

Wavelet-Based Functional Mixed Models

Jeffrey S. Morris ¹

The University of Texas MD Anderson Cancer Center, Houston, USA

and Raymond J. Carroll

Texas A&M University, College Station, USA

Summary: Increasingly, scientific studies yield functional data, in which the ideal units of observation are curves and the observed data consist of sets of curves sampled on a fine grid. In this paper, we present new methodology that generalizes the linear mixed model to the functional mixed model framework, with model fitting done using a Bayesian wavelet-based approach. This method is flexible, allowing functions of arbitrary form and the full range of fixed effects structures and between-curve covariance structures available in the mixed model framework. It yields nonparametric estimates of the fixed and random effects functions as well as the various between-curve and within-curve covariance matrices. The functional fixed effects are adaptively regularized as a result of the nonlinear shrinkage prior imposed on the fixed effects' wavelet coefficients, and the random effect functions experience a form of adaptively regularization because of the separately estimated variance components for each wavelet coefficient. Because we have posterior samples for all model quantities, we can perform pointwise or joint Bayesian inference or prediction on the quantities of the model. The adaptiveness of the method makes it especially appropriate for modeling irregular functional data characterized by numerous local features like peaks.

Keywords: Bayesian methods; Functional data analysis; Mixed models; Model averaging; Non-parametric regression; Proteomics; Wavelets

Short title: Wavelet-Based Functional Mixed Models

1 Introduction

Technological innovations in science and medicine have resulted in a growing number of scientific studies that yield functional data. Here, we consider data to be functional if (1) the ideal units of observation are curves and (2) the observed data consist of sets of curves sampled on a fine grid. Ramsay and Silverman coined “functional data analysis” (FDA, 1997) as an inclusive term for the analysis of data for which the ideal units are curves. They stated that the common thread uniting these methods is that they must deal with both replication, or combining information across N curves, and regularity, or exploiting the smoothness to borrow strength between the

¹*Address for correspondence:* Jeffrey S. Morris, University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Box 447, Houston, TX 77030-4009, USA.

Email/Web page: jeffmo@odin.mdacc.tmc.edu, <http://biostatistics.mdanderson.org/Morris>

measurements within a curve. The key challenge in FDA is to find effective ways to deal with both of these issues simultaneously.

Much of the existing FDA literature deals with exploratory analyses, and more work developing methodology to perform inference is needed. The complexity and high dimensionality of these data make them challenging to model, since it is difficult to construct models that are reasonably flexible, yet feasible to fit. When the observed functions are well-represented by simple parametric forms, parametric mixed models (Laird and Ware, 1982) can be used to model the functions (see Verbeke and Mohlenberghs, 2000). When simple parametric forms are insufficient, however, nonparametric approaches allowing arbitrary functional forms must be considered. There are numerous papers in the recent literature applying kernels or fixed-knot splines to this problem of modelling replicated functional data (e.g. Rice and Silverman, 1991; Shi, Weiss and Taylor, 1996; Zhang, et al., 1998; Wang, 1998; Staniswallis and Lee, 1998; Brumback and Rice, 1998; Rice and Wu, 2001; Wu and Zhang, 2002; Guo, 2002; Liang, Wu and Carroll, 2003; Wu and Liang, 2004). Some of these models are very flexible, with many allowing different fixed effect functions of arbitrary form and some also allowing random effect functions to be of arbitrary form. Among the most flexible of these is Guo (2002), who introduced a functional mixed model allowing functional fixed and random effect functions of arbitrary form, with the modeling done using smoothing splines. All of these approaches are based on smoothing methods using global bandwidths and penalties, so are not well-suited for modelling irregular functional data characterized by spatial heterogeneity and local features like peaks.

This type of functional data is frequently encountered in scientific research, for example, in biomarker assessments on a spatial axis on colonic crypts (Grambsch, et al., 1995; Morris, et al., 2003a), in measurements of activity levels using accelerometers (Gortmaker, et al. 1999), and mass spectrometry proteomics (Morris, et al., 2005). Our main focus in this paper is modelling functions of this type. In existing literature, data like these are successfully modelled in the single function setting using kernels with local bandwidths or splines with free knots or adaptive penalties. However, it is not straightforward to generalize these approaches to the multiple function setting, since the positions of the local features may differ across curves. It is possible for the mean functions to be spiky but the curve-to-curve deviations smooth, the mean functions to be smooth but the curve-to-curve deviations spiky, or for both the mean functions and the curve-to-curve deviations to be spiky. This requires flexible and adaptive modelling of both the mean and covariance structure of the data.

Wavelet regression is an alternative method that can effectively model spatially heterogeneous data in the single function setting (e.g. Donoho and Johnstone 1995). Morris, et al. (2003a) extended these ideas to a specific multiple function setting – hierarchical functional data – which consists of functions observed in a strictly nested design. The fully Bayesian modelling approach yielded adaptively regularized estimates of the mean functions in the model, estimates of random effect functions, and posterior samples which could be used for Bayesian inference. However, the method presented in that paper has limitations that prevent its more general use. It can only model nested designs, hence cannot be used to model functional effects for continuous covariates, functional main and interaction effects for crossed factors, and cannot jointly model the effects of multiple covariates. Also, it cannot handle other between-curve correlation structures, such as serial correlation that might occur in functions sampled sequentially over time. Further, Morris, et al. (2003a) make restrictive assumptions on the curve-to-curve variation that do not accommodate nonstationarities commonly encountered in these type of functional data, such as different variances and different degrees of smoothness at different locations in the curve-to-curve deviations (See Figure 1). Finally, Morris, et al. (2003a) do not provide general-use code that could be used to analyze other data sets.

In this paper, we develop a unified Bayesian wavelet-based approach for the much more general functional mixed models framework. This framework accommodates any number of fixed and random effect functions of arbitrary form, so can be used for the broad range of mean and between-curve correlation structures available in the mixed model setting. The random effect distributions are allowed to vary over strata, allowing different groups of curves to differ with respect to both their mean functions and covariance surfaces. We also make much less restrictive assumptions on the form of the curve-to-curve variability that accommodate important types of nonstationarity, and results in more adaptively regularized representations of the random effect functions. As in Morris, et al. (2003a), we obtain posterior samples of all model quantities, which can be used to perform any desired Bayesian inference. We also present a completely data-based method for selecting the regularization parameters of the method, which allows the procedure to be applied without any subjective prior elicitation, if desired, and these regularization parameters are allowed to differ across fixed effect functions. The additional flexibilities we have built into the method presented in this paper has led to increased computational challenges, but we have tackled these and developed general-use code for implementing the method that is efficient enough to handle extremely large data sets. We make this code freely available on the web (link

at <http://biostatistics.mdanderson.org/Morris/papers.html>), so researchers need not write their own code to implement our method.

The remainder of the paper is organized as follows. In Section 2, we introduce wavelets and wavelet regression. In Section 3, we describe our functional mixed model framework. In Section 4, we describe the wavelet-based functional mixed models methodology, presenting the wavelet-space model, describing the covariance assumptions we make, and specifying prior distributions. In Section 5, we describe the Markov Chain Monte Carlo procedure we use to obtain posterior samples of our model quantities and explain how we use these for inference. In Section 6, we apply the method to an example functional data set, and in Section 7, we present a discussion of the method. Technical details and derivations are in an appendix.

2 Wavelets and Wavelet Regression

Wavelets are families of orthonormal basis functions that can be used to represent other functions parsimoniously. For example, in $L^2(\mathfrak{R})$, an orthogonal wavelet basis is obtained by dilating and translating a *mother wavelet* ψ as $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$ with j, k integers. A function g can then be represented by the wavelet series $g(t) = \sum_{j,k \in \mathfrak{S}} d_{jk}\psi_{jk}(t)$, with wavelet coefficients $d_{jk} = \int g(t)\psi_{jk}(t)dt$ describing features of the function g at the spatial locations indexed by k and frequencies indexed by j . In this way, the wavelet decomposition provides a location and scale decomposition of the function.

Let $\mathbf{y} = (y_1, \dots, y_T)$ be a row vector containing values of a function taken at T equally spaced points. A fast algorithm, the *discrete wavelet transform* (DWT), exists for decomposing \mathbf{y} into a set of T wavelet and scaling coefficients (Mallat, 1989). This transform only requires $O(T)$ operations when T is a power of 2. The DWT can also be represented as matrix multiplication by an orthogonal matrix $W' = [W'_1, W'_2, \dots, W'_J, V'_J]$ where J is the coarsest level of the transform. A DWT applied to the vector \mathbf{y} of observations $\mathbf{d} = \mathbf{y}W'$ decomposes the data into sets of wavelet and scaling coefficients $\mathbf{d} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_J, \mathbf{c}_J]$, where $\mathbf{d}_j = \mathbf{y}W'_j$ are the wavelet coefficients at level or scale j , and $\mathbf{c}_J = \mathbf{y}V'_J$ the scaling coefficients. For simplicity, we refer to the entire set of wavelet and scaling coefficients \mathbf{d} as simply the wavelet coefficients. Each wavelet level j contains K_j coefficients. A similar algorithm for the inverse reconstruction, the inverse discrete wavelet transform (IDWT), also exists.

