Web Appendix for “Hierarchical Adaptive Regression Kernels for Regression with Functional Predictors” by D. B. Woodard, C. Crainiceanu, and D. Ruppert

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 $\theta_i = \theta(\omega_i)$, for each subject $i \in \{1, \ldots, n\}$. Here we assume the unnormalized Gaussian kernel (2).

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, $\{(\gamma_{im}, \mu_{im}, \sigma_{im}^2)\}_{m=1}^{M_i}$ of the kernel mixture representation are obtained using a thin plate spline fit $\hat{f}_i$ for the subject-specific function f_i, as described below. The smoothing parameter for the thin plate spline is obtained by applying restricted maximum likelihood estimation (Ruppert, Wand and Carroll 2003) to each subject, then taking the median of the estimated smoothing parameter across subjects to obtain a single value for the population; sensitivity to this choice is assessed in Section 5.1.

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 $\hat{\beta}_{0i}$. The location μ_{im} of the mixture component is taken to be the location of the local maximum/minimum. 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. When obtaining these estimates, we discard any peaks whose absolute height is less than $\varepsilon > 0$, or whose difference in height from the previous peak is less than ε, to avoid focusing on small peaks or small fluctuations in the signal. We take ε to be 1/5 of the difference between the

.05 and .95 quantile of all function observations $\{W_i(x_{ik})\}_{i,k}$.

B. SPECIFICATION OF PRIOR CONSTANTS

Here we specify the constants in the prior distribution of the functional data model (as defined in Section 2.3), using an empirical Bayes approach and assuming the Gaussian kernel (2). We will utilize the empirical estimate $\hat{\omega}_i = (\hat{\beta}_{0i}, \hat{\tau}_i^2, \{\hat{\gamma}_{im}, \hat{\mu}_{im}, \hat{\sigma}_{im}^2\}_{m=1}^{\hat{M}_i})$ 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 $i' \in \{1, \ldots, n\}$. Similarly, in order to select the hyperparameters ρ and α for the prior distribution of $|\gamma_{im}|$, we use the empirical estimates $\{|\hat{\gamma}_{i'm'}| : i' = 1, \ldots, n;\ m' = 1, \ldots, M_i\}$. We take the mean ("Mean$_\gamma$") and standard deviation ("SD$_\gamma$") of these values over all subjects i' and all components m' to be the prior mean and standard deviation for $|\gamma_{im}|$, i.e. taking $\rho = \text{Mean}_\gamma/\text{SD}_\gamma^2$ and $\alpha = \text{Mean}_\gamma^2/\text{SD}_\gamma^2$. As discussed in Section 2.3 it is desirable to have $\alpha > 1$; in practice α is typically well above 1, an effect that can be explained informally as follows. Recall from Appendix A that $|\hat{\gamma}_{i'm'}| \geq \frac{r}{5}$ where r is the difference between the .05 and .95 quantile of all function observations $\{W_i(x_{ik})\}_{i,k}$. Also, most of the $\hat{\gamma}_{i'm'}$ values satisfy $|\hat{\gamma}_{i'm'}| \leq \frac{r}{2}$, since the values $\hat{\gamma}_{i'm'}$ are obtained as deviations of the estimated function $\hat{f}_i$ from the mean $\hat{\beta}_{0i}$. For a set of values that fall in the interval $[\frac{r}{5}, \frac{r}{2}]$, the mean is certainly $\geq \frac{r}{5}$, and the standard deviation is $\leq \frac{1}{2}\left(\frac{r}{2} - \frac{r}{5}\right) < \frac{r}{5}$ by Lemma 2 of Audenaert (2010). So it is typically the case that $\text{Mean}_\gamma > \text{SD}_\gamma$, and thus that $\alpha > 1$.

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. We ensure efficiency of the Markov 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 $\{W_{i'}(x_{i'k})\}_{i',k}$. 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 $\{W_i(x_{ik})\}_{k=1}^{K_i}$. This choice yields fast mixing of the resulting Markov chains while still allowing the data to inform the posterior estimate of β_{0i}.

C. CONVERGENCE OF THE COMPUTATIONAL PROCEDURE

Here we show that for any $\xi > 0$ and any initialization $(\{\omega_i^{(0)}\}_{i=1}^n, \zeta^{(0)}, \eta^{(0)}, \psi^{(0)}, \phi^{2(0)})$, for all L large enough and all $N_1, \ldots, N_L$ large enough the total variation distance between $\tilde{\pi}$ (as defined in Equation (9) of the main paper) and the distribution of $(\{\omega_i^{(L)}\}_{i=1}^n, \zeta^{(L)}, \eta^{(L)}, \psi^{(L)}, \phi^{2(L)})$ is less than ξ. We require two regularity conditions.

