g_j(\theta_{ij}) = \eta_0 + \sum_{j=1}^{J}\sum_{p=1}^{P} \zeta_{jp}\eta_{jp}B_{jp}(\theta_{ij}),$$

where η_0 and η_{jp} are regression coefficients, and $\zeta_{jp} \in \{0, 1\}$ are unknown inclusion indicators for each term. Define

$$\begin{aligned}
\zeta &= (1, \zeta_{11}, \ldots, \zeta_{1P}, \zeta_{21}, \ldots, \zeta_{2P}, \ldots, \zeta_{J1}, \ldots, \zeta_{JP})\\
\eta &= (\eta_0, \eta_{11}, \ldots, \eta_{1P}, \eta_{21}, \ldots, \eta_{2P}, \ldots, \eta_{J1}, \ldots, \eta_{JP})\\
Z_i &= (1, Z_{i11}, \ldots, Z_{i1P}, Z_{i21}, \ldots, Z_{i2P}, \ldots, Z_{iJ1}, \ldots, Z_{iJP}), \quad \text{where} \quad Z_{ijp} = B_{jp}(\theta_{ij}),
\end{aligned}$$

and take Z_i^ζ and η^ζ to indicate the subvectors of Z_i and η, respectively, corresponding to the nonzero elements of ζ. The regression model (5) can then be written as a linear model:

$$Y_i = V_i\psi^T + Z_i^\zeta \eta^{\zeta T} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \phi^2). \tag{6}$$

Model selection will be done by estimating ζ jointly with the other unknowns. If $\sum_{p=1}^{P} \zeta_{jp} = 0$, then the jth predictor θ_{ij} drops out of the model. We specify the prior $\pi(\zeta)$ to give equal prior probability to a predictor being included in the model or not, so that $\pi(\sum_{p=1}^{P} \zeta_{jp} = 0) = 1/2$ for each j. Conditional on $\sum_{p=1}^{P} \zeta_{jp} \neq 0$, we place equal prior probability on all possible values of the vector $(\zeta_{j1}, \ldots, \zeta_{jP})$, yielding

$$\pi(\zeta) = \prod_{j=1}^{J} \frac{1}{2}\left(\frac{1}{2^P - 1}\right)^{\mathbf{1}_{\{\sum_p \zeta_{jp} \neq 0\}}} \propto (2^P - 1)^{\sum_{j=1}^{J} \mathbf{1}_{\{\sum_p \zeta_{jp} = 0\}}}.$$

As in DiMatteo, Genovese, and Kass (2001), we take the prior distribution $\pi(\phi^2) \propto 1/\phi^2$, which is considered to be noninformative since it corresponds to a uniform prior on $\log \phi$. We have also tried an empirical Bayes prior for ϕ^2, which gave nearly identical results. The prior on η^ζ, given ζ and ϕ^2, is

$$\eta^\zeta|\zeta, \phi^2 \sim N_{|\zeta|}(0, \phi^2 n(Z^{\zeta T} Z^\zeta)^{-1}),$$

where $|\zeta| = 1 + \sum_{j,p} \zeta_{jp}$ and $Z^\zeta = (Z_1^{\zeta T}, \ldots, Z_n^{\zeta T})^T$. This prior has been used by many authors and is called a "unit-information" prior because, loosely speaking, the amount of information in the prior is equal to the amount of information in a single observation (Smith and Kohn 1996; Pauler 1998; DiMatteo, Genovese, and Kass 2001). For the covariate regression coefficients ψ, one can either use a uniform prior distribution $\pi(\psi) \propto 1$ or a variable selection prior (George and McCulloch 1997).

Having specified both the prior and the likelihood structure, we can obtain the joint posterior distribution of all unknowns as follows. Denote prior, likelihood, and posterior by π as distinguished by their arguments, and let W_{ik} be shorthand for $W_i(x_{ik})$; then the joint posterior is

$$\begin{aligned} &\pi\left(\{\omega_i\}_{i=1}^n, \zeta, \eta, \psi, \phi^2|\{W_{ik}, Y_i\}_{i,k}\right) \\ &\propto \pi(\zeta, \eta, \psi, \phi^2) \prod_{i=1}^n \pi\left(\omega_i\right) \pi\left(\{W_{ik}\}_{k=1}^{K_i}|\omega_i\right) \pi(Y_i|\omega_i, \zeta, \eta, \psi, \phi^2). \end{aligned} \tag{7}$$

Here, $\pi(\zeta, \eta, \psi, \phi^2)$ and $\pi(\omega_i)$ are specified in this section and in Section 2.3, respectively; $\pi(\{W_{ik}\}_{k=1}^{K_i}|\omega_i)$ is given in (3), and $\pi(Y_i|\omega_i, \zeta, \eta, \psi, \phi^2)$ is specified by (6).

Estimation of any unknown quantity of interest is then based on the posterior distribution (7). We first select the model by choosing the maximum a posteriori (MAP) value $\hat{\zeta}$ of ζ (computed as described in Section 4). Then, inference for any function h of the remaining parameters is done conditional on $\hat{\zeta}$, by obtaining a point estimate (the posterior mean) or interval estimate [the $a/2$ and $1 - a/2$ posterior quantiles for $a \in (0, 1)$] of h. For example, we can obtain point and interval estimates of the regression function $g_j(\theta)$ evaluated at a specific value θ, or of the predictor function $f_i(x)$ at any location x. Computation of $\hat{\zeta}$ and of the posterior mean of an arbitrary function h conditional on $\hat{\zeta}$ are described in Section 4; posterior quantiles of h can be computed in the same way.

4. HARK COMPUTATION

We give a method for efficient posterior computation based on a two-stage approach that propagates the uncertainty from the first stage into the second stage. This approach is justified by an approximation to the posterior distribution based on *modularization* (Liu, Bayarri, and Berger 2009). This approximation avoids the potential computational pitfall of Bayesian inference for regression using an incompletely observed functional predictor, namely that the parameters of the functional signals are, in theory, dependent across subjects a posteriori; taking this dependence into account requires simultaneous estimation of all the functions. Although such joint estimation can be done in some cases for datasets with up to several hundred observations (Bigelow and Dunson 2009), such an approach is unlikely to scale well to thousands or tens of thousands of observations. Our approximation assumes that the functional data $\{W_i(x_{ik})\}_{k=1}^{K_i}$ contain far more information about the function f_i than does the outcome Y_i, so we can ignore Y_i when estimating f_i. This permits the function estimation to be performed independently (and in parallel) across subjects, which constitutes the first stage of computation. The second stage consists of inference for the regression model (5), using the posterior distribution of f_i for each i from the first stage.

Our modularization approach has the additional advantage that any potential misspecification of the regression model (5) does not negatively affect estimation of the subject functions f_i. As pointed out by Liu, Bayarri, and Berger (2009), misspecification in one part of a Bayesian model can, in some cases, have a dramatic effect on estimates in another component of the model, an effect that can be prevented by appropriate modularization. In our context, if there is lack of fit in either the additivity assumption or the normality assumption in (5), the estimates of f_i are unaffected. When using the modularization approach, we correct a problem in the original method: namely that the Markov chains used were not guaranteed to converge to the modularized approximation.