Wavelet regression is a nonparametric regression technique that is useful for modelling functional data that is spiky or otherwise characterized by local features. Suppose we observe a

response vector \mathbf{y} , represented by a row vector of length T on an equally-spaced grid \mathbf{t} and assumed to be some unspecified function of t plus white noise. That is, $\mathbf{y} = g(\mathbf{t}) + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim MVN(0, \sigma_e^2 I_T)$. Wavelet regression follows three steps. First, the data are projected into the wavelet space using the DWT. The corresponding wavelet space model is $\mathbf{d} = \boldsymbol{\theta} + \boldsymbol{\epsilon}^*$, where $\mathbf{d} = \mathbf{y}W'$ are the empirical wavelet coefficients, $\boldsymbol{\theta} = g(\mathbf{t})W'$ are the true function's wavelet coefficients, and $\boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}W' \sim MVN(0, \sigma_e^2 I_T)$ is the noise in the wavelet space.

Since the wavelet transform tends to distribute white noise equally among all wavelet coefficients but concentrates the signal on a small subset, most wavelet coefficients will tend to be small and consist almost entirely of noise, with the remaining few wavelet coefficients being large in magnitude and containing primarily signal. Thus, we can denoise the signal and regularize the observed function by taking the smallest wavelet coefficients and thresholding them or shrinking them strongly towards zero. This is done either using thresholding rules (e.g., Donoho and Johnstone, 1995), or by placing a mean zero shrinkage prior on the true wavelet coefficients (e.g. Abramovich, Sapatinas and Silverman 1998). An effective prior in this context should give rise to a nonlinear shrinkage profile, so that smaller coefficients are strongly shrunk while larger ones are left largely unaffected. This thresholding or shrinkage of the wavelet coefficients constitutes the second step of wavelet regression. Third, the thresholded or shrunk estimators of the true wavelet coefficients $\boldsymbol{\theta}$ are transformed back to the data space using the IDWT, yielding a nonparametric estimator of the function. This procedure accomplishes *adaptive regularization*, meaning that the functional estimates are denoised or regularized in a way that tends to retain dominant local features in the function. With the exception of Morris, et al. (2003a), previous literature on wavelet regression for functional responses has focused on the single function setting.

3 Functional Mixed Model

Here we introduce the functional mixed model framework upon which we base our methodology. This framework represents an extension of Laird and Ware (1982) to functional data, where the forms of the fixed and random effect functions are left completely unspecified. Other researchers (e.g. Shih, Weiss, and Taylor 1996, Brumback and Rice 1998, Rice and Wu 2001, Wu and Zhang 2002, Guo 2002, Wu and Liang 2004) have worked with similar models, although none have made the same modelling assumptions we describe here.

Suppose we observe a sample of N curves $Y_i(t), i = 1, \dots, N$ on a compact set \mathcal{T} , assumed

without a loss of generality to be $[0, 1]$. Our functional mixed model is given by

$$\mathbf{Y}(t) = X\mathbf{B}(t) + Z\mathbf{U}(t) + \mathbf{E}(t), \quad (1)$$

where $\mathbf{Y}(t) = \{Y_1(t), \dots, Y_N(t)\}'$ is a vector of observed functions, “stacked” as rows. Here, $\mathbf{B}(t) = \{B_1(t), \dots, B_p(t)\}'$ is a vector of fixed effect functions with corresponding $N \times p$ design matrix X , $\mathbf{U}(t) = \{U_1(t), \dots, U_m(t)\}'$ is a vector of random effect functions with corresponding $N \times m$ design matrix Z , and $\mathbf{E}(t) = \{E_1(t), \dots, E_N(t)\}'$ is a vector of functions representing the residual error processes.

Definition: A set of N stacked functions, $\mathbf{A}(t)$, all defined on the same compact set \mathcal{T} , is a realization from a *multivariate Gaussian process* with $N \times N$ between-row covariance matrix Λ and within-function covariance surface $\Sigma \in \mathcal{T} \times \mathcal{T}$, denoted $\mathbf{A}(t) \sim \mathcal{MG}\mathcal{P}(\Lambda, \Sigma)$, if the rows of $\Lambda^{-1/2}\mathbf{A}(t)$ are independent mean zero Gaussian processes with covariance surface $\Sigma(t_1, t_2)$, where $\Lambda^{-1/2}$ is the inverse matrix square root of Λ . This assumption implies that the covariance between $A_i(t_1)$ and $A_{i'}(t_2)$ is given by $\Lambda_{ii'}\Sigma(t_1, t_2)$. This distribution is the functional generalization of the matrix normal distribution (see Dawid, 1981). Note that a scalar identifiability condition must be set on either Λ or Σ , since letting $\Lambda = \Lambda/c$ and $\Sigma = \Sigma * c$ for some constant $c > 0$ yields the same likelihood. For example, one can set $\Lambda_{11} = 1$.

The set of random effect functions $\mathbf{U}(t)$ is assumed to be a realization from a multivariate Gaussian process with $m \times m$ between-function covariance matrix P and within-function covariance surface $Q(t_1, t_2)$, denoted by $\mathbf{U}(t) \sim \mathcal{MG}\mathcal{P}(P, Q)$. The residual errors are assumed to follow $\mathbf{E}(t) \sim \mathcal{MG}\mathcal{P}(R, S)$, independent of $\mathbf{U}(t)$.

This model is very general and includes many other models commonly used for functional data as special cases. For example, it reduces to a simple linear mixed model when the functional effects are represented by parametric linear functions. When $N = 1$, the model simplifies to a form in which traditional smoothing spline and wavelet regression models for single functions can be represented. If we omit the random effects and assume a factorial structure on the fixed effects, we get functional ANOVA models. Model (1) also includes the hierarchical functional model presented by Morris, et al. (2003a) as a special case.

This proposed model is very flexible. The fixed effects can be mean functions, functional main effects, functional interactions, functional linear coefficients for continuous covariates, interactions of functional coefficients with other effects, or any combination of these. The design matrix Z and between-curve correlation matrices P and R can be chosen to accommodate a myriad of different covariance structures between curves that may be suggested by the experimental

design. These include simple random effects, in which case $P = I$, as well as structures for functional data from nested designs, split-plot designs, subsampling designs, and designs involving repeated functions over time. The random effect portion of the model may be partitioned into $Z\mathbf{U}(t) = \sum_{h=1}^H Z_h \mathbf{U}_h(t)$ with $\mathbf{U}_h(t) \mathcal{MG}\mathcal{P}(P_h, Q_h)$, for example to allow multiple hierarchical levels of random effects or to allow different random effects distributions for different strata.

This model is similar to the functional mixed model in Guo (2002), with a couple of key differences. Guo (2002) assumes independent random effect functions ($P = R = I$ in our framework), while our model, by introducing P and R , can accommodate correlation across the functions. Also, Guo (2002) assumes a different structure on Q than we do here. For each level of random effects h , that paper assumes that $Q_h = L_h + \Sigma/\lambda_h$, where $L_h = \sigma_h^2 M' D M$ is the covariance induced by random intercept and linear terms whose design matrix is M , D is a structured 2×2 covariance matrix (assumed diagonal in their example), and σ_h^2 is a variance component estimated from the data. The parameter λ_h is a scalar smoothing parameter estimated from the data, and correlation matrix Σ is fixed based on the reproducing kernel for the chosen spline basis. Our assumptions on Q are described below in Section 4.2.

Of course, we can not directly fit model (1), since in practice we only observe samples of the continuous curves on some discrete grid. A discrete version of this model is given below, assuming that all observed functions are sampled on a common equally spaced grid $\mathbf{t} = (t_1 \dots t_T)'$. Recall that by our definition of functional data (sampled on a very fine grid), this assumption is not especially restrictive, since if the grid is fine enough, interpolation can be used to obtain a common grid without substantively changing the observed data. The model is

$$Y = XB + ZU + E, \tag{2}$$

where Y is an $N \times T$ matrix of observed curves on the grid \mathbf{t} , B is a $p \times T$ matrix of fixed effects, U is a $m \times T$ matrix of random effects, and E is an $N \times T$ matrix of residual errors. As defined above, X is an $N \times p$ matrix and Z is an $N \times m$ matrix, and the two are the design matrices for the fixed and random effect functions, respectively. Following the notation of Dawid (1981), U follows a matrix normal distribution with $m \times m$ between-row covariance matrix P and $T \times T$ between-column covariance matrix Q , which we denote by $U \sim \mathcal{MN}(P, Q)$. Another way to represent this structure is to say that $\text{vec}(U') \sim \mathcal{MVN}(0, P \otimes Q)$, where $\text{vec}(A)$ is the vectorized version of a matrix A obtained by stacking the columns, and \otimes is the Kronecker product, both defined as in Harville (1997). This assumption implies that the covariance between U_{ij} and $U_{i'j'}$ is $P_{ii'}Q_{jj'}$. The residual error matrix E is assumed to be $\mathcal{MN}(R, S)$. The within-random effect

curve covariance surface Q and residual error covariance surface S are $T \times T$ covariance matrices that are discrete approximations of the corresponding covariance surfaces in $\mathcal{T} \times \mathcal{T}$.