The sample vectors $\{\omega_i^{(\ell)}\}_{i=1}^n$ in Stage 1 form the iterations of a single Markov chain with invariant density $\pi(\{\omega_i\}_{i=1}^n | \{W_{ik}\}_{i,k})$; call the transition kernel of this Markov chain Q_0. Denote by Q_ℓ the Markov transition kernel used in the ℓth step of Stage 2, having invariant density $\pi\left(\zeta, \eta, \psi, \phi^2 | \{\omega_i^{(\ell)}, Y_i\}_{i=1}^n\right)$. Denote by λ the reference measure with respect to which the density $\tilde{\pi}$ is defined. The first regularity condition is on the distribution of $(\{\omega_i^{(L)}\}_{i=1}^n, \zeta^{(L)}, \eta^{(L)}, \psi^{(L)}, \phi^{2(L)})$. This distribution depends only on the specification of $Q_0, Q_1, \ldots, Q_L$ and on the distribution of the initial values $(\{\omega_i^{(0)}\}_{i=1}^n, \zeta^{(0)}, \eta^{(0)}, \psi^{(0)}, \phi^{2(0)})$.

A1. Assume that for all L large enough the distribution of $(\{\omega_i^{(L)}\}_{i=1}^n, \zeta^{(L)}, \eta^{(L)}, \psi^{(L)}, \phi^{2(L)})$ has a density with respect to λ, denoted $\nu_L(\{\omega_i\}_{i=1}^n, \zeta, \eta, \psi, \phi^2)$.

It is straightforward to show that Assumption A1 holds for the transition kernels $Q_0, Q_1, \ldots, Q_L$ defined in Sections 4.1-4.2, if the initial values $(\{\omega_i^{(0)}\}_{i=1}^n, \zeta^{(0)}, \eta^{(0)}, \psi^{(0)}, \phi^{2(0)})$ are drawn from a distribution that has a density with respect to λ.

Let $\tilde{\pi}_\omega(\{\omega_i\}_{i=1}^n)$ and $\nu_{L,\omega}(\{\omega_i\}_{i=1}^n)$ indicate the marginal density of $\{\omega_i\}_{i=1}^n$ under $\tilde{\pi}$ and ν_L, respectively. The density $\nu_{L,\omega}(\{\omega_i\}_{i=1}^n)$ is the density of the random vector $\{\omega_i^{(L)}\}_{i=1}^n$, and $\{\omega_i^{(L)}\}_{i=1}^n$ is generated by L iterations of Q_0 applied to the initialization $\{\omega_i^{(0)}\}_{i=1}^n$. Our second regularity condition is on $\nu_{L,\omega}$, and thus on Q_0 and $\{\omega_i^{(0)}\}_{i=1}^n$.

A2. Assume that for all L large enough, the support of $\nu_{L,\omega}(\{\omega_i\}_{i=1}^n)$ is the same as the support of $\tilde{\pi}_\omega(\{\omega_i\}_{i=1}^n)$.

Assumption A2 holds for the choice of Q_0 defined in Section 4.1, regardless of $\{\omega_i^{(0)}\}_{i=1}^n$. This is due to the following fact, noticing that $\tilde{\pi}_\omega(\{\omega_i\}_{i=1}^n) = \prod_{i=1}^n \pi(\omega_i | \{W_{ik}\}_{k=1}^{K_i})$ is the invariant density of Q_0. Regardless of $\{\omega_i^{(0)}\}_{i=1}^n$, after several iterations of Q_0 we have $\nu_{L,\omega}(\{\omega_i\}_{i=1}^n) > 0$ for all values of $\{\omega_i\}_{i=1}^n$ in the support of the invariant density $\tilde{\pi}_\omega$ of Q_0.

We will also need Lemma C.1, which uses the $\mathcal{L}_1$ norm on functions, denoted by $\|\cdot\|_{\mathcal{L}_1}$.

Lemma C.1. *Consider two probability densities $\mu^1(x,y)$ and $\mu^2(x,y)$ defined on a product space $\mathcal{X} \times \mathcal{Y}$ with product measure $\lambda_X \times \lambda_Y$. Let $\mu_X^1(x) = \int \mu^1(x,y)\lambda_Y(dy)$ and $\mu_X^2(x) = \int \mu^2(x,y)\lambda_Y(dy)$. Also, for x such that $\mu_X^1(x) > 0$ let $\mu_{Y|x}^1(y) = \mu^1(x,y)/\mu_X^1(x)$, and define $\mu_{Y|x}^2(y)$ analogously. For any $\xi > 0$, if*

1. $\{x : \mu_X^1(x) > 0\} = \{x : \mu_X^2(x) > 0\}$

2. $\|\mu_X^1 - \mu_X^2\|_{\mathcal{L}_1} < \xi/2$

3. $\|\mu_{Y|x}^1 - \mu_{Y|x}^2\|_{\mathcal{L}_1} < \xi/2$ *for every x such that $\mu_X^1(x) > 0$*

then $\|\mu^1 - \mu^2\|_{\mathcal{L}_1} < \xi$.

