Hierarchical Adaptive Regression Kernels for Regression with Functional Predictors

Dawn B. Woodard, Cornell University
Ciprian Crainiceanu, Johns Hopkins University
David Ruppert, Cornell University

Abstract

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 across subjects. We propose full 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, in order to better understand the relationship between sleep characteristics and health outcomes.

Keywords: Functional data analysis; nonparametric Bayes; kernel mixture; electroencephalogram; Lévy adaptive regression kernels; functional linear model.

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 (Quan et al. 1997) 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, Fearn, Miller and Douglas 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, Feder, Crainiceanu, Caffo and Reich 2010).

We introduce a new approach to regression with functional predictors. We represent the functional predictors using a parsimonious model designed for data that 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 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.

In order 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 (cf. 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 (Liu, Bayarri and Berger 2009).

The functional representation we utilize has been used previously for consistent nonparametric estimation of a *single function* under the name Lévy Adaptive Regression Kernels (LARK: Clyde, House, and Wolpert 2006; Wolpert, Clyde, and Tu 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; (3) introduce a method for efficient posterior computation for the joint functional

and exposure models, over the entire population.

We contrast our approach with state-of-the-art methods for regression with a functional predictor (Cardot, Ferraty and Sarda 2003; Reiss and Ogden 2007; Goldsmith et al. 2010). These methods assume that the functional predictor has a common domain across subjects (perhaps after domain warping), and that the outcome is linearly related to the function value $f_i(x)$ at each location x. We show that for simulated data that include features that occur at random locations, existing methods require a large amount of data to detect any relationship between the predictor and the outcome, and are unable to represent that relationship accurately for any sample size. By contrast, HARK can capture the relationship effectively even for small sample sizes (see Sections 3 and 7).

Our methodology is motivated by data from the Sleep Heart Health Study, and by interest in using these data to understand the relationship between sleep characteristics and health outcomes. We apply HARK towards this end, which requires fitting thousands of subject-specific curves, and handling missing data and complex variability patterns. We find that HARK provides a natural and effective solution. The representation of the functional data is both accurate and parsimonious. We find several important relationships, for instance that the frequency and magnitude of fluctuations in the EEG series is negatively correlated with the respiratory disturbance index of the subject.

We provide motivation in Section 2, by describing the sleep study, and in Section 3, by comparing HARK to existing methods using a simulation example. In Section 4 we introduce the model for the subject-specific functions, and in Section 5 we link the subject-level models hierarchically to a regression model. In Section 6 we describe our computational procedure. We provide more simulations in Section 7, and results for HARK on the sleep data in Section 8.

2. THE SLEEP HEART HEALTH STUDY

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, Crainiceanu, Caffo and Punjabi (2009), and Crainiceanu, Staicu and Di (2009). The SHHS is a multi-center cohort study that utilized the resources of existing, well-characterized, epidemiologic cohorts, and conducted further data collection, including measurements of sleep and breathing. Between 1995 and 1997, a sample of $6,441$ participants was recruited from the "parent" studies. Participants less than 65 years of age were over-sampled on self-reported snoring to augment the prevalence of sleep-disordered breathing (a condition where the airway of the throat collapses, triggering an arousal). Prevalent cardiovascular disease did not exclude potential participants and there was no upper age limit for enrollment.

In addition to the in-home polysomnogram (PSG), extensive data on sleep habits, blood pressure, anthropometrics, medication use, daytime sleep tendency, and quality of life (Medical Outcomes Study Short-form 36: SF-36, Ware and Sherbourne 1992) were collected. A PSG is a quasi-continuous multi-channel recording of physiological signals acquired during sleep that include 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 the physiological characteristics respiratory disturbance index and body mass index. The respiratory disturbance index, 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 Sleep Heart Health Study are mainly based on principal component regression and penalized splines (Di et al. 2009; Crainiceanu, Caffo, Di and Punjabi 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 minutes 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 the population. This type of behavior is hard or impossible to quantify using either principal components or standard smoothing techniques.

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, and does not require alignment of the subject-specific functions. We apply it to the sleep data in Section 8.

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

3. SIMULATION EXAMPLE: EXISTING METHODS VS. HARK

To make the benefits of HARK concrete, we use a simulated-data example to compare it to existing methods for regression with a functional predictor. Existing approaches, like HARK, summarize the functional data using a finite set of attributes that are then used as predictors in a regression model. For existing methods the most common choice of attributes is a set of coefficients from a linearly independent basis function representation of the predictor, such as principal component scores (cf. Müller and Stadtmüller 2005; Di, Crainiceanu, Caffo, and Punjabi 2009), spline coefficients (cf. James 2002), Fourier coefficients (cf. Ramsay and Silverman 2005), or partial least squares coefficients (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 et al. 2003).

The basis function approach, exemplified by the above citations, assumes that the func-

tional domain is common across subjects. In the case of a continuous outcome it also 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. Specifically, it takes the outcome to be 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 et al. 2003; Müller and Stadtmüller 2005). When the outcome is instead count-valued or has restricted domain the linearity assumption is made on some transformation of the expected response, and the framework is referred to as the *functional generalized linear model* (Cardot and Sarda 2005).

For data that exhibit features such as spikes that occur at random locations, application of the basis function approach means that many bases are needed to capture all possible occurrence locations of the features. As a result the power to detect a relationship between the characteristics of such features and the response can be low. Additionally, the functional linear model form can be inadequate to represent the predictor-outcome relationship, regardless of sample size.

To illustrate, we simulate data according to the following simple model, and compare HARK to principal component regression (PCR; cf. Cardot, Ferraty, and Sarda 1999; Di et al. 2009), using regularized estimates of the principal component loadings (Ramsay and Silverman 2005). Another state-of-the-art functional data approach is applied to the same example in Section 7.