We will show how to use Monte Carlo methods to efficiently compute an approximation to the MAP model $\hat{\zeta}$ and to the posterior mean of any quantity $h(\{\omega_i\}_{i=1}^n, \eta, \psi, \phi^2)$ of interest, conditional on $\hat{\zeta}$. This will be done by constructing a stochastic process with limiting distribution equal to an approximation $\tilde{\pi}$ of the posterior. This will yield sample vectors $(\{\omega_i^{(\ell)}\}_{i=1}^n, \zeta^{(\ell)}, \eta^{(\ell)}, \psi^{(\ell)}, \phi^{2(\ell)})$ indexed by $\ell = 1, \dots, L$ that converge in distribution to $\tilde{\pi}$. As in Markov chain Monte Carlo methods, we will then estimate $\hat{\zeta}$ as the most frequently occurring value in $\{\zeta^{(\ell)}\}_{\ell=1}^L$. Conditional on $\hat{\zeta}$, we estimate the approximate posterior mean $\mathsf{E}_{\tilde{\pi}}(h|\hat{\zeta})$ using the ergodic average:

$$\frac{\sum_{\ell=1}^L \mathbf{1}_{\{\zeta^{(\ell)}=\hat{\zeta}\}} h\big(\{\omega_i^{(\ell)}\}_{i=1}^n, \eta^{(\ell)}, \psi^{(\ell)}, \phi^{2(\ell)}\big)}{\sum_{\ell=1}^L \mathbf{1}_{\{\zeta^{(\ell)}=\hat{\zeta}\}}}. \tag{8}$$

Validity of this approach is discussed later.

We obtain our approximation $\tilde{\pi}$ by decomposing the joint posterior distribution (7) as

$$\begin{aligned}
&\pi\left(\{\omega_i\}_{i=1}^n, \zeta, \eta, \psi, \phi^2 | \{W_{ik}, Y_i\}_{i,k}\right)\\
&\quad = \pi\left(\{\omega_i\}_{i=1}^n \big| \{W_{ik}, Y_i\}_{i,k}\right) \ \pi(\zeta, \eta, \psi, \phi^2 | \{\omega_i, W_{ik}, Y_i\}_{i,k})\\
&\quad = \pi\left(\{\omega_i\}_{i=1}^n \big| \{W_{ik}, Y_i\}_{i,k}\right) \ \pi\left(\zeta, \eta, \psi, \phi^2 | \{\omega_i, Y_i\}_{i=1}^n\right)
\end{aligned}$$

and applying modularization:

$$\pi\left(\{\omega_i\}_{i=1}^n | \{W_{ik}, Y_i\}_{i,k}\right) \approx \pi\left(\{\omega_i\}_{i=1}^n \big| \{W_{ik}\}_{i,k}\right) = \prod_{i=1}^n \pi\big(\omega_i | \{W_{ik}\}_{k=1}^{K_i}\big).$$

The last equality holds because both the prior distribution for ω_i and the likelihood (3) for $\{W_{ik}\}_{k=1}^{K_i}$, given ω_i, are independent across i. The resulting approximate posterior density is

$$\tilde{\pi}\left(\{\omega_i\}_{i=1}^n, \zeta, \eta, \psi, \phi^2\right) = \pi\left(\zeta, \eta, \psi, \phi^2 | \{\omega_i, Y_i\}_{i=1}^n\right) \prod_{i=1}^n \pi\big(\omega_i \{W_{ik}\}_{k=1}^{K_i}\big). \tag{9}$$

This simplification allows us to propose a two-stage computational method.

Method for approximate posterior simulation

Stage 1. For each subject i, obtain L sample vectors $\omega_i^{(\ell)}$ as the iterations of an ergodic Markov chain with invariant density $\pi(\omega_i|\{W_{ik}\}_{k=1}^{K_i})$. This computation can be performed in parallel across subjects.

Stage 2. Take the set of ℓ-indexed sample vectors $\{\omega_i^{(\ell)}\}_{i=1}^n$ from Stage 1 and take arbitrary initial values $(\zeta^{(0)}, \eta^{(0)}, \psi^{(0)}, \phi^{2(0)})$. For each $\ell = 1, \ldots, L$: (1) starting at $(\zeta^{(\ell-1)}, \eta^{(\ell-1)}, \psi^{(\ell-1)}, \phi^{2(\ell-1)})$, simulate N_ℓ iterations of an ergodic Markov chain with invariant density $\pi(\zeta, \eta, \psi, \phi^2 \mid \{\omega_i^{(\ell)}, Y_i\}_{i=1}^n)$; (2) save the last value of this chain as $(\zeta^{(\ell)}, \eta^{(\ell)}, \psi^{(\ell)}, \phi^{2(\ell)})$.

The sample vectors $(\{\omega_i^{(\ell)}\}_{i=1}^n, \zeta^{(\ell)}, \eta^{(\ell)}, \psi^{(\ell)}, \phi^{2(\ell)})$ converge to $\tilde{\pi}$ in the following sense. For every $\xi > 0$, for all L large enough and all N_ℓ large enough for each ℓ, the total variation distance between $\tilde{\pi}$ and the distribution of $(\{\omega_i^{(L)}\}_{i=1}^n, \zeta^{(L)}, \eta^{(L)}, \psi^{(L)}, \phi^{2(L)})$ is less than ξ. This is proven in Appendix C in the supplementary materials, under mild regularity conditions on the Markov transition kernels in Stages 1 and 2. Intuitively, this result holds because for N_ℓ large, Stage 2 above is roughly the same as obtaining a single sample $(\zeta^{(\ell)}, \eta^{(\ell)}, \psi^{(\ell)}, \phi^{2(\ell)})$ from the full conditional density $\pi(\zeta, \eta, \psi, \phi^2 \mid \{\omega_i^{(\ell)}, Y_i\}_{i=1}^n)$. If one could instead obtain a sample precisely from this full conditional density, the resulting sample vectors $(\{\omega_i^{(\ell)}\}_{i=1}^n, \zeta^{(\ell)}, \eta^{(\ell)}, \psi^{(\ell)}, \phi^{2(\ell)})$ for $\ell = 1, \ldots, L$ would form the iterations of an ergodic Markov chain on the joint space, with limiting distribution $\tilde{\pi}$. This analogy suggests that our computational procedure is valid in a stronger sense, namely convergence of $\hat{\zeta}$ to the MAP estimate of ζ, and convergence of the Monte Carlo estimate (8) to the value $\mathsf{E}_{\tilde{\pi}}(h|\hat{\zeta})$. We are currently investigating these properties.

