Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis

Jeffrey S. MORRIS, Marina VANNUCCI, Philip J. BROWN, and Raymond J. CARROLL

In this article we develop new methods for analyzing the data from an experiment using rodent models to investigate the effect of type of dietary fat on O^6-methylguanine-DNA-methyltransferase (MGMT), an important biomarker in early colon carcinogenesis. The data consist of observed profiles over a spatial variable contained within a two-stage hierarchy, a structure that we dub *hierarchical functional data*. We present a new method providing a unified framework for modeling these data, simultaneously yielding estimates and posterior samples for mean, individual, and subsample-level profiles, as well as covariance parameters at the various hierarchical levels. Our method is nonparametric in that it does not require the prespecification of parametric forms for the functions and involves modeling in the wavelet space, which is especially effective for spatially heterogeneous functions as encountered in the MGMT data. Our approach is Bayesian; the only informative hyperparameters in our model are effectively smoothing parameters. Analysis of this dataset yields interesting new insights into how MGMT operates in early colon carcinogenesis, and how this may depend on diet. Our method is general, so it can be applied to other settings where hierarchical functional data are encountered.

KEY WORDS: Bayesian method; Carcinogenesis; Functional data analysis; Hierarchical model; Model averaging; Nonparametric regression; Wavelet.

1. INTRODUCTION

1.1 Colon Carcinogenesis Studies

Colorectal cancer is a major international health problem. It is the third most common cancer worldwide and the second-leading cause of cancer deaths in the United States. Because colon cancer is often asymptomatic until it is advanced, and current treatment of advanced disease has limited effectiveness, the development of preventive approaches is crucial in fighting this disease. A large part of this effort entails gaining a better understanding of the mechanisms underlying colon carcinogenesis, identifying important risk factors, and understanding how they operate.

There are indications that environmental factors (most notably, diet) play a primary role in the development of colon cancer (see, e.g., Giovannucci and Willet 1994). Carcinogen-induced colon cancer in rodent models are extensively used to delineate mechanisms in colon carcinogenesis. In these models, rodents are fed particular diets of interest for a specific period, exposed to a carcinogen known to induce colon cancer, and then later euthanized, with their colons removed and examined for carcinogenic responses.

In modeling biological mechanisms in the colon, it is important to consider the special architecture of cells within the colon. Colon cells replicate and spend their entire life cycles within *crypts*, fingerlike structures that grow into the wall of the colon. An individual cell is "born" in a region known as the stem cell region, toward the bottom of the crypt, and moves up the crypt wall as it matures and differentiates, until it is finally exfoliated out into the lumen at the end of its natural life cycle. This special cell-life sequence in the crypts suggests two important facts. First, cells at the same relative depths within different crypts will tend to share common biological characteristics. Second, cells at different depths of a given crypt are at different stages of maturity and could in principle react differently to carcinogens and other stimuli. As a result, it is important to study biological measurements in the colon as a function of relative cell position, because averaging over all crypt cells obscures any potential depth-specific effects. The relative cell position t is defined such that the bottom of each crypt has $t = 0$ and the top has $t = 1$, with positions in between coded proportionally.

1.2 Application

Epidemiologic and animal studies have suggested that diets high in fish oil fats, or n-3 polyunsaturated fatty acids, have a protective effect against colon cancer when compared with diets high in corn oil fats (e.g., Boyle, Zaridze, and Smans 1985). However, the biological mechanisms behind this observed effect remain unknown and are of considerable interest. Nutrition researchers at Texas A&M University (Hong et al. 2001) investigated how dietary fat type affects the initiation stage of colon carcinogenesis (the first few hours after exposure to a carcinogen). During this stage, the carcinogen exposure leads to damage to the cells' DNA, which, if not either repaired or removed, may eventually lead to cancer. The biological responses of interest in this study include DNA adduct levels, which quantify the amount of carcinogen-induced damage; O^6-methylguanine-DNA-methyltransferase (MGMT), measuring the amount of an enzyme that can repair this damage; apoptosis, the elimination of damaged cells; and BCL2, a protein related to apoptosis.

In this study, 30 rats were randomized to a diet high in either fish oil or corn oil. After being fed these diets for 2 weeks, each rat was exposed to the carcinogen azoxymethane (AOM)

Jeffrey S. Morris is Assistant Professor, Department of Biostatistics, University of Texas M.D. Anderson Cancer Center, Houston TX 77030 (E-mail: *jeffmo@mdanderson.org*). Marina Vannucci is Assistant Professor, Department of Statistics, Texas A&M University, College Station TX 77843 (E-mail: *mvannucci@stat.tamu.edu*). Philip J. Brown is the Pfizer Professor of Medical Statistics, Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2-7NZ, U.K. (E-mail: *Philip.J.Brown@ukc.ac.uk*). Raymond J. Carroll is Distinguished Professor, Department of Statistics and Department of Epidemiology and Biostatistics, Texas A&M University, College Station TX 77843 (E-mail: *carroll@stat.tamu.edu*). Vannucci's research was supported by National Science Foundation CAREER award DMS-0093208. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030), and by the Texas A&M Center for Environmental and Rural Health (CERH) via a grant from the National Institute of Health (P30-ES09106). The first author thanks Phil Brown for his hospitality while visiting the University of Kent at Canterbury under EPSRC Visiting Fellowship grant GR/R52770/01. The authors extend sincere thanks to Joanne Lupton, Robert Chapkin, Nancy Turner, and Mee Young Hong, who have helped invaluably with substantive issues involved with this work. They also thank the editor and referees, whose insightful comments and questions have dramatically improved this article.