For each of 5000 simulated subjects we generate a predictor on the (time) domain $\mathcal{X} = [1, 100]$; let $W_i(x_{ik})$ indicate the simulated predictor value for subject i at time x_{ik}. 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. I.e., we define

$$W_i(x_{ik}) = \beta_0 + \gamma_i \mathcal{K}(x_{ik}; \mu_i, \sigma^2) + \epsilon_{ik}$$

where $\epsilon_{ik} \sim N(0, \tau^2)$ and

$$\mathcal{K}(x; \mu, \sigma^2) = \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} - \exp\left\{-\frac{(x-(\mu+4))^2}{2\sigma^2}\right\}.$$

The amplitude of the "blip" is γ_i, which is sampled uniformly in the interval $(5, 20)$ for each i. To complete the specification, we take $\mu_i \overset{\text{iid}}{\sim} \text{Unif}(3, 93)$, $\sigma^2 = 4$, $\tau^2 = 2$, $\beta_0 = 0$, and $x_{ik} = k$ for $k = 1, \ldots, 100$. Simulated functions $W_i(x)$ are shown in Figure 2 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 $Y_i = \gamma_i$. We apply HARK and PCR for a range of sample sizes (the first $n = 20, 50, 100, 200, 500, 1000, 2000, 5000$ simulated subjects). HARK effectively captures the relationship between predictor and outcome, even for the smallest sample size. Representing the signal using a Gaussian kernel mixture (which can capture the spike and dip in $\mathcal{K}$), it finds that the average magnitude of the mixture components in the signal is positively correlated with the outcome. This relationship is found to be statistically significant using as few as $n = 20$ subjects. Details of the HARK analysis are left until Section 7.

For PCR, smoothed versions of the subject-specific functions are first obtained via penalized splines, and then the principal components of the smoothed functions are obtained. For all sample sizes we tried, 15 principal components (PCs) are required to capture 95% of the variability in the data (13 to capture 90%). These PCs are difficult to interpret; the loading functions of two PCs for the case $n = 1000$ are shown in Figure 3. For PCR we then regress the outcome Y_i on the scores of these first 15 PCs. For sample sizes up to $n = 1000$, none of the regression coefficients are significant at the $\alpha = .05$ level after Bonferroni correction.

For the largest sample sizes ($n = 2000$ and 5000), regression of Y_i on the PC scores yields one statistically significant regression coefficient. A plot of the loading function of this PC for the case $n = 5000$ is nearly identical to the right-hand plot in Figure 3. Although we have detected a relationship between the functional predictor and the outcome with this large sample size, the nature of this relationship is not clear. In particular, this analysis would

not lead to the conclusion that the outcome Y_i is positively correlated with the amplitude of a particular feature that occurs at variable time in the predictor.

4. FUNCTIONAL DATA MODEL

Next we describe the nonparametric functional representation we use. Consider that for each subject i we have noisy observations $\{W_i(x_{ik})\}_{k\in K_i}$ of a functional predictor $f_i(x)$ at locations x_{ik} in the domain $\mathcal{X}_i$. We utilize the following mixture form for the function $f_i(x)$:

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

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 the 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 $\beta_{0i}(x)$ would typically have a parametric form, such as a polynomial function of x; for simplicity we will 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).

The kernel mixture (1) can be viewed as representing the function via a basis expansion, where instead of orthogonal basis functions a larger, non-orthogonal dictionary of generating elements is used. Although the coefficients are no longer unique, one can obtain a much more parsimonious representation, by using fewer non-zero coefficients to attain the same accuracy. Sparsity is induced through the prior distribution; this effect is described in detail in 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 2006). LARK models have been applied to one-dimensional curve fitting and spatial and spatio-temporal modeling in Wolpert, Clyde and Tu (2006) and Woodard, Wolpert and O'Connell (2010). They have also been used for peak identification in mass spectroscopy (Clyde, House and Wolpert 2006; House, Clyde and Wolpert 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. We use a non-Lévy-process prior; the fact that we have a population of functions will allow us to utilize an empirical Bayes prior specification, and this prior specification will facilitate 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 8). One should choose the kernel form appropriately to the context; in the air pollution application of Wolpert et al. (2006), 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.

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 5.

4.1 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 4, by adding $N(0, 0.01)$ noise to the test function at 2048 equally-spaced locations in the domain. In order to evaluate whether LARK can recover the mixture components correctly, it was applied using kernels of the same form as in the test function, and the intercept $\beta_{0,i=1}$ was set equal to its true value of zero.

The LARK function estimate is shown in Figure 4, 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.

4.2 Likelihood

Next we specify the likelihood function, i.e. the probability density for the observations $\{W_i(x_{ik}) \approx f_i(x_{ik})\}_{k \in K_i}$ for each subject i. We use a normal error model $W_i(x_{ik}) \overset{\text{ind.}}{\sim} N\big(f_i(x_{ik}), \tau_i^2\big)$ for some variance parameter τ_i^2. This leads to the likelihood

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

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

4.3 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 4.2. One can obtain an empirical estimate of ω_i

for each subject i as described in Appendix A; the distribution of these estimates across subjects gives us a sense of what values of the parameters are reasonable, and will guide our prior specification. For instance, a Poisson prior distribution might be the obvious choice for a prior on $|M_i|$. Indeed, this is used in LARK prior specification since only a single functional observation is available so there is not enough information in the data to question this "default" 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 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 $\{|\hat{M}_i|\}_{i \in I}$.