4 Wavelet-Based Functional Mixed Model

Having presented a conceptual functional mixed model for correlated functional data, we now describe our nonparametric wavelet-based approach to fit it. Our approach consists of three basic steps: (1) Compute the empirical wavelet coefficients for each observed curve, which we think of as projecting the observed curves from the data space to the wavelet space. (2) Use Markov Chain Monte Carlo methods to obtain posterior samples for quantities in the wavelet-space version of the functional mixed model. Projecting to the wavelet space allows modelling to be done in a more parsimonious and computationally efficient manner, and causes regularization to be performed as a natural consequence of the modelling through shrinkage priors placed on the fixed effects portion of the model. (3) Transform the wavelet-space quantities back to the data space, yielding posterior samples of all quantities in the data space model, which can be used to perform Bayesian estimation, inference, and prediction.

The first step involves decomposing each observed function, sampled on an equally spaced grid of size T , into a set of T wavelet coefficients. This projection from the data space into the wavelet space is done by applying the discrete wavelet transform (DWT) to each row of Y , and can be conceptualized as the right matrix multiplication $D = YW'$, where W is the orthogonal DWT matrix. The $N \times T$ matrix D contains the empirical wavelet coefficients for all observed curves, with row i containing wavelet and scaling coefficients for curve i and the columns double-indexed by the scale j and location k , with $j = 1, \dots, J$ and $k = 1, \dots, K_j$.

4.1 Wavelet Space Model

Right matrix multiplication of both sides of model (2) by the DWT matrix W' yields a wavelet-space version of the model:

$$D = XB^* + ZU^* + E^*, \quad (3)$$

where X and Z are the design matrices as in model (2), $B^* = BW'$ is a $p \times T$ matrix whose rows contain the wavelet coefficients for the p fixed effect functions on the grid, $U^* = UW'$ is a $m \times T$ matrix whose rows contain the wavelet coefficients for the m random effect functions, and $E^* = EW'$ is a $N \times T$ matrix consisting of the residual errors in the wavelet space. Like D , the columns of B^* , U^* , and E^* are all double-indexed by the wavelet coefficients' scale j and location

k . The linearity of the DWT makes it easy to compute the induced distributional assumptions of the random matrices in the wavelet space, $U^* \sim \mathcal{MN}(P, Q^*)$ and $E^* \sim \mathcal{MN}(R, S^*)$, where $Q^* = WQW'$ and $S^* = WSW'$. Note that the between-row covariance structure is retained when projecting into the wavelet space; only the column covariance changes.

4.2 Covariance Assumptions

Before we fit model (3), it is necessary to specify some structure on the various covariance matrices since their large dimensions make it infeasible to estimate them in a completely unstructured fashion. We model P and R using parametrically structured covariance matrices as in linear mixed models, which can be chosen based on either the experimental design or an empirical investigation of the data. The vectors of the covariance parameters indexing these matrices are denoted by Ω_P and Ω_R , respectively.

For Q and S , we propose a parsimonious structure in the wavelet space that yields a flexible class of covariance surfaces in the data space. As is frequently done in wavelet regression, we assume that the wavelet coefficients within a given curve are independent across j and k , making Q^* and S^* diagonal. The heuristic justification frequently given for this assumption is the whitening property of the wavelet transform, which is discussed in Johnstone and Silverman (1997). The diagonal elements are allowed to vary across both wavelet scales j and locations k , yielding $Q^* = \text{diag}(q_{jk}^*)$ and $S^* = \text{diag}(s_{jk}^*)$. For convenience, we denote these sets of variance components by Ω_Q and Ω_S , respectively.

This structure requires only T parameters instead of the $T(T+1)/2$ parameters that would be required to estimate each of these matrices in an unstructured fashion, yet it is flexible enough to emulate a wide range of covariance structures commonly encountered in functional data. For example, when $T = 256$, only 256 parameters are required instead of the 32,896 for unstructured representation. Note that independence in the wavelet space does not imply independence in the data space unless the variance components are identical across all wavelet scales j and locations k , since heterogeneity in variances across wavelet coefficients at different levels induces serial dependencies in the data. In general, larger variances at low frequency scales correspond to stronger serial correlations, and thus smoother functions.

Further, since the variance components are free to vary across both scale j and location k , this structure accommodates nonstationarity, for example allowing the curve-to-curve variances and the smoothness in the curve-to-curve deviations to both vary over t . These types of nonstationarities are frequently encountered in complex functional data, but cannot be accommodated

when the variance components are only allowed to vary over j (see Figure 1). It is typical in existing wavelet regression literature for the wavelet-space variance components to vary over j , but not k . (e.g. Abramovich, Sapatinas, and Silverman 1998, Morris, et al. 2003, Abramovich and Angelini 2003, Antoniadis and Sapatinas 2004). This may be a necessary practical restriction in the single function case, but not in the multiple function case, since the replicate functions allow the variance components to be estimable even when they also vary by k . To our knowledge, this is the first paper allowing these variance components to depend on both j and k .

Another advantage of this independence assumption is that it allows us to fit the wavelet-space model (3) one column (wavelet coefficient) at a time. This greatly simplifies the computational procedure and allows much larger data sets to be fit using this method.

4.3 Adaptive Regularization Using a Multiple Shrinkage Prior

In order to obtain adaptively regularized representations of the fixed effect functions $B_i(t)$, as is standard in Bayesian implementations of wavelet regression, we place a mixture prior on B_{ijk}^* , the wavelet coefficient at scale j and location k for fixed effect i :

$$\begin{aligned} B_{ijk}^* &= \gamma_{ijk}^* \mathcal{N}(0, \tau_{ijk}) + (1 - \gamma_{ijk}^*) I_0, \\ \gamma_{ijk}^* &= \text{Bernoulli}(\pi_{ij}), \end{aligned} \tag{4}$$

where I_0 is a point mass at zero and γ_{ijk}^* an indicator of whether wavelet coefficient (j, k) is “important” for representing the signal for fixed effect function i . The hyperparameter π_{ij} is the prior probability that a wavelet coefficient at wavelet scale j is important for representing the fixed effect function i , and τ_{ijk} is the prior variance of any important wavelet coefficient at location k and level j for fixed effect i .

The quantities π_{ij} and τ_{ijk} are regularization parameters. For example, smaller π_{ij} will result in more attenuation in the features of fixed effect function i occurring at a frequency indicated by scale j . By indexing these parameters by i and j , we allow different degrees of regularization for different fixed effect functions and at different frequencies. See Morris, et al. (2003a) for a discussion of the intuition behind how this prior leads to adaptive regularization. It is possible to elicit values for these regularization parameters, taking into account some of the considerations discussed in Morris, et al. (2003a) or Abramovich, Sapatinas, and Silverman (1998), or to estimate them from the data using an empirical Bayes procedure. Section 4.4 describes one such procedure.

In this modelling framework, the random effect functions $U_i(t)$ are also regularized as a result

of the mean zero Gaussian distribution on their wavelet coefficients. Morris, et al. (2003b) described how the regularization of the random effect functions in their wavelet-based hierarchical functional model was governed by the relative sizes of corresponding variance components and residual errors. The same principles also apply here, although here our regularization is more adaptive than in Morris, et al. (2003a) since we allow the wavelet-space variance components for both the random effects and residual errors to depend on scale j and location k . To explain, wavelet coefficients indexed by (j, k) that tend to be important for representing even a small number of random effect functions will have relatively large subject-level variance components q_{jk} . These large variances will lead to less shrinkage of these coefficients, and thus the features represented by these coefficients will tend to be preserved in the regularized random effect function estimates. Wavelet coefficients that are unimportant for representing the random effect functions will be close to zero, leading to small variance components, strong shrinkage, and regularization of the features corresponding to these coefficients.

This regularization is adaptive enough to model very spiky random effect functions, as demonstrated in supplementary material available from a link on the first author's web site (<http://biostatistics.mdanderson.org/Morris/papers.html>). A major advantage of our approach is that the random effect functions' regularization parameters are simply the variance components of the model, which are directly estimated from the data, and thus need not be arbitrarily chosen. Further, in our Bayesian approach, the uncertainty of their estimation is automatically propagated throughout any inference that is done.

It may be possible to obtain even more adaptively randomized random effect functions by assuming a mixture prior like (4) on the wavelet coefficients for the random effect functions. However, by doing so, we would lose some of the coherency evident in models (1)-(3), since the random effect functions would no longer be Gaussian in the data space. Further, we would not be able to marginalize over the random effect functions in our model fitting (see Section 5), which would increase the computational burden for implementing the method. Since we are satisfied with the degree of adaptiveness afforded by our Gaussian assumptions with variances depending on j and k , we do not further pursue this idea in this paper.

4.4 Empirical Bayes for Selecting Shrinkage Hyperparameters

Here we present a data-based procedure for determining the shrinkage hyperparameters for the fixed effect functions in the wavelet-based functional mixed model. We estimate these hyperparameters using maximum likelihood while conditioning on consistent estimates of the

variance components in the model. This method is an extension of the work of Clyde and George (2000), which they later adapted to the hierarchical functional framework (Clyde and George, 2003).