Proof. We have that

$$\begin{aligned}
&\|\mu^1 - \mu^2\|_{\mathcal{L}_1} \\
&= \int \left|\mu^1(x,y) - \mu^2(x,y)\right| \lambda_X(dx)\,\lambda_Y(dy) \\
&= \int \left|\mu_X^1(x)\mu_{Y|x}^1(y) - \mu_X^1(x)\mu_{Y|x}^2(y) + \mu_X^1(x)\mu_{Y|x}^2(y) - \mu_X^2(x)\mu_{Y|x}^2(y)\right| \lambda_X(dx)\,\lambda_Y(dy) \\
&\le \int \left|\mu_X^1(x)\mu_{Y|x}^1(y) - \mu_X^1(x)\mu_{Y|x}^2(y)\right| \lambda_X(dx)\,\lambda_Y(dy) + \int \left|\mu_X^1(x)\mu_{Y|x}^2(y) - \mu_X^2(x)\mu_{Y|x}^2(y)\right| \lambda_X(dx)\,\lambda_Y(dy) \\
&= \int \mu_X^1(x) \int \left|\mu_{Y|x}^1(y) - \mu_{Y|x}^2(y)\right| \lambda_Y(dy)\,\lambda_X(dx) + \int \left|\mu_X^1(x) - \mu_X^2(x)\right| \int \mu_{Y|x}^2(y)\,\lambda_Y(dy)\,\lambda_X(dx) \\
&= \int \mu_X^1(x)\,\|\mu_{Y|x}^1 - \mu_{Y|x}^2\|_{\mathcal{L}_1}\,\lambda_X(dx) + \|\mu_X^1 - \mu_X^2\|_{\mathcal{L}_1} \qquad < \xi.
\end{aligned}$$

□

We will take L and $N_1, \ldots, N_{L-1}$ to be fixed values and N_L to be a function of the random variables $\{\omega_i^{(L)}\}_{i=1}^n$ and $(\zeta^{(L-1)}, \eta^{(L-1)}, \psi^{(L-1)}, \phi^{2(L-1)})$. Recall that Q_0 is an ergodic Markov chain with invariant density $\tilde{\pi}_\omega$ and that $\nu_{L,\omega}$ is the density of $\{\omega_i\}_{i=1}^n$ after L iterations of Q_0. So using Assumption A1, for all L large enough

$$\|\nu_{L,\omega} - \tilde{\pi}_\omega\|_{\mathcal{L}_1} < \xi/2. \tag{B.1}$$

This is given by results in, e.g., (Roberts and Rosenthal 2004), recalling that the $\mathcal{L}_1$ distance between probability densities is equal to the total variation distance between the corresponding probability measures.

Take $N_1, \ldots, N_{L-1}$ to be arbitrary fixed values ≥ 1. Recall that Q_L is an ergodic Markov chain with invariant density $\pi\left(\zeta, \eta, \psi, \phi^2 \middle| \{\omega_i, Y_i\}_{i=1}^n\right)$ where $\{\omega_i\}_{i=1}^n = \{\omega_i^{(L)}\}_{i=1}^n$. Define

$$\begin{aligned} \nu_{L,\cdot|\omega}\left(\zeta, \eta, \psi, \phi^2 \middle| \{\omega_i\}_{i=1}^n\right) &= \nu_L\left(\{\omega_i\}_{i=1}^n, \zeta, \eta, \psi, \phi^2\right) / \nu_{L,\omega}\left(\{\omega_i\}_{i=1}^n\right) \\ \tilde{\pi}_{\cdot|\omega}(\zeta, \eta, \psi, \phi^2 | \{\omega_i\}_{i=1}^n) &= \tilde{\pi}(\{\omega_i\}_{i=1}^n, \zeta, \eta, \psi, \phi^2)/\tilde{\pi}_\omega(\{\omega_i\}_{i=1}^n) \\ &= \pi\left(\zeta, \eta, \psi, \phi^2 \middle| \{\omega_i, Y_i\}_{i=1}^n\right). \end{aligned}$$

Since $\tilde{\pi}_{\cdot|\omega}$ is the invariant density of Q_L and $\nu_{L,\cdot|\omega}$ is the density of $(\zeta, \eta, \psi, \phi^2)$ after N_L steps of Q_L, for all N_L large enough

$$\|\nu_{L,\cdot|\omega} - \tilde{\pi}_{\cdot|\omega}\|_{\mathcal{L}_1} < \xi/2 \tag{B.2}$$

by the same argument as (B.1).

Combining (B.1)-(B.2) with Assumption A2 and Lemma C.1, we have that

$$\|\nu_L - \tilde{\pi}\|_{\mathcal{L}_1} < \xi.$$

Again using the fact that the total variation distance between two distributions is equal to the $\mathcal{L}_1$ distance between the corresponding densities, this gives the desired result.

REFERENCES

Audenaert, K. M. R. (2010), "Variance bounds, with an application to norm bounds for commutators," *Linear Algebra and its Applications*, 432, 1126–1143.

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

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