With this choice the function f_i is C^∞ smooth, since the kernel (2) is C^∞ smooth and $|M_i| < \infty$ almost surely. Conditional on $|M_i|$, the γ_{im} values are assumed to be independently distributed according to a symmetric gamma distribution:

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

i.e. a gamma distribution for $|\gamma_{im}|$. Since this prior is symmetric about $\gamma = 0$, we have that $\mathsf{E}(f_i(x)) = \beta_{0i}$ for all x. The μ_{im} values are assumed to be independently uniformly distributed on the domain $\mathcal{X}_i$. The σ^2_{im} parameters are assumed to independently have an

inverse gamma distribution with shape and scale parameters $\alpha_\sigma > 0$ and $\rho_\sigma > 0$, i.e.

$$\pi(\sigma^2) = \frac{\rho_\sigma^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)}(\sigma^2)^{-(\alpha_\sigma+1)}e^{-\rho_\sigma/\sigma^2}.$$

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 an empirical Bayes approach (cf. Carlin and Louis 2008) as described in Appendix B.

5. 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, by assuming additivity of their effects. Take a vector of summary statistics of the functional representation, e.g. $\theta(\omega_i) = (1, \beta_{0i}, \tau_i^2, |M_i|, \bar{\gamma}_i, \bar{\sigma}_i^2)$, where $\bar{\gamma}_i = \mathbf{1}_{\{|M_i|>0\}} \sum_{m\in M_i} |\gamma_{im}|/|M_i|$ and $\bar{\sigma}_i^2 = \mathbf{1}_{\{|M_i|>0\}} \sum_{m\in M_i} \sigma_{im}^2/|M_i|$. We assume a linear regression model for the (continuous) outcome Y_i conditional on the parameter vector $\theta_i = \theta(\omega_i)$:

$$Y_i \overset{\text{ind.}}{\sim} N(\theta_i\eta, \phi^2) \tag{5}$$

where η is a regression coefficient vector and $\phi^2 > 0$ is the residual variance.

We use the default prior specification $\pi(\eta, \phi^2) \propto \phi^{-2}$. This can be considered to be noninformative, since it is uniform on η and $\log\phi$. Although this prior is nonintegrable, the

posterior distribution is proper as long as the number of subjects exceeds the length of the vector θ (Gelman et al. 2004, Sec. 14.2).

Having specified both the prior and 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

$$\begin{aligned}&\pi\left(\{\omega_i\}_{i\in I}, \eta, \phi^2 | \{W_{ik}, Y_i\}_{i\in I, k\in K_i}\right)\\ &\quad\propto \pi(\eta,\phi^2)\prod_{i\in I}\pi\left(\omega_i\right)\pi\left(\{W_{ik}\}_{k\in K_i}|\omega_i\right)\pi\left(Y_i|\omega_i,\eta,\phi^2\right). \end{aligned} \tag{6}$$

Here $\pi(\eta,\phi^2)\propto\phi^{-2}$, the quantity $\pi(\omega_i)$ is specified in Section 4.3, $\pi\left(\{W_{ik}\}_{k\in K_i}|\omega_i\right)$ is given in (3), and $\pi\left(Y_i|\omega_i,\eta,\phi^2\right)$ is specified by (5).

Estimation of any unknown quantity of interest is then based on the posterior distribution (6). For any function g of the parameters, one can obtain a point estimate (the posterior mean) or interval estimate (the $a/2$ and $1-a/2$ posterior quantiles for $a\in(0,0.5)$) of g. For example, we can obtain point and interval estimates of each regression coefficient η_j, or of $f_i(x)$ at any location x. Computation of the posterior mean of an arbitrary function g is described in the next section, and posterior quantiles of g can be computed in the same way.

6. 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 et al. 2009). This approximation avoids the potential computational pitfall of Bayesian inference for regression using a 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. Our approximation assumes that the functional data $\{W_i(x_{ik})\}_{k\in K_i}$ contain far more information about the function f_i

than does the outcome Y_i, so that we can ignore Y_i when estimating f_i. This two-stage approach permits the function estimation to be performed independently (and in parallel) across subjects.

We will show how to use Monte Carlo methods to compute the approximate posterior mean of any quantity g of interest. This will be done by constructing an ergodic Markov chain with limiting distribution equal to an approximation $\tilde{\pi}$ of the posterior. Such a Markov chain yields sample vectors $(\{\omega_i^{(\ell)}\}_{i\in I}, \eta^{(\ell)}, \phi^{2(\ell)})$ indexed by ℓ that approach $\tilde{\pi}$ in distribution. Then one can evaluate the approximate posterior mean $\mathsf{E}_{\tilde{\pi}}(g)$ using the ergodic average

$$\frac{1}{L}\sum_{\ell=1}^{L} g\left(\{\omega_i^{(\ell)}\}_{i\in I}, \eta^{(\ell)}, \phi^{2(\ell)}\right) \overset{L\to\infty}{\longrightarrow} \mathsf{E}_{\tilde{\pi}}\left[g\left(\{\omega_i\}_{i\in I}, \eta, \phi^2\right)\right]$$

(cf. Roberts and Rosenthal 2004).

We obtain our approximation $\tilde{\pi}$ by decomposing the joint posterior distribution (6) as

$$\begin{aligned}
&\pi\left(\{\omega_i\}_{i\in I}, \eta, \phi^2 | \{W_{ik}, Y_i\}_{i\in I, k\in K_i}\right) \\
=&\pi\left(\{\omega_i\}_{i\in I} | \{W_{ik}, Y_i\}_{i\in I, k\in K_i}\right)\ \pi\left(\eta, \phi^2 | \{\omega_i, W_{ik}, Y_i\}_{i\in I, k\in K_i}\right) \\
=&\pi\left(\{\omega_i\}_{i\in I} | \{W_{ik}, Y_i\}_{i\in I, k\in K_i}\right)\ \pi\left(\eta, \phi^2 | \{\omega_i, Y_i\}_{i\in I}\right)
\end{aligned}$$

and applying modularization:

$$\begin{aligned}
\pi\left(\{\omega_i\}_{i\in I} | \{W_{ik}, Y_i\}_{i\in I, k\in K_i}\right) \approx &\pi\left(\{\omega_i\}_{i\in I} | \{W_{ik}\}_{i\in I, k\in K_i}\right) \\
= &\prod_{i\in I} \pi\left(\omega_i | \{W_{ik}\}_{k\in K_i}\right).
\end{aligned}$$

The resulting approximate posterior distribution is denoted $\tilde{\pi}$.