Our computational method corrects a problem with the modularization approach as implemented by Liu, Bayarri, and Berger (2009), Lunn et al. (2009), and McCandless et al. (2010), and even in the popular software package WinBUGS (Spiegelhalter et al. 2003). These authors considered the posterior density $\pi(x, y|\mathcal{D}_1, \mathcal{D}_2)$ of two vectors of unknowns x and y, conditional on two statistics $\mathcal{D}_1, \mathcal{D}_2$. This can be written as

$$\pi(x, y|\mathcal{D}_1, \mathcal{D}_2) = \pi(x|\mathcal{D}_1, \mathcal{D}_2)\pi(y|x, \mathcal{D}_1, \mathcal{D}_2).$$

Due to various modeling and/or computational considerations, they (implicitly or explicitly) replaced this with the following valid joint density on x and y:

$$\nu(x, y) = \pi(x|\mathcal{D}_1)\pi(y|x, \mathcal{D}_1, \mathcal{D}_2),$$

just as we do to obtain (9). To perform inference, they simulated a Markov chain that repeats the steps: (1) update x via single iteration of a Markov kernel Q_x with invariant density $\pi(x|\mathcal{D}_1)$; (2) update y via a single iteration of a Markov kernel $Q_{y|x}$ with invariant density $\pi(y|x, \mathcal{D}_1, \mathcal{D}_2)$. This seems natural since it imitates a standard Gibbs or Metropolis-within-Gibbs sampler for x and y, but uses only information from $\mathcal{D}_1$ when updating x. However, note that $\pi(x|\mathcal{D}_1) = \nu(x) = \int \nu(x, y)dy$ and that $\pi(y|x, \mathcal{D}_1, \mathcal{D}_2) = \nu(y|x) = \nu(x, y)/\nu(x)$. So, their procedure corresponds to first updating x according to a Markov kernel with invariant density $\nu(x)$ and then updating y according to a Markov kernel with invariant density $\nu(y|x)$. This procedure does not, in general, converge to the joint density $\nu(x, y)$, and is not known to be valid in any sense (Gelfand and Smith 1990). Lunn et al. (2009) acknowledged such a concern: "it is possible that a joint distribution with the simulated properties does not even exist." We consider our procedure above to be a corrected version of this existing approach, in a specific context.

Choices of L and N_ℓ are described in the following subsections. In our simulations and data analysis, the computation time for each Markov chain from Stage 1 is less than 10 min on a single processor with 2.66-GHz speed and 2-GB memory. Stage 2 takes about 15 min for the simulation studies, and about an hour for the data analysis (which has thousands of subjects). So, the total run time of our method, in the parallel environments characteristic of modern computing (Chappell 2010), is less than 75 min in these examples. Software to implement HARK is provided in the online supplementary materials.

4.1 STAGE 1 COMPUTATION

Clyde, House, and Wolpert (2006) and Wolpert, Clyde, and Tu (2011) provided robust methods for simulation from the posterior distribution of a LARK model for a single function. In Stage 1, we use a very similar Markov chain method to sample from the posterior $\pi(\omega_i|\{W_{ik}\}_{k=1}^{K_i})$ of our subject-specific functional model.

This approach uses the reversible jump extension (Green 1995) of the Metropolis–Hastings algorithm (cf. Tierney 1994). In each iteration of the Markov chain, one of the parameters τ_i^2, β_{0i}, $\{(\gamma_{im}, s_{im})\}_{m=1}^{M_i}$ is updated or sampled from its conditional posterior distribution. An update of $\{(\gamma_{im}, s_{im})\}_{m=1}^{M_i}$ can involve (1) a change in the magnitude γ_{im} or the parameters s_{im} of a single mixture component, (2) the addition or deletion of a mixture component, or (3) the merging of two components or splitting of a single component into two. Split/merge moves are not strictly necessary (without these moves, the chain is still irreducible), but greatly improve the convergence and mixing of the Markov chain.

We choose the number of iterations L by evaluating convergence diagnostics and Monte Carlo standard error estimates for the elements of the summary vector θ_i, as defined in Section 3, as well as for the log-likelihood obtained from (3). We use Geweke's diagnostic (Geweke 1992) and estimate the Monte Carlo standard error using consistent batch means (cf. Flegal, Haran, and Jones 2008). We take L to be large enough so that the Geweke p-values are greater than 0.05 after correction for multiplicity, and that the estimated Monte Carlo standard error is less than 0.5% of the parameter estimate.

Such standard error estimation relies on geometric ergodicity of the Markov chain. We use visual inspection of time series plots of ergodic averages to verify that the Markov chains do not exhibit behavior characteristic of nongeometric convergence, but leave formal proof of geometric ergodicity as future work.

4.2 STAGE 2 COMPUTATION

Next, we define the Markov chain used in Stage 2, having invariant density $\pi(\zeta, \eta, \psi, \phi^2|\{\omega_i^{(\ell)}, Y_i\}_{i=1}^n)$. We omit ψ for notational simplicity, since there are no scalar covariates V_i in our simulation study or data analysis; however, in the presence of scalar covariates, one can concatenate V_i with Z_i^ζ and ψ with η^ζ and proceed as follows.

The parameters η and ϕ^2 can be integrated analytically out of the conditional posterior density $\pi(\zeta, \eta, \phi^2 | \{\omega_i^{(\ell)}, Y_i\}_{i=1}^n)$, yielding

$$\pi\left(\zeta \middle| \left\{\omega_i^{(\ell)}, Y_i\right\}_{i=1}^n\right) \propto \pi(\zeta)\,(n+1)^{-|\zeta|/2}(d^*)^{-c^*}, \tag{10}$$

where

$$c^* = n/2, \quad d^* = \frac{1}{2}[Y^TY - \nu^{*T}\Sigma^{*-1}\nu^*], \quad Y = (Y_1, \ldots, Y_n)^T,$$

$$\nu^* = \Sigma^*(Z^{\zeta T}Y), \quad \Sigma^* = \left(\frac{n+1}{n} Z^{\zeta T} Z^\zeta\right)^{-1}, \quad |\zeta| = 1 + \sum \zeta_{jp}$$

(Smith and Kohn 1996). We will construct a Markov transition kernel Q_ζ with invariant density (10). To use this to obtain a Markov kernel with invariant density $\pi(\zeta, \eta, \phi^2 | \{\omega_i^{(\ell)}, Y_i\}_{i=1}^n)$, one can transition according to Q_ζ and then sample according to the conditional posterior density

$$\begin{aligned}
&\pi\left(\eta, \phi^2 | \zeta, \left\{\omega_i^{(\ell)}, Y_i\right\}_{i=1}^n\right) \\
&\quad = \text{NormalInvGamma}(\nu^*, \Sigma^*, c^*, d^*) \\
&\quad \propto \left(\frac{1}{\phi^2}\right)^{c^*+|\zeta|/2+1} \times \exp\left[-\frac{1}{\phi^2}\left(d^* + \frac{1}{2}(\eta-\nu^*)^T\Sigma^{*-1}(\eta-\nu^*)\right)\right].
\end{aligned}$$

We take Q_ζ to be a Gibbs sampler that updates each element of the vector ζ, conditional on the remaining elements, by sampling according to the conditional posterior distribution for that element as implied by (10).

To choose N_ℓ for each ℓ, we take a common value $N_\ell = N$, chosen to satisfy the convergence diagnostics described in Section 4.1, as applied to the sequence of sample vectors $(\zeta^{(\ell)}, \eta^{(\ell)}, \phi^{2(\ell)})$. Typical values of N are 200–2000 in the simulation study and 20,000 in the data analysis.