First we introduce some notation. Consider the following quantities:

$$\widehat{B}_{ijk,MLE}^* = \{X_i'(\Sigma_{jk})^{-1}X_i\}^{-1}X_i'(\Sigma_{jk})^{-1}\{\mathbf{d}_{jk} - \mathbf{X}_{(-i)}\widehat{B}_{(-i)jk,MLE}^*\} \quad (5)$$

$$V_{ijk} = \text{var}(\widehat{B}_{ijk,MLE}^*) = \{X_i'(\Sigma_{jk})^{-1}X_i\}^{-1}, \quad (6)$$

where X_i is the i^{th} column of the design matrix and $X_{(-i)}$ is the design matrix with column i omitted, and

$$\Sigma_{jk} = ZP(\boldsymbol{\Omega}_P)Z' * q_{jk}^* + R(\boldsymbol{\Omega}_R) * s_{jk}^*. \quad (7)$$

is the marginal variance of \mathbf{d}_{jk} . Note that $\widehat{B}_{ijk,MLE}^*$ is the maximum likelihood estimator of B_{ijk}^* conditional on the covariance parameters and the other fixed effects, and $\sqrt{V_{ijk}}$ is the standard error of the MLE. Taking their ratio yields

$$\zeta_{ijk} = \widehat{B}_{ijk,MLE}^*/\sqrt{V_{ijk}}, \quad (8)$$

which can be thought of as a standardized score for the wavelet coefficient at scale j and location k from fixed effect function i .

We assume that $\tau_{ijk} = V_{ijk}\Upsilon_{ij}$ for some parameters Υ_{ij} , allowing full flexibility in these regularization parameters across different scales, but making the ratio of regularization parameters within a given scale proportional to the size of the variance of the MLE for that coefficient. This allows us to estimate Υ_{ij} from the data. Assuming knowledge of V_{ijk} , it can be shown that the likelihood for Υ_{ij} and π_{ij} can be represented by

$$l(\Upsilon_{ij}, \pi_{ij}) \propto (1 + \Upsilon_{ij})^{(-\sum_{k=1}^{K_j} \gamma_{ijk}^*)/2} \left[\exp \left\{ -1/2 \sum_{k=1}^{K_j} \zeta_{ijk}^2 \gamma_{ijk}^* / (1 + \Upsilon_{ij}) \right\} \right] \times (\pi_{ij})^{\sum_{k=1}^{K_j} \gamma_{ijk}^*} (1 - \pi_{ij})^{K_j - \sum_{k=1}^{K_j} \gamma_{ijk}^*}. \quad (9)$$

Based on this likelihood, local maximum likelihood estimates of π_{ij} and Υ_{ij} can be obtained by iterating through the following steps until convergence is achieved.

$$\begin{aligned} \widehat{\gamma}_{ijk}^* &= \widehat{O}_{ijk}/(1 + \widehat{O}_{ijk}) \\ \widehat{O}_{ijk} &= \left\{ \frac{\widehat{\pi}_{ij}}{1 - \widehat{\pi}_{ij}} \right\} (1 + \widehat{\Upsilon}_{ij})^{-1/2} \exp \left(1/2 \zeta_{ijk}^2 \frac{\widehat{\Upsilon}_{ij}}{1 + \widehat{\Upsilon}_{ij}} \right) \end{aligned}$$

$$\hat{\Upsilon}_{ij} = \max \left\{ 0, \frac{\sum_{k=1}^{K_j} \hat{\gamma}_{ijk}^* \zeta_{ijk}^2}{\sum_{k=1}^{K_j} \hat{\gamma}_{ijk}^*} - 1 \right\}$$

$$\hat{\pi}_{ij} = \sum_{k=1}^{K_j} \hat{\gamma}_{ijk}^* / K_j.$$

This procedure can be applied while conditioning on consistent estimators of the variance components, e.g. method of moment or maximum likelihood estimators, giving \hat{V}_{ijk} of V_{ijk} . Then the empirical Bayes estimates of π_{ij} and τ_{ijk} are given by $\hat{\pi}_{ij}$ and $\hat{V}_{ijk} * \hat{\Upsilon}_{ij}$, respectively.

5 Posterior Sampling Using MCMC

After specifying diffuse proper priors for the variance components, we are left with a fully specified Bayesian model for the functional data. Since the posterior distributions of parameters are not available in closed form, we use Markov Chain Monte Carlo (MCMC) to obtain posterior samples for all parameters in model (3). We work with the marginalized likelihood where the random effects have been integrated out, which improves the mixing properties of the sampler over a naive Gibbs sampler. We alternate between sampling the fixed effects B^* and the covariance parameters $\mathbf{\Omega}$, then later sample the random effects U^* whenever they are of interest. Following are the details of the sampling procedure we use.

1. For each wavelet coefficient (j, k) , sample fixed effect i from $f(B_{ijk}^* | \mathbf{d}_{jk}, B_{(-i)jk}^*, \mathbf{\Omega})$, where $B_{(-i)jk}^*$ is the set of all fixed effects coefficients at scale j and location k except the i^{th} one. As shown in the appendix, this distribution is a mixture of a point mass at 0 and a normal distribution, with the normal mixture proportion α_{ijk} and the mean and variances of the normal μ_{ijk} and v_{ijk} , respectively, given by:

$$\alpha_{ijk} = \Pr(\gamma_{ijk} = 1 | \mathbf{d}_{jk}, B_{(-i)jk}^*, \mathbf{\Omega}) = O_{ijk} / (O_{ijk} + 1), \quad (10)$$

$$O_{ijk} = \pi_{ij} / (1 - \pi_{ij}) BF_{ijk},$$

$$BF_{ijk} = (1 + \tau_{ijk} / V_{ijk})^{-1/2} \exp\{1/2(\zeta_{ijk})^2(1 + V_{ijk} / \tau_{ijk})\}, \quad (11)$$

$$\mu_{ijk} = \hat{B}_{ijk,MLE} (1 + V_{ijk} / \tau_{ijk})^{-1}, \quad (12)$$

$$v_{ijk} = V_{ijk} (1 + V_{ijk} / \tau_{ijk})^{-1}, \quad (13)$$

where $\hat{B}_{ijk,MLE}^*$, V_{ijk} , Σ_{jk} , and ζ_{ijk} are defined as in (5) – (8) above. Note that O_{ijk} and BF_{ijk} have an interesting interpretation. They are the posterior odds and Bayes factor, respectively, for deciding whether wavelet coefficient (j, k) is important for representing

function i , conditional on the covariance parameters Ω and other fixed effects. The posterior means of the B_{ijk} will be Bayesian model averaged estimators that have averaged over models where B_{ijk} is either 0 or not. Alternatively, a soft thresholding approach could be used whereby $\widehat{B}_{ijk} = 0$ if the estimated posterior probability that $|B_{ijk}| > 0$ (that is, $\gamma_{ijk} = 1$) from the MCMC is less than some threshold.

2. For each wavelet coefficient (j, k) , sample the elements q_{jk}^* and s_{jk}^* of Ω_Q and Ω_S using a random walk Metropolis-Hastings. The objective function is $f(q_{jk}^*, s_{jk}^* | \mathbf{d}_{jk}, B_{jk}^*, \Omega_P, \Omega_R) \propto |\Sigma_{jk}|^{-1/2} [\exp\{-1/2(\mathbf{d}_{jk} - XB_{jk}^*)'(\Sigma_{jk})^{-1}(\mathbf{d}_{jk} - XB_{jk}^*)\}] f(q_{jk}^*, s_{jk}^*)$. We use an independent Gaussian density, truncated at zero and centered at the previous parameter values, as the proposal for each parameter. We automatically estimate the proposal variance from the data using estimates of the variance of the maximum likelihood estimates. Wolfinger, et al. (1994) provide details of how to compute maximum likelihood estimates and their standard errors in linear mixed models. The details of the Metropolis-Hastings procedure are available as supplementary material on the first authors' web site (<http://biostatistics.mdanderson.org/Morris/papers.html>).
3. Sample the between-curve covariance parameters Ω_P and Ω_R using a single random walk Metropolis-Hastings step. If the random effects and residual errors are assumed to be independent and homoscedastic across samples ($P = I$ and $R = I$), then there are no parameters to update in this step. The assumption of independence among the wavelet coefficients allows the Metropolis-Hastings objective function to factor into the product of independent pieces for each wavelet coefficient: $f(\Omega_P, \Omega_R | D, B^*, \Omega_Q, \Omega_S) \propto \prod_{j,k} |\Sigma_{jk}|^{-1/2} [\exp\{-1/2(\mathbf{d}_{jk} - XB_{jk}^*)'(\Sigma_{jk})^{-1}(\mathbf{d}_{jk} - XB_{jk}^*)\}] f(\Omega_P, \Omega_R)$, where Σ_{jk} is given by equation (7) above. The implementation details are similar to those for the previous step. Again, we use an independent truncated Gaussian with mean at the previous parameter values for the proposal distribution, with the proposal variance automatically determined from the data.
4. Sample the random effects \mathbf{u}_{jk}^* for each (j, k) from their full conditional $f(\mathbf{u}_{jk}^* | \mathbf{d}_{jk}, B_{jk}^*, \Omega)$, which is easily seen to be Gaussian with mean $\{\Psi_{jk}^{-1} + (q_{jk}^* * P)^{-1}\}^{-1} \Psi_{jk}^{-1} \widehat{\mathbf{u}}_{NS,jk}$ and variance $\{\Psi_{jk}^{-1} + (q_{jk}^* * P)^{-1}\}^{-1}$, where $\Psi_{jk} = \{Z'(s_{jk}^* * R)^{-1}Z\}^{-1}$ and $\widehat{\mathbf{u}}_{NS,jk} = \{Z'(s_{jk}^* * R)^{-1}Z\}^{-1}Z'(s_{jk}^* * R)^{-1}(\mathbf{d}_{jk} - XB_{jk}^*)$. Note that if the random effects are not desired, we can omit this step and thus speed up the MCMC, since the previous steps work with the

marginalized likelihood.