This simplification allows two-stage computation:

Method for Approximate Posterior Simulation

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

2. Take the set of ℓ-indexed sample vectors $\{\omega_i^{(\ell)}\}_{i \in I}$ from Stage 1. For each ℓ, obtain a single sample of the parameters $(\eta^{(\ell)}, \phi^{2(\ell)})$ from their (closed-form) conditional posterior distribution $\pi\left(\eta, \phi^2 | \{\omega_i^{(\ell)}, Y_i\}_{i \in I}\right)$.

The resulting sample vectors $(\{\omega_i^{(\ell)}\}_{i \in I}, \eta^{(\ell)}, \phi^{2(\ell)})$ for $\ell = 1, \ldots, L$ thus are technically the iterations of an ergodic Markov chain on the joint space, with limiting distribution $\tilde{\pi}$ as desired. Stage 2 can also be replaced with a Rao-Blackwellization step, if the conditional expectation $\mathsf{E}[g\left(\{\omega_i\}_{i \in I}, \eta, \phi^2\right) | \{\omega_i, Y_i\}_{i \in I}]$ is available in closed form.

6.1 Stage 1 Computation

Clyde et al. (2006) and Wolpert et al. (2006) provide 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 $\pi\left(\omega_i | \{W_{ik}\}_{k \in K_i}\right)$ of our subject-specific functional model.

This approach uses the reversible jump variant (Green 1995) of the Metropolis-Hastings algorithm (cf. Tierney 1994). In each iteration of the Markov chain one of the parameters $\tau_i^2, \beta_{0i}, \{(\gamma_{im}, s_{im})\}_{m \in M_i}$ is updated or sampled from its conditional posterior distribution. An update of $\{(\gamma_{im}, s_{im})\}_{m \in M_i}$ can involve (a) a change in the magnitude γ_{im} or the parameters s_{im} of a single mixture component, (b) the addition or deletion of a mixture component, or (c) the merge of two components or split 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 evaluate convergence diagnostics and Monte Carlo standard error estimates for the

elements of the summary vector θ_i as defined in Section 5, 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 verify 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 non-geometric convergence, but leave formal proof of geometric ergodicity as future work.

6.2 Stage 2 Computation

The regression model (5) has a closed-form posterior distribution. Sampling from this distribution is done as described in Gelman, Carlin, Stern and Rubin (2004), Sec 14.2.

7. SIMULATION STUDY

We return to the example of Section 3, first giving the HARK results in more detail, then investigating the efficacy of another state-of-the-art functional data analysis approach (Goldsmith et al. 2010). To implement HARK for this example, we take $\theta_i = (1, \beta_{0i}, \tau_i^2, |M_i|, \bar{\gamma}_i, \bar{\sigma}_i^2)$ as suggested in Section 5; let the associated regression coefficient vector η be indexed from zero. Fitting the HARK model with $n = 20$ simulated subjects, the posterior estimate of the regression coefficient η_4 associated with $\bar{\gamma}_i$ is 0.94. The 95% posterior interval is $(0.82, 1.06)$, which is entirely above zero, leading to a conclusion of statistical significance.

The 95% posterior intervals for the regression coefficients η_j corresponding to β_{0i}, τ_i^2, $|M_i|$, and $\bar{\sigma}_i^2$ (as well as for the intercept η_0) all contain zero. We also performed a backward elimination procedure by repeatedly applying HARK, each time removing the least significant predictor until none of the 95% posterior intervals contain zero. This yields the model having

only η_4.

The conclusion is that the outcome Y_i is positively correlated with the magnitude of the mixture components of the predictor function f_i, and that the regression coefficient associated with this relationship is close to one. This conclusion is accurate, since in truth Y_i is exactly equal to the magnitude of those mixture components. The same results are obtained for larger sample sizes ($n > 20$), except that the point estimate of η_4 gets closer to one and the posterior interval for η_4 becomes narrower as n grows.

Next we apply the method of Goldsmith et al. (2010). This method represents the functional predictor using a much larger number of principal components than PCR, 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 et al. (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; (2) utilizing a connection to mixed effects models that provides a framework for generalization and a method for stable and efficient computation; (3) automatic selection of the smoothing parameter.

Although the method of Goldsmith et al. (2010) has been shown to work well for other types of simulated data 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 functions $f_i(x)$ are first estimated via regularized PCA. We regularize by smoothing the predictor functions before applying PCA; an alternative is to smooth the covariance matrix, which gives virtually identical results for this example. Then the outcome Y_i is regressed on the first K principal component scores for some large K. We take $K = 35$; this truncation level is suggested by Goldsmith et al. (2010), and captures 99.999% of the variability in the simulated functional data. Estimation in the regression model is performed by representing the function $\beta(x)$ by a power series spline basis and estimating it via penalized likelihood maximization.

For $n = 1000$ subjects, the principal component loading functions look similar to those in Figure 3 and are thus difficult to interpret. 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. The function has periodic behavior, and the confidence intervals all include zero.

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 point-wise 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. (2010) method are very similar for other sample sizes ($n = 200$, 500, 2000 and 5000).

8. RESULTS FOR THE SLEEP DATA

Next we use the Sleep Heart Health Study data to relate sleep patterns, as measured by the EEG time series, to respiratory disturbance index (RDI) and body mass index (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 RDI and BMI.