5. SIMULATION STUDY

Next, we show that HARK is able to efficiently estimate the regression functions g_j for data simulated according to the model, and we compare HARK to PFR (Goldsmith et al. 2011) on data simulated both from the HARK model and the functional linear model. PFR is a state-of-the-art method for estimation in the functional linear model framework described in Section 2, and is implemented in the R package "refund" (Crainiceanu and Reiss 2011). It represents the functional predictor using a much larger number of PCs than is done in PC regression, capturing nearly all of the information in the functional data. In the resulting regression model, smoothing of the coefficient function $\beta(x)$ is used to enforce parsimony. This method is most closely related to the functional regression framework of Cardot, Ferraty, and Sarda (2003) and Cardot and Sarda (2005), but improves upon it in a number of ways, including: (1) handling functions $f_i(x)$ that are observed with error or missingness, (2) using a connection to mixed-effects models that provides a framework for generalization and a method for stable and efficient computation, and (3) automatic selection of the smoothing parameter.

Although PFR has been shown to work well for data simulated from the functional linear model and for a diffusion tensor imaging study, it is not designed for prediction when the functional data include features occurring at random times, and we will see that it fares poorly on such data. To apply this method, the first K PC functions are first estimated via regularized principal component analysis (PCA), where K is some truncation level. We regularize by smoothing the predictor functions before applying PCA; an alternative is to smooth the covariance matrix, which gives virtually identical results for the following examples. We take $K = 35$; this truncation level was suggested by Goldsmith et al. (2011) and captures 99.9% of the variability in our simulated functional data. After obtaining the PC functions, the outcome Y_i is regressed on the functions $f_i(\cdot)$, as represented in the PC basis. Estimation in the regression model is performed by representing $\beta(x)$ by a power series spline basis and using penalized likelihood maximization.

The basis function approach, as exemplified by PFR, assumes that the expected response is linear and additive in the functional predictor $f_i(x)$ at each location x, rather than being controlled by a highly nonlinear quantity, such as the maximum of f_i or the location of the maximum. While some nonparametric extensions of these methods have been developed, these methods still assume that the functions are aligned across subjects, and (with the exception of Ferraty and Vieu 2004) that the expected response is additive in either $f_i(x)$ at each location x or in m-ary products $\prod_{j=1}^{m} f_i(x_j)$ (Yao and Müller 2010). For data that exhibit features such as spikes that occur at random locations, such models can be insufficient to capture the relationship between the predictor and the outcome. Also, application of the basis function approach means that many bases are needed to capture the many possible occurrence locations of the features.

5.1 SIMULATING FROM THE HARK MODEL

We first generate the predictor for each subject i on the common domain $\mathcal{X} = [1, 100]$. For each subject, we take the expectation $\mathsf{E}(W_i(x)|\mu_i)$ to be flat as a function of x, except for a single "blip" that occurs at the random time μ_i. That is, we define $W_i(x_{ik}) = \beta_0 + \gamma_i \mathcal{K}(x_{ik}; \mu_i, \sigma^2) + \varepsilon_{ik}$, where $\varepsilon_{ik} \sim N(0, \tau^2)$ and

$$\mathcal{K}(x; \mu, \sigma^2) = \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} - \exp\left\{-\frac{(x-(\mu+10))^2}{2\sigma^2}\right\}.$$

The amplitude of the "blip" is γ_i, which is sampled uniformly in the interval (10, 20) for each i. To complete the specification, we take $\mu_i \overset{\text{iid}}{\sim} \text{Unif}(3, 87)$, $\sigma^2 = 5$, $\tau^2 = 1$, $\beta_0 = 0$, and $x_{ik} = k$ for $k = 1, \ldots, 100$. Simulated functions $W_i(x)$ are shown in Figure 3 for several subjects; the observations are shown as points, while $\mathsf{E}(W_i(x)|\mu_i)$ is shown as a curve.

For each subject, we take the outcome to be $Y_i \sim N(\gamma_i, 5)$. We apply HARK and PFR to 10 simulated datasets for each sample size considered ($n = 50, 100, 200, 500$ subjects). For PFR, smoothed versions of the subject-specific functions are first obtained via penalized splines and then the principal components of the smoothed functions are obtained. Between 6 and 10 PCs are required to capture 95% of the variability in the data, depending on the

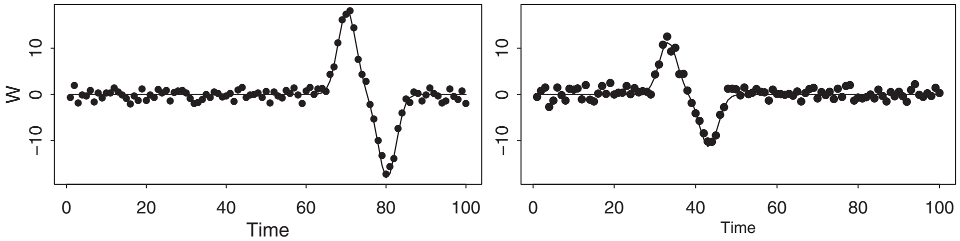

Figure 3. Simulated functional predictor for two subjects.

simulation and the sample size n (6–8 for $n = 50$ and 9–10 for $n = 500$). Between 6 and 9 PCs are required to capture 90% of the variability. These PCs are difficult to interpret; the loading functions of two PCs for a randomly chosen simulation in the case $n = 500$ are shown in Figure 4.

Before we apply smoothing splines to estimate the coefficient function $\beta(x)$, we explore the unsmoothed estimate. This estimate is shown in the left panel of Figure 5, along with point-wise 95% confidence bounds, for a randomly chosen simulation with $n = 500$. The function has periodic behavior, and its interpretation is not clear.

Smoothing a periodic function such as this one is questionable; however, we show the smoothed estimate of $\beta(x)$ in the right panel of Figure 5. Bias-adjusted pointwise confidence bounds are also shown. A significant relationship has been detected between the predictor and the outcome, since the confidence bounds exclude zero for large portions of the domain. The smoothed estimate of $\beta(x)$ suggests that high values of the predictor at the beginning of the time series, and low values of the predictor at the end of the time series, may be associated with higher values of the outcome. This effect is technically correct and is an artifact of the shape of the "blip," namely an upward spike, followed by a downward spike. However, this result does not capture the crucial fact: that the outcome is highly correlated with the magnitude of a particular feature that occurs at a variable time. Results of the Goldsmith et al. (2011) method are very similar for other sample sizes ($n = 50, 100, 200$). Although in this simple simulation, one could use registration to align the subject-specific functions, this would not be the case in a slightly more complex example, for instance, if we simulated a random number of "blips" for each subject.