Links to code for applying this method is available on the first author's web site, <http://biostatistics.mdanderson.org/Morris/papers.html>.

5.1 Bayesian Inference and Prediction

The MCMC described above yields posterior samples for all quantities in the wavelet-space mixed model (3). These posterior samples can then be projected back into the data space using the IDWT, yielding posterior samples of the quantities in model (2). Specifically, posterior samples for each fixed effect function $B_i(t)$ on the grid \mathbf{t} are obtained by applying the IDWT to each posterior sample of the corresponding vector of wavelet coefficients $B_i^* = \{B_{i11}^*, \dots, B_{iJK_J}^*\}$, and similarly for the random effect functions. Further, posterior samples of the covariance matrices Q and S are obtained by applying the 2-dimensional IDWT to the posterior samples of the diagonal matrices Q^* and S^* , following Vannucci and Corradi (1999).

Given the posterior samples, we are then able to construct any Bayesian estimators and perform any desired Bayesian inference. See Gelman et al. (2004) for an overview of Bayesian analysis and inference, and a description of the types of inference possible given posterior samples. For example, we can construct pointwise credible intervals for fixed effect functions or compute posterior probabilities for any hypotheses of interest. These can involve any transformation or combination of the parameters in the model. Since we have posterior samples for entire functions, marginal inference can be done for single locations on the function or joint inference can be done over regions of the function. It is also straightforward to compute posterior predictive distributions $f(Y^*|Y)$ for a future observed curve Y^* given data Y , since $f(Y^*|Y) = \int f(Y^*|B, U, \Omega) f(B, U, \Omega|Y) dBdUd\Omega$, which can be estimated via Monte Carlo integration using the posterior samples as: $G^{-1} \sum_g f(Y^*|B^{(g)}, U^{(g)}, \Omega^{(g)})$, where the superscript (g) indicates the posterior sample from iteration g of the MCMC. This inference and prediction appropriately accounts for all sources of variation in the model. For example, it does not condition on estimates of the variance components as if they were known, but automatically propagates the uncertainty of their estimation throughout inference. This is one of the advantages of using a unified Bayesian modelling approach.

6 Example

Nutrition researchers at Texas A&M University conducted a rat carcinogenesis experiment to investigate whether the type of dietary fat (fish oil or corn oil) plays a role in modulating

important colon cancer biomarkers during the initiation stage of carcinogenesis, the first hours after carcinogen exposure. In this study, they fed 30 rats one of the two diets for 14 days, exposed them to a carcinogen, then sacrificed them at one of 5 times after exposure to the carcinogen (0, 3, 6, 9, or 12 hours). They removed and dissected each rat's colon, then used immunohistochemical staining to obtain measurements of various cancer biomarkers, including the DNA adduct level, a measurement of the amount of DNA damage occurring from the carcinogen exposure; MGMT, a DNA repair enzyme that repairs this carcinogen-induced damage; and apoptosis, the selective elimination of damaged cells.

They quantified each biomarker for a separate set of roughly 25 crypts in the distal region of each rat's colon. Crypts are finger-like structures extending into the colon wall in which all colon cells reside. A cell's relative depth within its crypt is related to its age and stage in the cell cycle, so it is an important factor to consider when assessing biomarker modulation. Using image analysis software, they quantified the MGMT levels on a fine grid along the side of each selected crypt, resulting in an observed curve for each crypt containing the biomarker quantifications as a function of relative depth within the crypt. The relative depth in the crypt was coded such that an observation at the base of the crypt was relative cell position 0, while an observation at the luminal surface was relative cell position 1. Figure 2 contains the observed curves from two crypts from two rats. Note that these functions appear very irregular, with many spikes presumably corresponding to local areas in the crypt with high biomarker levels (Morris, et al. 2003a), e.g., the nuclei of the cells. The full data set consists of 738 such observed curves, each sampled on an equally-spaced 256-unit grid.

The MGMT data were analyzed by Morris, et al. (2003a), and it was found that corn oil-fed rats had lower MGMT expression near the luminal surface at 12 hours after carcinogen exposure than did fish oil-fed rats. Our goal here is to relate the levels of the other biomarkers to the MGMT expression levels, and see if this 12 hour effect remains after adjusting for these other biomarkers as covariates. For each rat, we obtained measurements of the continuous covariates mean DNA adduct level and apoptotic index (percentage of cells undergoing apoptosis) across its crypts in the upper 1/3 compartment, i.e., the compartment closest to the luminal surface. We would like to assess whether there is a relationship between the amount of DNA damage and/or amount of apoptosis near the luminal surface of the crypts and the levels of MGMT, and whether these relationships depend on relative cell position and/or diet. These covariates were not considered in Morris, et al. (2003a), and could not be accommodated by their hierarchical

functional model.

Our design matrix X had $p = 14$ columns, with the first 10 indicating the rat's diet by time group. Columns 11 and 12 contained the mean DNA adduct level in the upper 1/3 of the crypt for rats fed the fish and corn oil diets, respectively. These columns were standardized to have mean 0 and standard deviation 1. Columns 13 and 14 contained the apoptotic index in the upper 1/3 of the crypt for rats fed the fish and corn oil diets, respectively. To model the correlation between crypts from the same rat, we included random effect functions for each rat. The residual errors represented the sum of the crypt-to-crypt variability and any within-function noise. We assumed rats and crypts within rats were independent and identically distributed, so we let $P = R = I$. We used the Daubechies wavelet with 8 vanishing moments (Daubechies, 1992) at $J = 8$ levels. Other wavelet bases yielded substantively equivalent results. After a burn-in of 1000, we ran the MCMC for 20,000 iterations, keeping every 10. The Metropolis Hastings acceptance probabilities for the variance components were all between 0.12 and 0.39. Trace plots of the model parameters are available from the first author (<http://biostatistics.mdanderson.org/Morris/papers.html>), and reveal the MCMC converged and mixed very well.

Figure 3 contains the posterior mean functional coefficients corresponding to the DNA adduct level and apoptotic index covariates for fish and corn oil rats. The estimate for the DNA adduct level-top coefficient was negative near the luminal surface for rats fed fish oil or corn oil, meaning that animals with high levels of DNA damage near the luminal surface tended to also have lower levels of MGMT near the luminal surface. The posterior probabilities that the coefficient at the top of the crypt was less than zero were 0.947 and 0.989 for fish and corn oil, respectively. This negative relationship extended to the middle of the crypts for corn oil rats, but not for fish oil rats, for whom the estimate was positive. The posterior probability that the fish oil coefficient at the middle of the crypt (relative cell position 0.5) was greater than that for the corn oil coefficient was 0.9965.

For fish oil-fed rats, the apoptotic index-top coefficient was positive throughout nearly the entire crypt, with the coefficient increasing in a roughly linear fashion moving up the crypt. The posterior probability that this coefficient was greater than 0 at the luminal surface for fish and corn oil-fed rats was > 0.9995 and 0.612, respectively, while the posterior probability that fish > corn was 0.9815. The interpretation of these results is that the fish oil-fed animals who had a large amount of apoptosis near their luminal surface also had high levels of the DNA repair enzyme MGMT near their luminal surface, meaning the two major mechanisms for dealing with

DNA damage were correlated. This relationship was not so strong for corn oil-fed animals.

With DNA adduct level and apoptotic index and their interactions with diet included in the model, the difference between fish oil and corn oil at 12hr near the luminal surface found in Morris, et al. (2003a) was no longer evident (posterior probability that fish>corn was only 0.674, while it was > 0.9995 without covariates in the model). One interpretation of this result is that the differences in MGMT between diets at the luminal surface may be explained by the previously observed DNA adduct level and apoptosis effects (Hong, et al. 2000), whereby rats on fish oil diets had lower DNA adduct levels and higher apoptotic rates at the lumen surface than rats fed corn oil diets.

7 Discussion

Functional data are increasingly encountered in scientific studies, and there is a need for systematic methods for analyzing these complex and large data sets and extracting the meaningful information contained inside them. In this paper, we have introduced a unified Bayesian wavelet-based modelling approach for functional data that is a vast extension over the hierarchical functional method introduced by Morris, et al. (2003a). Although applied to just one example here, our approach is flexible enough to be applied to a very broad range of functional data sets and address a large number of potential research questions. Note that if we substitute higher-dimensional wavelet transforms for the one-dimensional transforms described here, our methodology is immediately extendable to higher dimensional functional data, e.g. image data.