The δ power (EEG) series are defined on a common time domain $\mathcal{X}_i = \mathcal{X}$ (the function domain is the first four hours of sleep; we make $\mathcal{X}$ slightly larger than this interval when applying HARK in order to mitigate edge effects). The (normalized) δ power has a range of

zero to one, so we take the observations $W_i(x_{ik})$ to be the logit transformation of δ power.

We will apply HARK using the choice of summary vector $\theta_i = (1, \beta_{0i}, \tau_i^2, |M_i|, \bar{\gamma}_i, \bar{\sigma}_i^2)$ suggested in Section 5. However, we consider transforming the elements of this vector to ensure that the linear regression model (5) holds. To check these assumptions, and to explore the data, we start by performing a classical linear regression of the outcome on an empirical estimate of the vector θ_i. Empirical estimates of the elements of θ_i are obtained as described in Appendix A; call these estimates $\hat{\beta}_{0i}$, $\hat{\tau}_i^2$, $|\hat{M}_i|$, $\hat{\gamma}_i$, and $\hat{\sigma}_i^2$.

We find that log transformations of $\hat{\tau}_i^2$ and $\hat{\sigma}_i^2$ and a square root transformation of $\hat{\gamma}_i$, combined with log transformations of the outcomes RDI and BMI, are most appropriate to satisfy the assumptions of the linear regression model. Classical estimation for the regression model finds that $\hat{\beta}_{0i}$, $|\hat{M}_i|$, and $\hat{\gamma}_i^{1/2}$ are significant predictors of log RDI, and that the linear trends are negative in all three of these cases. For the log BMI outcome, $\hat{\beta}_{0i}$ and $\log \hat{\tau}_i^2$ are significant predictors, in both cases having a negative relationship to the outcome. Regression coefficient estimates are given in Table 1.

This informal procedure suggests important relationships between the EEG signal and the outcomes, and we turn to HARK for more formal results. HARK is applied as described in Sections 4-6. 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 models.

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 ex-

pressed 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 $1 + 3|M_i|$ (each mixture component has three parameters), so that the number of parameters in this posterior sample is 19 and 13, respectively, for the two subjects.

Overall the Gaussian kernel captures the characteristics of the δ-power signal well. However, when there is a sudden change in the δ-power, such as at times 1 and 2.5 in the top-right panel of Figure 6, the Gaussian kernel can only capture the sudden change by using two mixture components that overlap substantially. This could be addressed by using the more flexible asymmetric Gaussian kernel (cf. Kato, Omachi and Aso (2002)).

In the HARK model for each outcome, we perform a backward elimination procedure (repeatedly applying HARK, each time removing the least significant predictor until none of the 95% posterior intervals contain zero). The resulting hierarchical regression model for log BMI includes the intercept plus the predictors β_{0i} and $\log \tau_i^2$. The final model for log RDI includes the intercept plus the predictors β_{0i}, $|M_i|$ and $\bar{\gamma}_i^{1/2}$. Regression coefficient estimates and 95% posterior intervals are given in Table 2.

The predictor β_{0i} measures the average logit-δ-power; our analysis has shown that this is negatively associated with body mass index. We have also found that the residual variability τ_i^2 is negatively associated with BMI, i.e. that individuals with higher BMI tend to have less measurement error and/or small-scale fluctuations in their δ-power time series. This is not surprising since the measurement error in EEG is known to be affected by a number of physiological factors, including skull thickness and skin properties.

Our analysis has also found that respiratory disturbance index 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.

The HARK results (Table 2) agree with those from the regression model that uses an empirical estimate of θ_i (Table 1). The same set of significant predictors is found, and the signs of the coefficient estimates from HARK agree with those from the empirical regression. However, most of the HARK coefficient estimates are closer to zero than the coefficient estimates from the empirical regression.

The similarity of the results from HARK and the empirical regression is not due to the fact that empirical Bayes prior specification is used for the functional predictors in HARK. The prior on the functional representation ω_i is the same for every i (with the exception of the parameter β_{0i}). For instance, the prior on $|M_i|$ is the same for each i. If we had instead centered the prior for $|M_i|$ at the estimate $|\hat{M}_i|$, and assigned a small prior variance, then the results from HARK would match the empirical regression very closely, simply due to the prior specification. In fact, we do this for β_{0i} for computational reasons (Appendix B), probably accounting for the similarity of the regression coefficient estimates for β_{0i} between the two methods. However, for the other parts of the functional representation ω_i, it is the likelihood function and not the prior that leads to the similarity of the HARK results to those of the empirical regression.

9. 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.

The HARK model for continuous outcomes naturally generalizes to outcomes that are count or binary-valued or have restricted ranges, by utilizing the generalized linear model framework. However, the construction of an efficient computational method is more subtle in this case; in Stage 2 of the computation there is no longer a closed-form conditional posterior distribution. A Markov chain must be used instead, and the validity of the two-stage Markov chain computation would need to be established.

10. ACKNOWLEDGEMENTS

Dr.s Crainiceanu and Ruppert are supported by NIH grant R01NS060910. All authors thank Jeff Goldsmith for providing software and assistance in using the method of Goldsmith et al. (2010).

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.

Cardot, H., Ferraty, F., and Sarda, P. (1999), "Functional linear model," *Statistics & Probability Letters*, 45, 11–22.

Cardot, H., Ferraty, F., and Sarda, P. (2003), "Spline estimators for the functional linear model," *Statistica Sinica*, 13, 571–591.

Cardot, H., and Sarda, P. (2005), "Estimation in generalized linear models for functional data via penalized likelihood," *Journal of Multivariate Analysis*, 92, 24–41.

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

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 University Press, pp. 293–308.

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 University Press, Oxford, UK, pp. 1–24.

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.