To implement HARK for this example, we take $\theta_i = (\beta_{0i}, \tau_i^2, M_i, \bar{\gamma}_i, \bar{\mu}_i, \bar{\sigma}_i^2)$, as suggested in Section 3. We also use a Gaussian kernel, which is not identical to the kernel used

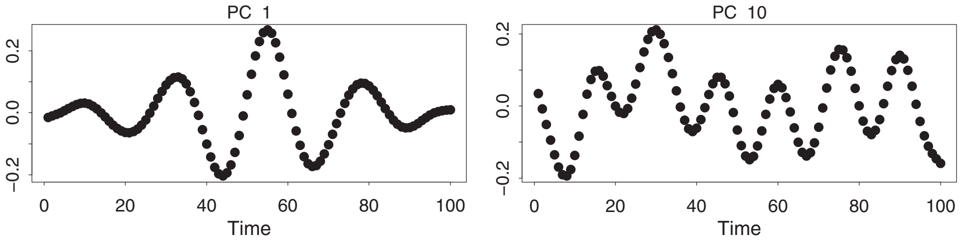

Figure 4. Loadings for two principal components of the simulated functional data.

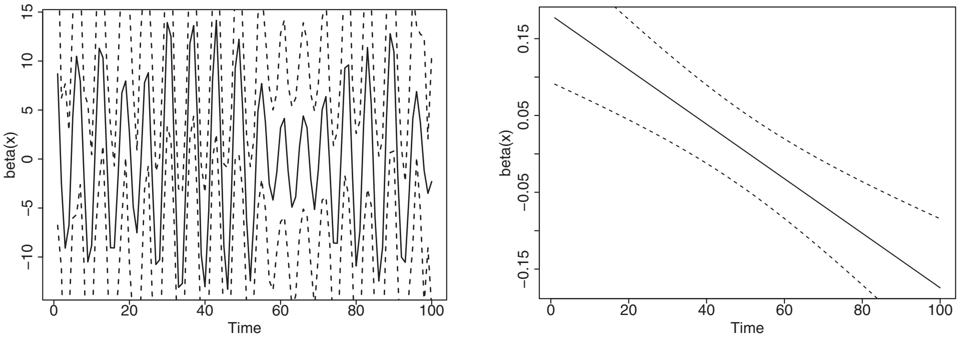

Figure 5. Estimated coefficient function $\beta(x)$ (solid curves) and pointwise 95% confidence bounds (dotted curves) for PFR on the simulated data. Left: without smoothing. Right: with smoothing (note change in the y-axis scale).

to generate the data but can represent it accurately. To do model selection, we use the MAP estimate $\hat{\zeta}$ of the vector ζ, as described in Section 3.

The true model for this example includes the predictor $\theta_{i4} = \bar{\gamma}_i$ and omits θ_{ij} for $j \neq 4$. The true function $g_4(\theta_{i4})$ is $g_4(\theta_{i4}) = \theta_{i4}$. The HARK results are summarized in Table 1, for the different sample sizes, with 10 simulated datasets for each sample size. Column 2 gives the percentage of simulated datasets for which HARK estimates $\sum_{p=1}^{P} \zeta_{4p} \neq 0$, that is, for which the HARK estimated model correctly includes the predictor θ_{i4}. Column 3 gives the percentage of simulations in which HARK incorrectly estimates $\sum_{p=1}^{P} \zeta_{jp} \neq 0$ for some other predictor $j \neq 4$ (a false positive). Column 4 reports the root mean squared error (RMSE) of the estimator of g_4, conditioning on the true model and averaging the RMSE over the domain of g_4. The percentage of that domain for which the 95% posterior bands for g_4 cover the true value is given in column 5. The width of the posterior bands for g_4, averaged over the domain, is given in column 6. The values in columns 4–6 are means over the simulations, with standard deviations shown in parentheses.

We see that the model selection accuracy is remarkably good: for sample sizes over 100, the estimated model is equal to the true model for all simulated datasets, while for the smaller sample sizes, there are some cases in which θ_{i4} is not included in the model, and only a single case in which other predictors are incorrectly included in the model. The function g_4 is also estimated accurately; the RMSE of its estimator decreases quickly as the sample size increases, and is small for all sample sizes [<2; compare with the range

Table 1. Results of HARK in the first simulation study

Sample size	% simulations with g_4 in model	% simulations with g_j in model, $j \neq 4$	RMSE of $\hat{g}_4$	Coverage of g_4 95% bands	Avg. width of g_4 95% bands
50	40	10	1.51 (0.49)	94.9 (8.2)	6.71 (1.03)
100	90	0	1.20 (0.28)	94.4 (7.8)	4.93 (0.61)
200	100	0	0.93 (0.09)	96.2 (6.1)	3.63 (0.28)
500	100	0	0.58 (0.13)	95.6 (6.9)	2.16 (0.12)

NOTE: Standard deviations across simulated datasets are shown in parentheses.

of the function g_4, which is the interval $(10, 20)$]. The coverage of the posterior bands is close to 95% for all sample sizes. The width of those bands drops quickly as the sample size increases, to a value of just 2.16 when $n = 500$.

In summary, the HARK model is able to detect the true relationship between the functional predictor and the outcome, even at the smallest sample sizes and with a very low false-positive rate. The posterior mean estimate of g_4 is typically close to the true value of g_4, and gets closer as the sample size increases. One would conclude from the HARK results that the outcome Y_i is positively associated with the magnitude of the mixture components of the predictor function f_i, which is accurate.

These conclusions are not sensitive to the choice of smoothing parameter in the empirical Bayes prior specification (Appendices A and B in the supplementary materials). Multiplying or dividing the smoothing parameter by a factor of 2 does not change any of the qualitative conclusions. The quantitative results after multiplying the smoothing parameter by 2 are nearly indistinguishable from the original results; the only change in terms of model selection results is that there are two false positives instead of one at the smallest sample size, and that for the sample size $n = 100$, the model selection accuracy increases to 100%. The other numerical results are nearly unchanged. When the smoothing parameter is divided by 2, the model selection results are again not much affected; the only change being that the first entry of column 2 in Table 1 decreases from 40% to 30%. The RMSE and the width of the posterior bands do increase noticeably; the RMSE increases by 26.7%, on average, across the simulations and sample sizes, while the width of the posterior bands increases by 10.2%, on average.

5.2 Simulating From the Functional Linear Model

Next, we generate data according to the functional linear model example of Goldsmith et al. (2011). The predictor functions $f_i(\cdot)$, noisy functional observations $W_i(x_k)$, and scalar outcomes Y_i are generated as follows, where $x_k = \frac{k}{10} : k = 0, \ldots, 100$ on the interval $[0, 10]$:

$$f_i(x_k) = u_{i1} + u_{i2}x_k + \sum_{\ell=1}^{10}\left[v_{i\ell 1}\sin\left(\frac{2\pi\ell}{10}x_k\right) + v_{i\ell 2}\cos\left(\frac{2\pi\ell}{10}x_k\right)\right],$$

$$W_i(x_k) \sim N(f_i(x_k), 1.0), \quad Y_i = \frac{1}{101}\sum_{k=0}^{100} f_i(x_k)\beta(x_k) + \varepsilon_i, \quad i = 1, \ldots, n.$$

Here, $\varepsilon_i \sim N(0, 0.25)$, $u_{i1} \sim N(0, 25)$, $u_{i2} \sim N(0, 0.04)$, and $v_{i\ell 1}, v_{i\ell 2} \sim N(0, 1/\ell^2)$. Results for PFR for this example were given by Goldsmith et al. (2011), where PFR is shown to recover the true $\beta(x)$ coefficient function accurately and efficiently. To implement HARK for this example, we again take $\theta_i = (\beta_{0i}, \tau_i^2, M_i, \bar{\gamma}_i, \bar{\mu}_i, \bar{\sigma}_i^2)$ and use a Gaussian kernel since it is appropriate to represent the smoothly varying predictor functions. For each of the sample sizes $n = 50, 100, 200, 500$ and a single simulation, HARK finds no significant relationship between the functional predictor and the outcome; repeating the simulation yields the same result. So, HARK is unable to recover the function $\beta(x)$, since it assumes a different model that is unrelated to the functional linear model.