The underlying functional mixed models framework is very flexible, allowing the same wide range of mean and covariance structures as in mixed effects models, while allowing functional fixed and random effects of unspecified form. We perform our modelling in the wavelet space, which provides a natural mechanism for adaptive regularization using mixture prior distributions, and also allows us to model the high dimensional covariance matrices Q and S describing the form of the curve-to-curve deviations in a parsimonious manner. As in much work in wavelet regression, we assume independence in the wavelet space, but unlike existing work in wavelet regression, we allow the wavelet space variance components to vary across both scale j and location k . This provides a great deal of flexibility, accommodating various types of nonstationarity commonly encountered in functional data, including heteroscedasticity and varying degrees of smoothness at different locations in the curve-to-curve deviations, see Figure 1. This flexibility allows us to model many different types of functional data, and also results in more adaptive regularization in the representations of the fixed and random effect functions. This approach

is able to effectively accommodate spiky fixed effect functions and/or spiky random effect functions. In our example, the fixed effect and rat-level random effect functions were smooth, but the crypt-level deviations were spiky.

After running an MCMC, we obtain posterior samples of the fixed and random effect functions and various covariance matrices in the model, which can be used to perform any desired Bayesian estimation, inference, or prediction. Credible intervals can be constructed and posterior probabilities of hypotheses can be computed for any transformation or function of the model parameters, for example averaging over different intervals or looking at specific locations of interest. Also, predictive densities for future curves can be estimated. While our method is Bayesian, the only informative priors we use in our analyses involve the shrinkage hyperparameters, which can be estimated from the data using the empirical Bayes method we describe, if desired. Another advantage of the Bayesian approach is that there is a natural mechanism for handling measurement error or missingness, both in covariates and in the functional responses, since the missing or error prone data can simply be treated as parameters that are updated from their complete conditional distributions as part of the MCMC. Also, the structure of our framework makes it possible to consider functional hypothesis testing using Bayes Factors or mixture priors with positive probabilities placed on zero functions. These ideas require further development, however, so are beyond the scope of this paper and topics of future investigation.

There is some recent and ongoing related work on functional ANOVA using wavelets. Unlike here, the major focus in these papers is on developing frequentist functional hypothesis tests. A paper by Fan and Lin (1998) presented methods for functional testing using wavelets, although their framework did not include random effects. Abramovich and Angelini (2003) allowed functional random effects, but only dealt with one-way ANOVA mean structures. Antoniadis and Sapatinas (2004) also allowed functional random effects, and they described a functional mixed modelling framework similar to (1), but did not accommodate correlated random effect functions.

There are other important differences between our modelling framework and those used in these papers. While we let the wavelet space variance components depend on scale j and location k , these papers only allow them to depend on j , which places strong restrictions on functional forms of the between-curve deviations (see Figure 1), which we expect should affect any subsequent inference. Also, since we specify diffuse proper priors for the wavelet-space variance components for the random effects and update them within the MCMC, we estimate these parameters from the data and propagate the uncertainty of their estimation throughout

subsequent inference. These variance components both model the curve-to-curve variability and serve as regularization parameters for the random effect functions. In Antoniadis and Sapatinas (2004), the user simply fixes the relative sizes of these variance components across different wavelet scales j , then only estimates a single scalar variance component from the data. Abramovich and Angelini (2003) describe a data-based method for estimating them, but they condition on these estimates as though they were known, and thus the inference they describe does not account for their estimation error.

These papers focus on functional hypothesis testing for fixed effect functions and, in Antoniadis and Sapatinas (2004), random effect functions. This is clearly of interest in many contexts, but is not the only relevant question with functional data. For example, the primary interest in many applications is not simply testing whether the function is identically 0, but rather identifying specific regions or features of the curves that differ from zero. No inferential procedures for these questions are described in these papers. One example is mass spectrometry proteomics, where the functions are characterized by many peaks corresponding to different proteins present in the sample. The primary goal is not to simply decide whether there are any systematic differences in the mean curves for different groups of patients, but rather to identify which regions of the curves demonstrate differences. These specific regions can subsequently be mapped to individual proteins that may serve as useful biomarkers in medical applications.

We have developed easy-to-use code for implementing our method that we make freely available via a link on the first author's web site (<http://biostatistics.mdanderson.org/Morris/papers.html>). The minimum information a user needs to supply includes a matrix of observed functions Y , fixed and random design matrices X and Z , and a specification of the desired covariance structures and wavelet bases to use. Method of moments and generalized least squares starting values, vague proper priors on the variance components, and empirical Bayes values for the hyperparameters are all automatically computed by the program and can be used, if desired. The program also contains an automatic, data-based method for determining the proposal variances necessary for the Metropolis-Hastings steps used to sample the large number of covariance parameters in the model. This method appears to work very well with none of the fine tuning normally required when implementing random-walk Metropolis-Hastings algorithms. This feature is key in making our method practically implementable for high dimensional functional data.

8 Acknowledgements

We thank Phil Brown, Marina Vannucci, Louise Ryan, Kevin Coombes, Keith Baggerly, Peter Mueller and Yuan Ji for useful discussions regarding this work. We also thank Joanne Lupton, Rob Chapkin, Nancy Turner, and Meeyoung Hong for the colon carcinogenesis data, and Dick Herrick for his assistance in helping deal with various computational issues that arose in coding the method. We also would like to thank the associate editor and referees, whose questions and insightful comments have led to a much improved paper. Morris' effort was supported by the National Cancer Institute (CA-107304). Carroll's research was supported by the National Cancer Institute (CA-57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

9 Appendix: Conditional Distribution for Fixed Effects

Here we show that the conditional distribution $(B_{ijk}^* | \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega)$ is a mixture of a point mass at zero and a normal, with normal mixing proportion α_{ijk} given by (10) and the mean and variances of the normal μ_{ijk} and v_{ijk} given by (12) and (13), respectively.

Recall that after integrating the random effects out of model (3), we have $\mathbf{d}_{jk} \sim MVN(X\mathbf{B}_{jk}^*, \Sigma_{jk})$ where $\Sigma_{jk} = ZP(\boldsymbol{\Omega}_P)Z' * q_{jk}^* + R(\boldsymbol{\Omega}_R) * s_{jk}^*$ as defined in (7). The prior for B_{ijk}^* is given by (4), which is a mixture of a $N(0, \tau_{ijk})$ and a point mass at 0, with γ_{ijk}^* the indicator for the normal component of the mixture, which itself has a Bernoulli(π_{ij}) prior distribution.

Note that we can write:

$$\begin{aligned} f(B_{ijk}^* | \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) &= \int f(B_{ijk}^* | \gamma_{ijk}^*, \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) f(\gamma_{ijk}^* | \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) d\gamma_{ijk}^* \\ &= f(B_{ijk}^* | \gamma_{ijk}^* = 1, \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) \Pr(\gamma_{ijk}^* = 1 | \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) \quad (\text{A.1}) \\ &\quad + f(B_{ijk}^* | \gamma_{ijk}^* = 0, \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) \Pr(\gamma_{ijk}^* = 0 | \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) \quad (\text{A.2}) \end{aligned}$$

We will first show that $f(B_{ijk}^* | \gamma_{ijk}^* = 1, \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega)$ in (A.1) is normal with mean μ_{ijk} and variance v_{ijk} . Second, we will show that $\Pr(\gamma_{ijk}^* = 1 | \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega)$ in (A.1) is equal to α_{ijk} . It is trivial to show that in (A.2), $f(B_{ijk}^* | \gamma_{ijk}^* = 0, \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) = I_0$ and $\Pr(\gamma_{ijk}^* = 0 | \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) = 1 - \alpha_{ijk}$.

First note that

$$\begin{aligned} f(B_{ijk}^* | \gamma_{ijk}^* = 1, \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega) &\propto f(\mathbf{d}_{jk} | B_{ijk}^*, B_{(-i)jk}^*, \Omega) f(B_{ijk}^* | \gamma_{ijk}^* = 1) \\ &\propto \exp[-1/2 * \{(\mathbf{d}_{jk}^* - X_i B_{ijk}^*)' \Sigma_{jk}^{-1} (\mathbf{d}_{jk}^* - X_i B_{ijk}^*)\}] \quad (\text{A.3}) \\ &\quad * \exp[-1/2 * \{(B_{ijk}^*)^2 / \tau_{ijk}\}], \quad (\text{A.4}) \end{aligned}$$

where $\mathbf{d}_{jk}^* = (\mathbf{d}_{jk} - X_{(-i)}B_{(-i)jk}^*)$ are the ‘‘residuals’’ after conditioning on the other fixed effect parameters. Multiplying (A.3) by the constant term $\exp[-1/2 * \text{tr}\{(X_i'\Sigma_{jk}^{-1}X_i)(X_i'\Sigma_{jk}^{-1}X_i)^{-1}(X_i'\Sigma_{jk}^{-1}X_i)(X_i'\Sigma_{jk}^{-1}X_i)^{-1}\}]$, reorganizing the terms within the trace and simplifying yields

$$\exp[-1/2 * \{(B_{ijk}^* - \widehat{B}_{ijk,MLE}^*)'V_{ijk}^{-1}(B_{ijk}^* - \widehat{B}_{ijk,MLE}^*)\}], \quad (\text{A.5})$$

where $\widehat{B}_{ijk,MLE}^* = (X_i'\Sigma_{jk}^{-1}X_i)^{-1}X_i'\Sigma_{jk}^{-1}\mathbf{d}_{jk}^*$ and $V_{ijk} = (X_i'\Sigma_{jk}^{-1}X_i)^{-1}$, as defined in (5) and (6). Combining the terms in (A.5) and (A.4) and completing the square leaves us with $\exp[-1/2 * (B_{ijk}^* - \mu_{ijk})^2/v_{ijk}]$, which is the kernel of a $N(\mu_{ijk}, v_{ijk})$, thus proving the first part.