Crainiceanu, C. M., Staicu, A., and Di, C. (2009), "Generalized multilevel functional regression," *Journal of the American Statistical Association*, 104, 1550–1561.

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.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.

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.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, 2nd edn, Boca Raton, FL: Chapman and Hall.

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 University Press, Oxford.

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

Goldsmith, J., Feder, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2010), "Penalized functional regression," Working Paper 2010-204, Johns Hopkins University Dept. of Biostatistics.

Goutis, C., and Fearn, T. (1996), "Partial least squares regression on smooth factors," *Journal of the American Statistical Association*, 91, 627–632.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

House, L. L., Clyde, M. A., and Wolpert, R. L. (2010), "Nonparametric models for peak identification and quantification in MALDI-TOF mass spectroscopy," Submitted.

James, G. M. (2002), "Generalized linear models with functional predictors," *Journal of the Royal Statistical Society, Series B*, 64, 411–432.

Kato, T., Omachi, S., and Aso, H. (2002), "Asymmetric Gaussian and its application to pattern recognition," in *Structural, Syntactic, and Statistical Pattern Recognition*, eds. T. Caelli, A. Amin, R. P. W. Duin, M. Kamel, and D. de Ridder, Springer-Verlag, Heidelberg, pp. 405–413.

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.

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.

Müller, H., and Stadtmüller, U. (2005), "Generalized functional linear models," *Annals of Statistics*, 33, 774–805.

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.

Pillai, N. (2008), "Lévy random measures: posterior consistency and applications," PhD thesis, Department of Statistical Science, Duke University, Durham, NC.

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.

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

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.

Roberts, G. O., and Rosenthal, J. S. (2004), "General state space Markov chains and MCMC algorithms," *Probability Surveys*, 1, 20–71.

Swihart, B. J. (2009), "Modeling multilevel sleep transitional data via Poisson log-linear multilevel models," Technical report, COBRA Preprint Series, Article 64.

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, 22, 1701–1762.

Ware, J. E., and Sherbourne, C. D. (1992), "The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection." *Medical Care*, 30, 473–483.

Wolpert, R. L., Clyde, M. A., and Tu, C. (2006), "Lévy Adaptive Regression Kernels," Discussion Paper 2006-08, Duke University, Dept. of Statistical Science. Revised, Feb. 2010.

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*, In press.

A. EMPIRICAL ESTIMATE OF THE KERNEL MIXTURE

Here we describe how to obtain an empirical estimate of the kernel mixture representation ω_i, and thus an estimate of the summary vector $\theta_i = \theta(\omega_i)$. The natural estimate of β_{0i} for a particular subject i is the average value of the functional predictor $W_i(x_{ik})$ over observations k; call this estimate $\hat{\beta}_{0i}$. Estimates of the other elements τ_i^2, $\{(\gamma_{im}, s_{im})\}_{m \in M_i}$ of the kernel mixture representation are obtained using a penalized spline fit $\hat{f}_i$ for the subject-specific function f_i. For instance, to obtain an estimate $\hat{\tau}_i^2$ of τ_i^2 we use the mean squared error of the residuals $[W_i(x_{ik}) - \hat{f}_i(x_{ik})]$.

Similarly, to find an estimate $|\hat{M}_i|$ of $|M_i|$ we count the number of local maxima of $\hat{f}_i$ that are above the mean $\hat{\beta}_{0i}$ and local minima that are below the mean. For each of these maxima

and minima, we estimate the magnitude γ_{im} of the mixture component by the height of the maximum/minimum minus $\hat{\beta}_{0i}$. We estimate σ_{im}^2 for the mixture component by finding the closest intersection of $\hat{f}_i$ with the mean line both before and after the maximum/minimum. The difference between the time of occurrence of these intersection points is roughly four times the standard deviation σ_{im} associated with the peak or dip.

B. SPECIFICATION OF PRIOR CONSTANTS

Here we specify the constants in the prior distribution of the functional data model (as defined in Section 4.3), using an empirical Bayes approach (cf. Carlin and Louis 2009). We will utilize an empirical estimate of the functional representation ω_i for each i as obtained in Appendix A.

We set the prior mean and variance of τ_i^2 to be the mean and variance of the empirical estimates $\hat{\tau}_{i'}^2$ over $i' \in I$. Similarly, in order to select the hyperparameters ρ and α for the prior distribution of $|\gamma_{im}|$, we use the empirical estimates of the $|\gamma_{i'm'}|$. We take the mean and standard deviation of these values over all subjects i' and all components m' to be the prior mean and standard deviation for $|\gamma_{im}|$. Also, to select the hyperparameters ρ_σ and α_σ for the inverse gamma prior for σ_{im}^2, we use the empirical estimates of the $\sigma_{i'm'}^2$ values. We use the mean and variance of these empirical values as the prior mean and variance for σ_{im}^2.

One might consider specifying the prior mean and variance of the parameter β_{0i} in the analogous fashion. However, this tends to yield a multimodal posterior distribution for β_{0i}, due to the fact that there are often multiple ranges of β_{0i} that are consistent with the data. Specifically, β_{0i} may be close to the minimum value of the observed time series, and all of the γ_{im} may be positive; alternatively, β_{0i} may be close to the maximum value of the observed time series, and all of the γ_{im} may be negative; or, β_{0i} may take some intermediate value and there may be some positive and some negative values of γ_{im}.