6. APPLICATION TO THE SLEEP HEART HEALTH STUDY

6.1 Background

The Sleep Heart Health Study (SHHS) is a landmark study of sleep and its impacts on health outcomes. A detailed description of the SHHS can be found in Quan et al. (1997), Di et al. (2009), and Crainiceanu, Staicu, and Di (2009). Between 1995 and 1997, a sample of 6441 participants was recruited; participants less than 65 years of age were oversampled on self-reported snoring to augment the prevalence of sleep-disordered breathing (a condition where the airway of the throat collapses, triggering an arousal). In addition to the in-home polysomnogram (PSG), data on sleep habits, blood pressure, anthropometrics, medication use, daytime sleep tendency, and quality of life were collected. A PSG is a quasi-continuous multichannel recording of physiological signals acquired during sleep, which includes two surface electroencephalograms (EEGs).

It is of interest to understand how physiological characteristics may be related to sleep patterns, as measured using the EEG data. We focus on two physiological characteristics: RDI and body mass index (BMI). The RDI, or apnea/hypopnea index, is a measure of sleep-disordered breathing. The methods currently in use for relating physiological outcomes to the EEG data in the SHHS are mainly based on PC regression and penalized splines (Di et al. 2009; Crainiceanu et al. 2009; Crainiceanu, Staicu, and Di 2009).

We will relate the physiological characteristics to the time series of normalized δ-power, an indicator of slow neuronal firing that is a summary of the EEG signal. The δ-power time series is measured from sleep onset, and so, is initially synchronized across patients. It tends to go up for all subjects in the first 30–45 min of sleep; this corresponds to a dominance of slow-wave brain firing, characterizing the period immediately following sleep onset. As the night progresses, subjects go through sleep cycles, whose length, size, and number may vary across the population. Thus, subject δ-power patterns and cycles may become desynchronized in time across subjects.

The number or magnitude of fluctuations in the time series may have physiological importance, and may be related to the outcomes. HARK is designed to capture this type of variability; while traditional approaches regress the outcome Y_i on $f_i(x)$ for each x, HARK regresses Y_i on the parameters of kernels that can occur at variable locations. For this reason, the functions $f_i(\cdot)$ do not have to be aligned across subjects when applying HARK, that is, at any fixed time x, the subjects can be in different parts of their sleep cycle.

For each subject, we compute the normalized δ-power as described in Crainiceanu et al. (2009), aggregating at the 1-min level. Figure 1 shows the resulting time series for four subjects, along with smoothed curves obtained by penalized splines.

6.2 Application of HARK

Next, we use the SHHS data to relate sleep patterns, as measured by the EEG time series, to the RDI and BMI. We find that HARK represents the functional data both accurately and parsimoniously, and detects important and previously undescribed relationships between the sleep EEG data and both the RDI and the BMI.

The δ-power (EEG) series are defined on a common time domain $\mathcal{X}_i = \mathcal{X}$ (the function domain is the first 4 hr of sleep; we make $\mathcal{X}$ slightly larger than this interval when applying

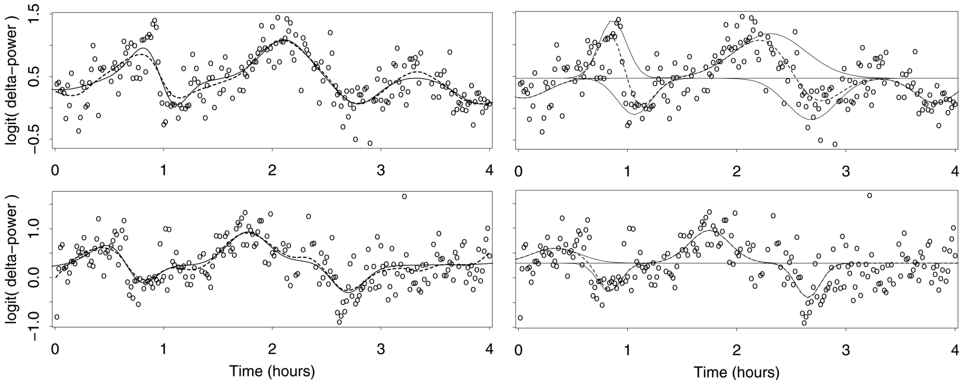

Figure 6. Left panels: HARK (solid curve) and penalized spline (dashed curve) functional estimates of the EEG δ-power signals for two randomly selected subjects. Right panels: mean line β_{0i} and mixture components (solid curves) from a single HARK posterior sample, for the same two subjects. The function f_i for the same posterior sample is also shown (dashed curve).

HARK to mitigate edge effects). The (normalized) δ-power has a range of 0 to 1, so we take the observations $W_i(x_{ik})$ to be the logit transformation of δ-power.

HARK is applied as described in Sections 2–4, using a Gaussian kernel and taking $\theta_i = (\beta_{0i}, \tau_i^2, M_i, \bar{\gamma}_i, \bar{\mu}_i, \bar{\sigma}_i^2)$. The resulting posterior mean estimate of the function f_i is shown in the left panels of Figure 6 for two randomly selected subjects i. The estimates are similar to penalized spline estimates of the same time series, also shown. They sometimes differ substantially near the start and end of the time series, where the difference is due to the edge effects of the two estimators.

In addition to yielding similar functional estimates to other methods, the HARK functional representation is parsimonious. The right panels of Figure 6 show the function representation from a single posterior sample, for the same two subjects. The horizontal line shows the mean $\beta_{0i}^{(\ell)}$ for this posterior sample ℓ, and the mixture components are shown deviating from this mean line. The function $f_i^{(\ell)}$ for this posterior sample is shown as a dashed curve; it is simply the sum of $\beta_{0i}^{(\ell)}$ and the mixture components, this sum being expressed in (1). The function is estimated using few mixture components; for the (randomly selected) posterior sample shown in the figure, $M_i^{(\ell)}$ is equal to 6 and 4, respectively, for the two subjects. The total number of parameters in the representation of f_i is $1 + 3M_i$ (each mixture component has three parameters), so the number of parameters in this posterior sample is 19 and 13, respectively, for the two subjects.