For the second part, note that $\Pr(\gamma_{ijk}^* = 1 | \mathbf{d}_{jk}, B_{(-i)jk}^*, \Omega)$ can be written as $O_{ijk}/(O_{ijk} + 1)$, where O_{ijk} is the conditional odds of $\gamma_{ijk}^* = 1$ vs. $\gamma_{ijk}^* = 0$, which can be written as a product of the prior odds $\pi_{ij}/(1 - \pi_{ij})$ and the conditional Bayes Factor

$$BF_{ijk} = \frac{f(\mathbf{d}_{jk} | \gamma_{ijk}^* = 1, B_{(-i)jk}^*, \Omega)}{f(\mathbf{d}_{jk} | \gamma_{ijk}^* = 0, B_{(-i)jk}^*, \Omega)}. \quad (\text{A.6})$$

All that needs to be done is show that BF_{ijk} simplifies into the expression given by (11).

Consider the numerator of (A.6), which is $f(\mathbf{d}_{jk} | \gamma_{ijk}^* = 1, B_{(-i)jk}^*, \Omega) = \int f(\mathbf{d}_{jk} | B_{ijk}^*, B_{(-i)jk}^*, \Omega) f(B_{ijk}^* | \gamma_{ijk}^* = 1) dB_{ijk}^*$. Given that $(\mathbf{d}_{jk} | B_{ijk}^*, B_{(-i)jk}^*, \Omega) \sim MVN(X_i B_{ijk}^* + X_{(-i)} B_{(-i)jk}^*, \Sigma_{jk})$ and $(B_{ijk}^* | \gamma_{ijk}^* = 1) \sim N(0, \tau_{ijk})$, some algebraic rearrangements and simplifications followed by the integration with respect to B_{ijk}^* reveals that $(\mathbf{d}_{jk} | \gamma_{ijk}^* = 1, B_{(-i)jk}^*, \Omega) \sim MVN(X_{(-i)} B_{(-i)jk}^*, \Sigma_{jk} + X_i X_i' \tau_{ijk})$, or equivalently $(\mathbf{d}_{jk}^* | \gamma_{ijk}^* = 1, B_{(-i)jk}^*, \Omega) \sim MVN(0, \Sigma_{jk} + X_i X_i' \tau_{ijk})$. It is trivial to show that $f(\mathbf{d}_{jk} | \gamma_{ijk}^* = 0, B_{(-i)jk}^*, \Omega)$ in the denominator of (A.6) is a $MVN(X_{(-i)} B_{(-i)jk}^*, \Sigma_{jk})$ density. Thus, we can write the conditional Bayes Factor BF_{ijk} as

$$BF_{ijk} = \frac{|\Sigma_{jk} + X_i X_i' \tau_{ijk}|^{-1/2}}{|\Sigma_{jk}|^{-1/2}} \exp[-1/2 (\mathbf{d}_{jk}^*)' \{(\Sigma_{jk} + X_i X_i' \tau_{ijk})^{-1} - (\Sigma_{jk})^{-1}\} \mathbf{d}_{jk}^*]. \quad (\text{A.7})$$

Consider the first part of (A.7). Multiplying the numerator and denominator by $|\Sigma_{jk}^{-1}|^{-1/2}$, this simplifies to $|I_N + \tau_{ijk} X_i X_i' \Sigma_{jk}^{-1}|^{-1/2}$, where I_N is an N -by- N identity matrix, and recall N is the number of observed functions. By the properties of determinants, we can rewrite this as the scalar quantity $(1 + \tau_{ijk} X_i' \Sigma_{jk}^{-1} X_i)^{-1/2}$, which is the first part of (11).

Now consider the exponent in (A.7). Using the well-known identity $\Sigma_1^{-1} - \Sigma_0^{-1} = -\Sigma_0^{-1} u v' \Sigma_0^{-1} / (1 + v' \Sigma_0^{-1} u)$ that holds whenever $\Sigma_1 = \Sigma_0 + u v'$, we can rewrite this expression and perform a series of simplifications

$$\begin{aligned} &= \exp[\tau_{ijk}/2 * (1 + \tau_{ijk} X_i' \Sigma_{jk}^{-1} X_i)^{-1} \{(\mathbf{d}_{jk}^*)' (\Sigma_{jk}^{-1} X_i X_i' \Sigma_{jk}^{-1}) (\mathbf{d}_{jk}^*)\}] \\ &= \exp[\tau_{ijk}/2 * \{(X_i' \Sigma_{jk}^{-1} X_i)^{-1} ((X_i' \Sigma_{jk}^{-1} X_i)' + \tau_{ijk})^{-1} (\mathbf{d}_{jk}^*)' \Sigma_{jk}^{-1} X_i X_i' \Sigma_{jk}^{-1} (\mathbf{d}_{jk}^*)\}] \end{aligned}$$

$$\begin{aligned}
&= \exp \left[1/2 * \frac{(\mathbf{d}_{jk}^*)' \Sigma_{jk}^{-1} X_i (X_i' \Sigma_{jk}^{-1} X_i)^{-1} (X_i' \Sigma_{jk}^{-1} X_i)^{-1} X_i' \Sigma_{jk}^{-1} (\mathbf{d}_{jk}^*) \tau_{ijk}}{(X_i' \Sigma_{jk}^{-1} X_i)^{-1} \{\tau_{ijk} + (X_i' \Sigma_{jk}^{-1} X_i)\}} \right] \\
&= \exp[1/2 * \{(\hat{B}_{ijk,MLE}^*)^2 / V_{ijk}\} \{1 + V_{ijk} / \tau_{ijk}\}^{-1}],
\end{aligned}$$

which, by letting $\zeta_{ijk} = \hat{B}_{ijk,MLE}^* / \sqrt{V_{ijk}}$, gives us the second part of (11).

10 References

- Abramovich, F. and Angelini, C. (2003) Testing in mixed-effects FANOVA models. *Technical Report RP SOR-03-03*, Department of Statistics and Operations Research, Tel Aviv University.
- Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, **60**, 725–749.
- Antoniadis, A. and Sapatinas, T. (2004) Estimation and inference in functional mixed-effects models. *Technical Report TR-15-2004*, Department of Mathematics and Statistics, University of Cyprus, Cyprus.
- Brumback, B. A. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961–976.
- Clyde, M. and George, E. I. (2000) Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society, Series B*, **60**, 681–698.
- Clyde, M. and George, E. I. (2003) Discussion of “Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis”. *Journal of the American Statistical Association*, **98**, 584–585.
- Daubechies, I. (1992) *Ten Lectures on Wavelets*, Philadelphia: Society for Industrial and Applied Mathematics.
- Dawid, A. P. (1981) Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, **68**, 265–274.
- Donoho, D. and Johnstone, I. M. (1995) Adapting to unknown smoothness by wavelet shrinkage. *Journal of the American Statistical Association*, **90**, 1200–1224.
- Fan, J. and Lin, S. K. (1998) Tests of significance when data are curves. *Journal of the American Statistical Association*, **93**, 1007–1021.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis.*, 2nd edition, Chapman and Hall, New York.
- Gortmaker, S., Peterson, K., Wiecha, J., Sobol, A., Dixit, S., Fox, M. and Laird, N. (1999) Reducing obesity via a school-based interdisciplinary intervention among youth: Planet health. *Archives of Pediatrics and Adolescent Medicine*, **153**, 409–418.
- Grambsch, P. M., Randall, B. L., Bostick, R. M., Potter, J. D. and Louis, T. A. (1995) Modeling the labeling index distribution: An application of functional data analysis. *Journal of the American Statistical Association*, **90**, 813–821.
- Guo, W. (2002) Functional mixed effects models. *Biometrics*, **58**, 121–128.
- Harville, D. (1997) *Matrix Algebra from a Statistician’s Perspective*, Springer, New York.
- Hong, M. Y., Lupton, J. R., Morris, J. S., Wang, N., Carroll, R. J., Davidson, L. A., Elder, R. and Chapkin, R. S. (2000) Dietary fish oil reduces O⁶-methylguanine DNA adduct levels in the rat colon in part by increasing apoptosis during tumor initiation. *Cancer Epidemiology, Biomarkers and Prevention*, **9**, 819–826.

- Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, **59**, 319–351.
- Laird, N. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
- Liang, H., Wu, H., and Carroll, R. J. (2003) The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics*, **4**, 297–312.
- Mallat, S. G. (1989) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.
- Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003a) Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, **98**, 573–583.
- Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003b) Rejoinder to “Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis”. *Journal of the American Statistical Association*, **98**, 591–597.
- Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*. Springer, New York.
- Rice, J. A., and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B*, **53**, 233–243.
- Rice, J. A., and Wu, C. O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.
- Shi, M., Weiss, R. E. and Taylor, J. M. G. (1996) An analysis of pediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics*, **45**, 151–163.
- Staniswalis, J. G., and Lee, J. J. (1998) Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, **93**, 1403–1418.
- Vannucci, M. and Corradi, F. (1999) Covariance structure of wavelet coefficients: Theory and models in a Bayesian perspective. *Journal of the Royal Statistical Society, Series B*, **61**, 971–986.
- Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. Springer, New York.
- Wang, Y. (1998) Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, **60**, 159–174.
- Wolfinger, R., Tobias, R. and Sall, J. (1994) Computing Gaussian Likelihoods and their Derivatives for general linear mixed models. *SIAM Journal of Scientific Computing*, **15**, 1294–1310.
- Wu, H. and Zhang, J. T. (2002) Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, **97**, 883–897.
- Wu, H. and Liang, H. (2004) Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scandinavian Journal of Statistics*, **31**, 3–20.
- Zhang, D., Lin X., Raz J. and Sowers, M. F. (1998) Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, **93**, 710–719.

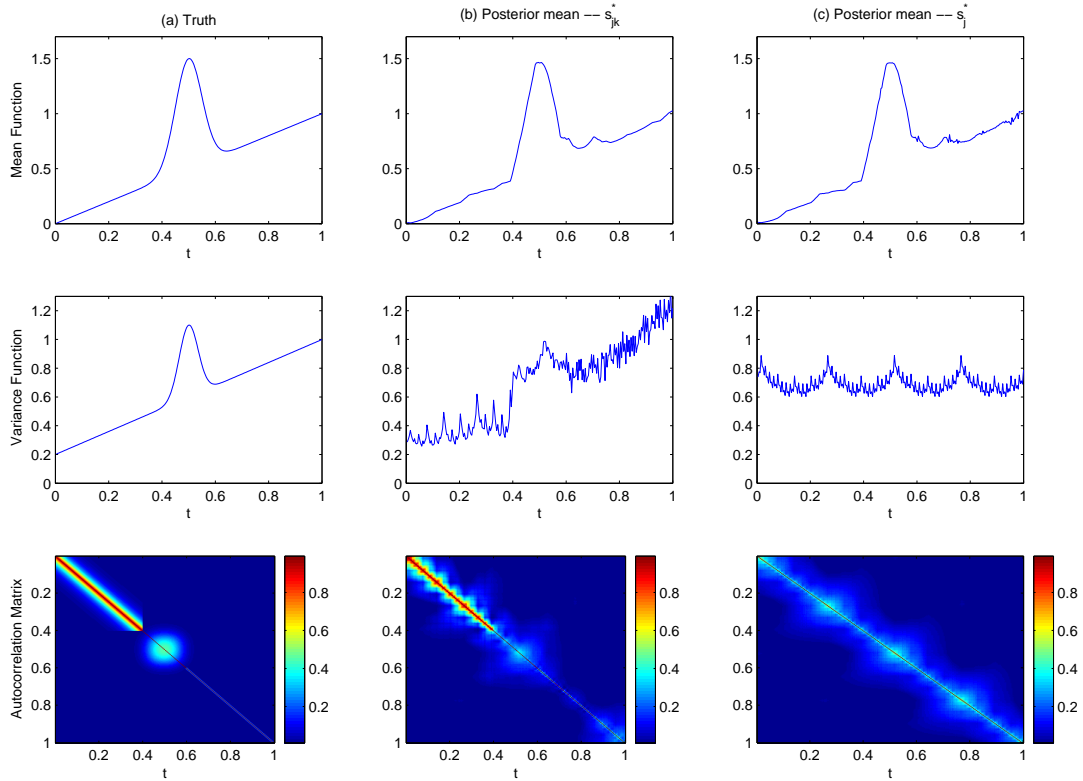


Figure 1: *Simulated Data*. We randomly generated 200 realizations from a Gaussian process with mean $\mu(t)$ and covariance $S(t_1, t_2)$ on an equally-spaced grid of length 256 on $(0, 1)$. From top to bottom, column (a) contains the true mean function $\mu(t)$, the true variance function $v(t) = \text{diag}(S)$, and the true autocorrelation surface $\rho_S(t_1, t_2) = v^{-1/2} S v^{-1/2}$. Columns (b) and (c) contain the posterior mean estimates of these quantities using wavelet-based methods. Both assume independence across wavelet coefficients, but (b) allows the wavelet-space variance components to vary across scale j and location k , and (c) only allows them to vary across j , as assumed in Morris, et al. (2003a) and other work involving wavelet regression. Note that the framework used in (b) is sufficiently flexible to pick up on the nonstationary features of S , while (c) is not. Specifically, it is able to model the increasing variance in t , the extra variance near the peak at 0.5, the different degrees of smoothness in the region $(0, 0.4)$ and $(0.6, 1)$, and the extra autocorrelation from the peak at 0.5. Also note it appears to have done a marginally better job of denoising the estimate of the mean function. These same principles apply to the covariance across random effect functions.

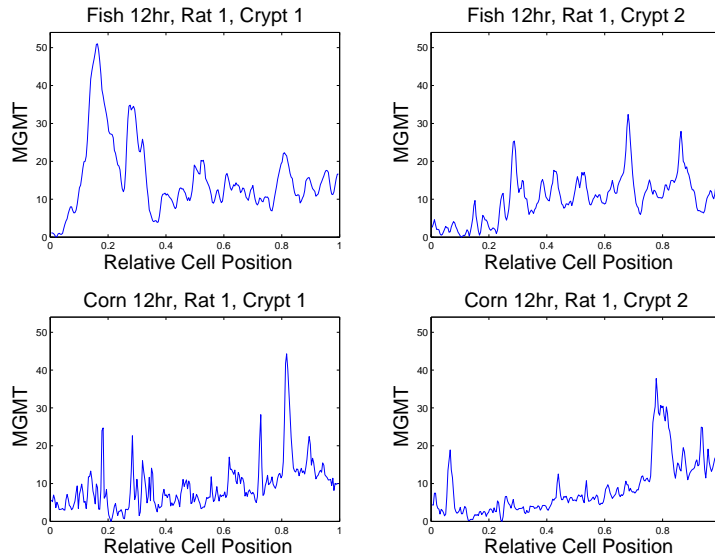


Figure 2: *Sample Curves*. Sample curves of MGMT intensity levels as a function of relative depth within the crypts. Two crypts from each of two rats are shown here.

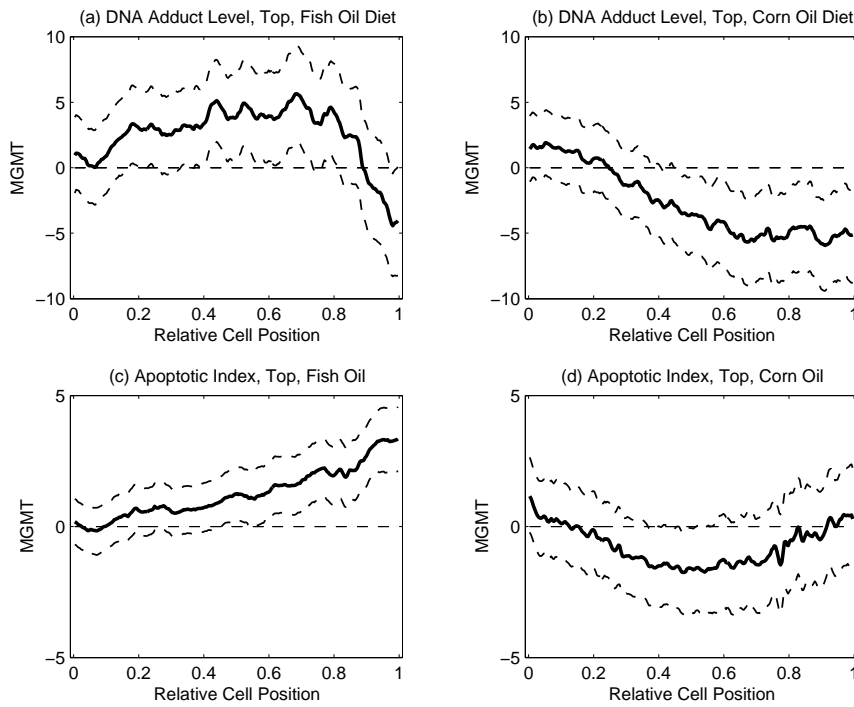


Figure 3: *MGMT Results*. Posterior mean and 95% pointwise posterior credible intervals for functional linear coefficients relating MGMT levels and (a) DNA adduct level, top 1/3 of crypt, fish oil diet, (b) DNA adduct level, top 1/3 of crypt, corn oil diet, (c) Apoptotic index, top 1/3 of crypt, fish oil diet, (d) Apoptotic index, top 1/3 of crypt, corn oil diet. These are functional linear coefficients for the corresponding continuous covariates in a functional mixed model that also includes categorical effects for the 10 diet \times time combinations and random effect functions for each rat.