The presence of multiple reasonable hypotheses does not invalidate posterior inferences;

however, it does cause relatively slow mixing of the Markov chain, since switching between these hypotheses happens infrequently. Since we have a large number of Markov chains to simulate, we ensure efficiency of each chain by putting an informative prior on β_{0i}, effectively giving high prior weight to the last of the three hypotheses above. We take the prior mean of β_{0i} to be equal to the empirical estimate $\hat{\beta}_{0i}$, and obtain the prior standard deviation of β_{0i} as follows. We calculate the standard deviation SD_{tot} of $\{W_{i'}(x_{i'k})\}_{i' \in I, k \in K_i}$. If we used SD_{tot} as the prior standard deviation of β_{0i} we would essentially be allowing β_{0i} to take values within several standard deviations of $\hat{\beta}_{0i}$, giving high prior weight to all three of the hypotheses above. In order to put most of the prior weight on the third hypothesis, we divide SD_{tot} by 10, meaning that much of the prior weight is on values of β_{0i} in the middle tenth of the range of $\{W_i(x_{ik})\}_{k \in K_i}$. This choice yields fast mixing of the resulting Markov chains while still allowing the data to inform the posterior estimate of β_{0i}.

Outcome	Predictor	Coef. Est.	95% Confidence Int.
log(RDI $+0.5$)	$\hat{\beta}_{0i}$	-0.236	(-0.328,-0.143)
	$\lvert\hat{M}_i\rvert$	-0.067	(-0.096,-0.038)
	$\hat{\gamma}_i^{1/2}$	-0.927	(-1.223,-0.631)
log BMI	$\hat{\beta}_{0i}$	-0.025	(-0.039, -0.011)
	$\log \hat{\tau}_i^2$	-0.058	(-0.089,-0.027)

Table 1: Coefficient estimates from regression models for RDI and BMI, where the predictors are empirical estimates of the functional summaries.

Outcome	Predictor	Coef. Est.	95% Posterior Int.
log(RDI $+0.5$)	β_{0i}	-0.210	(-0.304, -0.117)
	$\lvert M_i\rvert$	-0.058	(-0.096, -0.020)
	$\bar{\gamma}_i^{1/2}$	-0.835	(-1.279, -0.401)
log BMI	β_{0i}	-0.026	(-0.039,-0.012)
	$\log \tau_i^2$	-0.041	(-0.073,-0.009)

Table 2: Regression coefficient estimates in the HARK models for RDI and BMI.

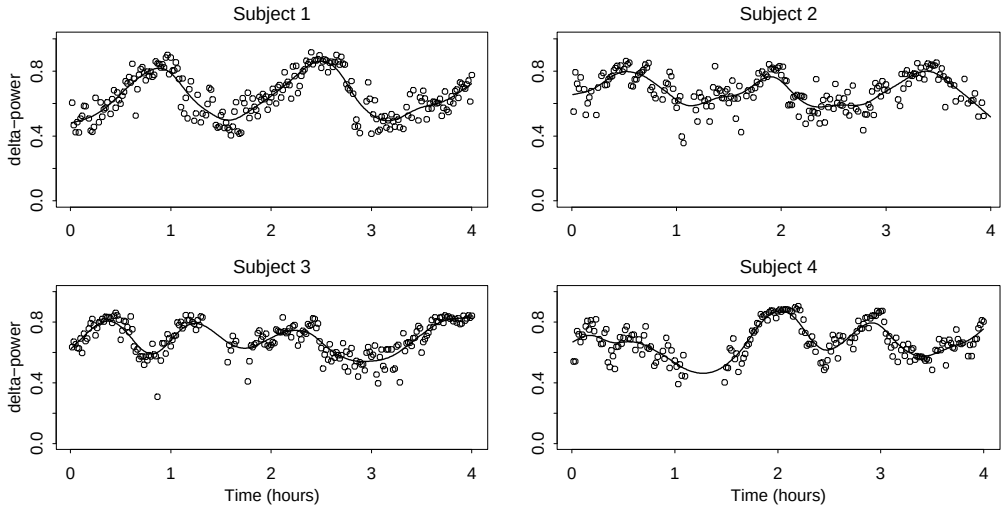

Figure 1: EEG (δ power) series for four subjects, with penalized spline approximations.

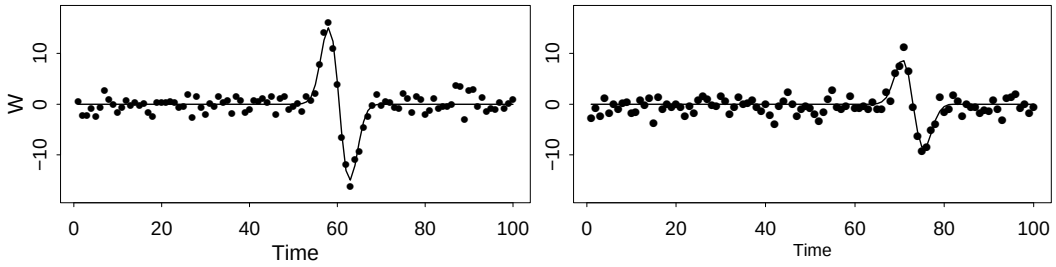

Figure 2: Simulated functional predictor for two subjects.

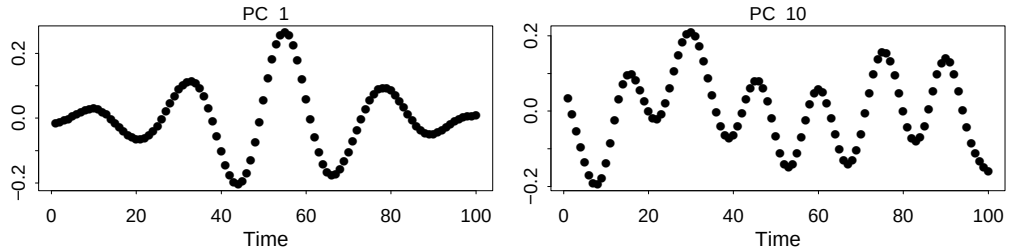

Figure 3: Loadings for two principal components of the simulated functional data.