For the outcome variable log BMI, the model selected by HARK includes the single predictor β_{0i}; for the outcome variable log RDI, the model selected includes the three predictor variables β_{0i}, M_i, and $\bar{\gamma}_i$. The estimated functions g_j relating these predictors to the outcome variables are shown in Figure 7, along with 95% posterior bands capturing the uncertainty in these functions. Since M_i is an integer-valued variable, the function g_3 relating it to log RDI is shown only at the discrete values where M_i is defined. In these plots, we show the functions on the domain defined by the 0.025 and 0.975 quantiles of the predictor values (more precisely, on point estimates of those predictor values taken to be their posterior mean, given $\{W_{ik}\}_{k=1}^{K_i}$). Outside of this region, there is less information

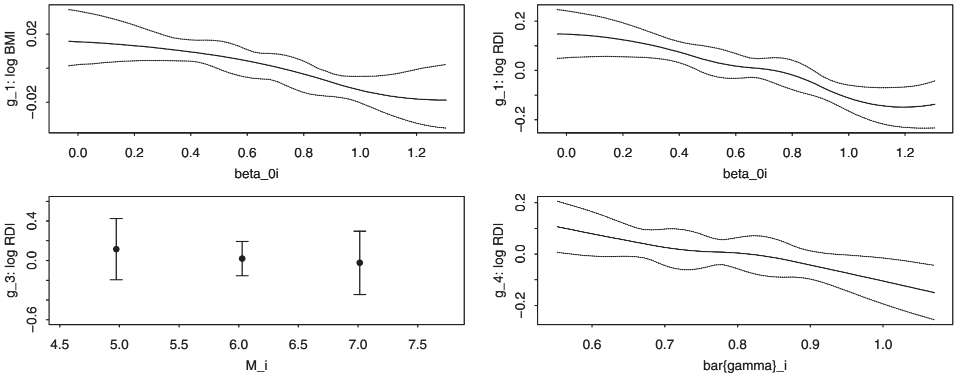

Figure 7. Estimated functions g_j relating the sleep study outcome variables to the predictors. Top left: the function relating log BMI to the predictor β_{0i}; clockwise from top right: the functions relating log RDI to the predictors β_{0i}, $\bar{\gamma}_i$, and M_i, respectively.

in the data regarding the functions g_j, so the posterior bands are wide and the function estimates are primarily driven by modeling choices rather than by the data.

The estimated functions are smooth, while there is some nonsmoothness of the posterior bands due to having fixed the knot locations. Placing a prior distribution on the knot locations would create smoother posterior bands, at the cost of some additional computation time.

The predictor β_{0i} measures the average logit δ-power; Figure 7 shows that this is negatively associated with the BMI. Additionally, the RDI is negatively associated with the average logit δ-power, the number of mixture components M_i, and the average magnitude $\bar{\gamma}_i$ of those mixture components. Since the kernel form is Gaussian, the mixture components represent bumps or dips in the δ-power series; this means that subjects with higher RDI tend to have fewer and less pronounced fluctuations in δ-power, a measure of slow neuronal firing. This contrasts with the fact such individuals are known to have more transitions between sleep states (Swihart 2009). These results are not contradictory, in part because transitions between sleep states occur at a shorter time scale than the δ-power fluctuations.

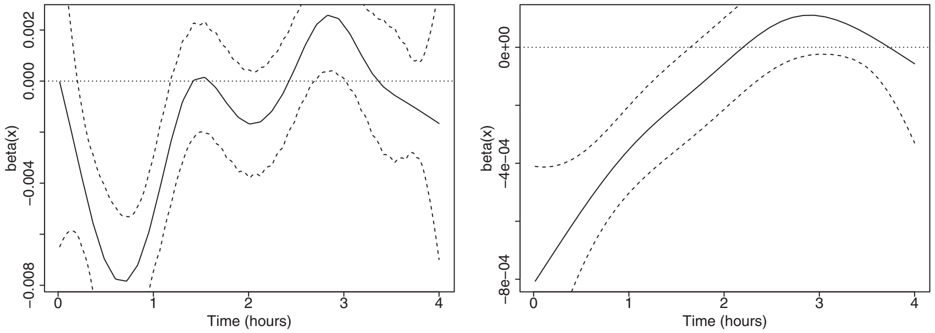

Figure 8. Estimated coefficient function $\beta(x)$ (solid curves) and pointwise 95% confidence bounds (dotted curves) for PFR on the sleep data. Left: log RDI outcome variable. Right: log BMI outcome variable.

6.3 Application of PFR

For comparison, we apply PFR to the sleep data. The estimated coefficient function $\beta(\cdot)$ is shown in Figure 8, for the two outcomes log RDI and log BMI. Both log RDI and log BMI are negatively associated with δ-power during the first 2 hr; after 2 hr, there is no statistically significant relationship between the δ-power and the outcomes. The δ-power signals are less synchronized after 2 hr, attenuating the predictive power of PFR.

7. CONCLUSIONS

HARK provides a method for relating continuous outcomes to functional predictors, based on a nonparametric kernel mixture representation of the predictors. It is appropriate when one hypothesizes that the functional data may include features such as bumps or plateaus occurring at varying locations, and that the frequency and characteristics of those features may be related to the outcome. We introduced a novel two-stage computational method that can handle large numbers of subjects, and that propagates the uncertainty from the first into the second stage. The validity of this computational method does not rely on the specifics of the HARK model, and we are currently formulating our method for use in a large class of Bayesian models.

SUPPLEMENTARY MATERIALS

Appendices: Technical Appendices A–C. (webAppendix.pdf; pdf file)

Software: S-PLUS package "hark" to implement HARK, along with code files to run the simulation study. Tested for Linux; see readme.pdf in the base directory for instructions on installation and use. (hark.tar.gz, GNU zipped tar file)

ACKNOWLEDGMENTS

The authors thank the referees and Associate Editor for their thorough reading and excellent recommendations, and Jeff Goldsmith for providing software and assistance in using PFR. Drs Crainiceanu and Ruppert are supported by the grant R01NS060910 from the National Institute of Neurological Disorders and Stroke, and Dr Woodard is supported by the National Science Foundation grant CMMI-0926814. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke or the National Institutes of Health.

[Received July 2011. Revised May 2012.]