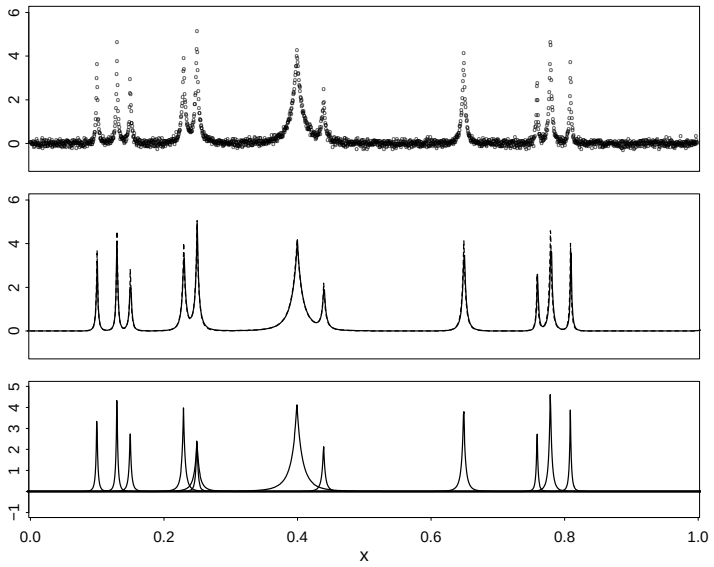

Figure 4: Estimation of a test function. Top: Simulated data. Middle: Test function (dashed curve) and LARK estimate (solid curve) are indistinguishable. Bottom: LARK representation, given by the mixture components from a single posterior sample.

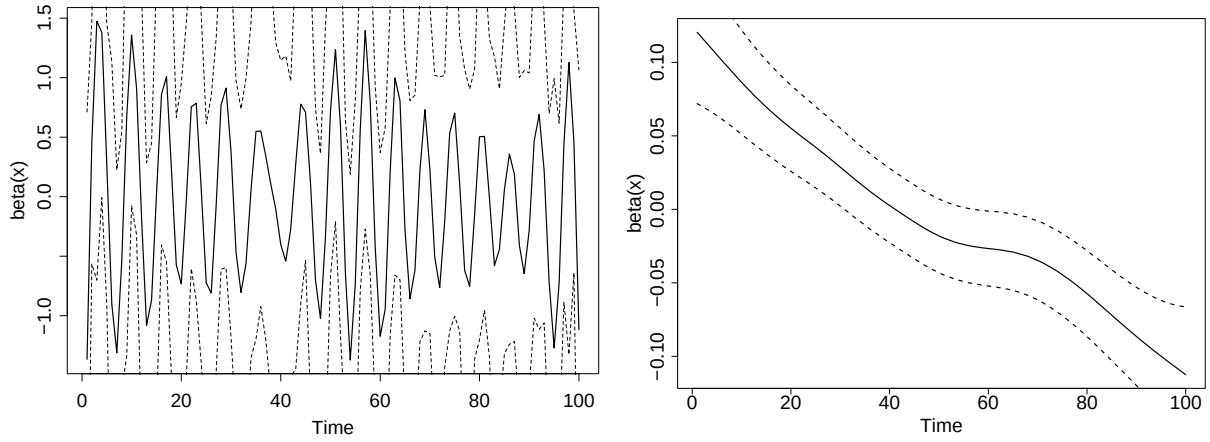

Figure 5: Estimated coefficient function $\beta(x)$ (solid curves) and point-wise 95% confidence bounds (dotted curves) for the simulated data. Left: without smoothing. Right: with smoothing (note change in y-axis scale).

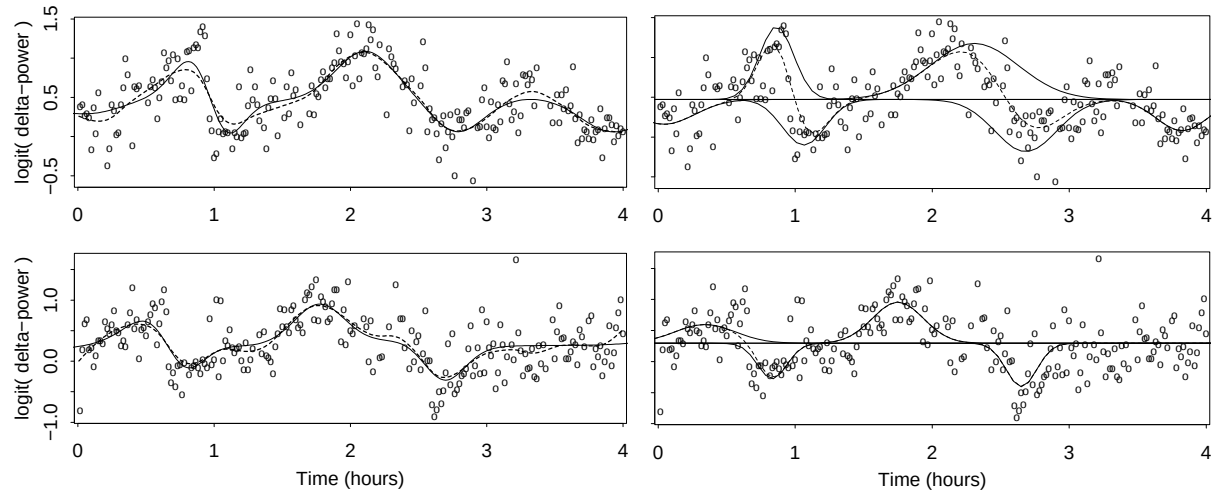

Figure 6: Left Panels: HARK (solid curve) and penalized spline (dashed curve) functional estimates of the EEG δ-power signals for two randomly selected subjects. Right Panels: Mean line β_{0i} and mixture components (solid curves) from a single posterior sample, for the same two subjects. The function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