REFERENCES

Best, N. G., Ickstadt, K., and Wolpert, R. L. (2000), "Spatial Poisson Regression for Health and Exposure Data Measured at Disparate Resolutions," *Journal of the American Statistical Association*, 95, 1076–1088. [4,8]

Bigelow, J. L., and Dunson, D. B. (2007), "Bayesian Adaptive Regression Splines for Hierarchical Data," *Biometrics*, 63, 724–732. [5,8]

——— (2009), "Bayesian Semiparametric Joint Models for Functional Predictors," *Journal of the American Statistical Association*, 104, 26–36. [3,10]

Cardot, H., Ferraty, F., and Sarda, P. (2003), "Spline Estimators for the Functional Linear Model," *Statistica Sinica*, 13, 571–591. [3,4,14]

Cardot, H., and Sarda, P. (2005), "Estimation in Generalized Linear Models for Functional Data via Penalized Likelihood," *Journal of Multivariate Analysis*, 92, 24–41. [14]

Carlin, B. P., and Louis, T. A. (2008), *Bayesian Methods for Data Analysis* (3rd ed.), Boca Raton, FL: Chapman & Hall. [8]

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models* (2nd ed.), Boca Raton, FL: Chapman & Hall. [3]

Chappell, D. (2010), *Introducing the Windows Azure Platform*, San Francisco, CA: David Chappell and Associates. Available at *http://www.windowsazure.com/en-us/develop/net/fundamentals/intro-to-windows-azure/* [13]

Clyde, M. A., House, L. L., and Wolpert, R. L. (2006), "Nonparametric Models for Proteomic Peak Identification and Quantification," in *Bayesian Inference for Gene Expression and Proteomics*, eds. K. A. Do, P. Muller, and M. Vannucci, Cambridge: Cambridge University Press, pp. 293–308. [3,5,7,13]

Clyde, M. A., and Wolpert, R. L. (2007), "Nonparametric Function Estimation Using Overcomplete Dictionaries," in *Bayesian Statistics 8*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: Oxford University Press, pp. 1–24. [5]

Crainiceanu, C. M., Caffo, B., Di, C., and Punjabi, N. M. (2009), "Nonparametric Signal Extraction and Measurement Error in the Analysis of Electroencephalographic Activity During Sleep," *Journal of the American Statistical Association*, 104, 541–555. [19]

Crainiceanu, C. M., and Reiss, P. (2011), *Refund: Regression With Functional Data*, R package version 0.1-5. Available at *http://cran.r-project.org/web/packages/refund/index.html* [3,14]

Crainiceanu, C. M., Staicu, A., and Di, C. (2009), "Generalized Multilevel Functional Regression," *Journal of the American Statistical Association*, 104, 1550–1561. [3,19]

de Boor, C. (2001), *A Practical Guide to Splines* (Rev. ed.), New York: Springer-Verlag. [9]

Dey, D. K., Ghosh, S. K., and Mallick, B. K. (eds.) (2000), *Generalized Linear Models: A Bayesian Perspective*, New York: Marcel Dekker. [9]

Di, C., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009), "Multilevel Functional Principal Component Analysis," *Annals of Applied Statistics*, 3, 458–488. [1,19]

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001), "Bayesian Curve-Fitting With Free-Knot Splines," *Biometrika*, 88, 1055–1071. [9,10]

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455. [6]

Dunson, D. B. (2010), "Multivariate Kernel Partition Process Mixtures," *Statistica Sinica*, 20, 1395–1422. [3]

Ferraty, F., and Vieu, P. (2004), "Nonparametric Models for Functional Data, With Application in Regression, Time-Series Prediction and Curve Discrimination," *Nonparametric Statistics*, 16, 111–125. [15]

Flegal, J. M., Haran, M., and Jones, G. L. (2008), "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?," *Statistical Science*, 23, 250–260. [13]

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409. [13]

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [10]

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 169–193. [13]

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996), *Markov Chain Monte Carlo in Practice*, New York: Chapman & Hall. [8]

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011), "Penalized Functional Regression," *Journal of Computational and Graphical Statistics*, 20, 830–851. [1,3,4,14,15,16,18]

Goutis, C., and Fearn, T. (1996), "Partial Least Squares Regression on Smooth Factors," *Journal of the American Statistical Association*, 91, 627–632. [4]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [13]

House, L. L. (2006), "Non-Parametric Bayesian Models in Expression Proteomic Applications," unpublished Ph.D. dissertation, Institute of Statistics and Decision Sciences, Duke University. [5]

House, L. L., Clyde, M. A., and Wolpert, R. L. (2010), "Bayesian Nonparametric Models for Peak Identification in MALDI-TOF Mass Spectroscopy," *Annals of Applied Statistics*, 5, 1488–1511. [5]

James, G. M. (2002), "Generalized Linear Models With Functional Predictors," *Journal of the Royal Statistical Society,* Series B, 64, 411–432. [4]

Lang, S., and Brezger, A. (2004), "Bayesian P-Splines," *Journal of Computational and Graphical Statistics*, 13, 183–212. [8,9]

Liu, F., Bayarri, M. J., and Berger, J. O. (2009), "Modularization in Bayesian Analysis, With Emphasis on Analysis of Computer Models," *Bayesian Analysis*, 4, 119–150. [3,10,11,12]

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009), "Combining MCMC With "Sequential" PKPD Modelling," *Journal of Pharmacokinetics and Pharmacodynamics*, 36, 19–38. [3,12]

MacLehose, R. F., and Dunson, D. B. (2009), "Nonparametric Bayes Kernel-Based Priors for Functional Data Analysis," *Statistica Sinica*, 19, 611–629. [5]

Marx, B. D., and Eilers, P. H. C. (1999), "Generalized Linear Regression on Sampled Signals and Curves: A *P*-Spline Approach," *Technometrics*, 41, 1–13. [4]

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010), "Cutting Feedback in Bayesian Regression Adjustment for the Propensity Score," *International Journal of Biostatistics*, 6, Article 16. [3,12]

Müller, H., and Stadtmüller, U. (2005), "Generalized Functional Linear Models," *The Annals of Statistics*, 33, 774–805. [4]

Osbourne, B. G., Fearn, T., Miller, A. R., and Douglas, S. (1984), "Application of Near Infrared Reflectance Spectroscopy to the Compositional Analysis of Biscuits and Biscuit Dough," *Journal of the Science of Food and Agriculture*, 35, 99–105. [1]

Pauler, D. K. (1998), "The Schwarz Criterion and Related Methods for Normal Linear Models," *Biometrika*, 85, 13–27. [10]

Pillai, N. (2008), "Lévy Random Measures: Posterior Consistency and Applications," Ph.D. dissertation, Durham, NC: Department of Statistical Science, Duke University. [5]

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., and Wahl, P. W. (1997), "The Sleep Heart Health Study: Design, Rationale, and Methods," *Sleep*, 20, 1077–1085. [1,19]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer. [1,2,4]

Reiss, P. T., and Ogden, R. T. (2007), "Functional Principal Component Regression and Functional Partial Least Squares," *Journal of the American Statistical Association*, 102, 984–996. [3,4]

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343. [10,14]

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003), *WinBUGS User Manual*, Available at *http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf* [12]

Swihart, B. J., Caffo, B., Crainiceanu, C. M., and Punjabi, N. M. (2009), "Modeling Multilevel Sleep Transitional Data via Poisson Log-Linear Multilevel Models," Technical report, COBRA Preprint Series, Article 64. Available at *http://works.bepress.com/ciprian_crainiceanu/32* [21]

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762. [13]

Wolpert, R. L., Clyde, M. A., and Tu, C. (2011), "Stochastic Expansions Using Continuous Dictionaries: Lévy Adaptive Regression Kernels," *The Annals of Statistics*, 39, 1916–1962. [5,7,8,13]

Woodard, D. B., Wolpert, R. L., and O'Connell, M. A. (2010), "Spatial Inference of Nitrate Concentrations in Groundwater," *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 209–227. [5,8]

Yao, F., and Müller, H. (2010), "Functional Quadratic Regression," *Biometrika*, 97, 49–64. [15]