Journal of the American Statistical Association
???? 2003, Vol. 0, No. 0, Applications and Case Studies
DOI 10.1198/00

and euthanized at one of five randomly chosen time points: 0, 3, 6, 9, or 12 hours after exposure to the carcinogen. For each response, 20–25 crypts were selected from each rat, and the response was quantified at fixed units of distance (*pixel*) along the left side of each selected crypt using immunhistochemical staining. The corresponding relative cell position for each pixel was computed as the ratio of the number of pixels closer to the crypt base and the total number of pixels in the crypt. These data have a natural hierarchical structure (treatment groups, rats within treatment, crypts within rats, and pixels within crypts), and the data for each crypt can be viewed as a single longitudinal profile over the spatial variable relative cell position.

Many important scientific questions can be answered using these data. For this article we focus on examining the MGMT response. We would like to obtain estimates of the mean MGMT profiles for each diet/time group as a function of relative cell position, from which we can perform various inferences to detect whether the MGMT expression profiles differ by diet or across time and whether the MGMT levels are higher at particular depths within the crypts. Second, we would like to obtain estimates of the variability at the various hierarchical levels, allowing us to determine whether the diets differ in the patterns of variability in MGMT expression and providing valuable information for efficiently designing future studies. Third, we would like to obtain estimates of the rat- and crypt-level profiles, to provide assurance that any diet/time level results are not being driven by outliers. We would like to answer these questions in a unified fashion using a single model, if possible.

1.3 Hierarchical Functional Data

The data that we have described consist of longitudinal profiles for each crypt, contained within a larger hierarchy including rats and treatment groups. We refer to data with this structure as *hierarchical functional data*. Here we present a unified model for hierarchical functional data that is essentially a hierarchical multilevel random-effects model, but with the base level of the hierarchy being noisy functions observed on a grid of spatial or time points:

$$\begin{aligned} \mathbf{y}_{abc} &= g_{abc}(\mathbf{t}) + \boldsymbol{\epsilon}_{abc}, \\ g_{abc}(\mathbf{t}) &= g_{ab}(\mathbf{t}) + \eta_{abc}(\mathbf{t}), \\ g_{ab}(\mathbf{t}) &= g_a(\mathbf{t}) + \xi_{ab}(\mathbf{t}). \end{aligned} \tag{1}$$

Here $\mathbf{y}_{abc}$ is the observed vector of responses for crypt c ($c = 1, \ldots, C_{ab}$) from rat b ($b = 1, \ldots, B_a$) of treatment group a ($a = 1, \ldots, A$), corresponding to a grid of cell positions $\mathbf{t}$ of length n. The functions $g_{abc}(\cdot)$, $g_{ab}(\cdot)$, and $g_a(\cdot)$ represent true underlying functions of cell position for an individual crypt, rat, and treatment level. The terms $\boldsymbol{\epsilon}_{abc}$ are the measurement errors, assumed to be $N(\mathbf{0}, \sigma_\epsilon^2 I_n)$, and $\eta_{abc}(\mathbf{t})$ and $\xi_{ab}(\mathbf{t})$ are crypt-specific and rat-specific errors, assumed to be mean-0 Gaussian random variables with covariance matrices $\Sigma_1(\mathbf{t})$ and $\Sigma_2(\mathbf{t})$, with $\boldsymbol{\epsilon}_{abc}$, $\eta_{abc}(\mathbf{t})$, and $\xi_{ab}(\mathbf{t})$ all mutually independent. Note that with their form left unstructured, the matrices $\boldsymbol{\Sigma}_1(\mathbf{t})$ and $\boldsymbol{\Sigma}_2(\mathbf{t})$ each contain $n(n+1)/2$ unique parameters. Clearly, further assumptions must be made before this model can be fit.

If one is willing to assume a parametric structure on the form of the g functions, then this model can be fit using standard mixed-model methodology. However, some responses in carcinogenesis, including MGMT, give rise to irregular functions that are not well represented by a parametric model. Figure 1 shows the observed MGMT functions for selected crypts. Note the spike-like features in these observed profiles. Clearly, a model that is nonparametric in t is needed to represent these data.

A number of methods in the literature deal with nonparametric estimation of replicated functions (e.g., Rice and Wu 2001; Fan and Zhang 2000), but most limit their scope to the single-level hierarchical setting and use kernels or splines, which can have difficulty modeling curves that change rapidly in t, such as the MGMT data. New methods are needed to model these data effectively. Wavelet regression is a nonparametric regression technique that is particularly effective in estimating functions that exhibit this type of local behavior. But most of the published work in wavelet regression is limited to the single function setting, and so cannot be directly applied to our setting. The work of Brown, Fearn, and Vannucci (2001) involves modeling sets of curves resulting from near-infrared spectroscopy, but in their problem the functions are predictors in a multivariate regression setting.

In this article we present a unified framework for fitting hierarchical functional data. Our modeling is performed in the wavelet space, which is useful for modeling spatially heterogeneous functions like those encountered here. Our approach is nonparametric and Bayesian, with the only informative hyperparameters being effective smoothing parameters. Our method makes full use of the data and simultaneously yields estimates and posterior samples for profiles at all hierarchical levels, as well as all variance components.

The remainder of this article is organized as follows. In Section 2 we provide an introduction to wavelets. In Section 3 we describe our method, presenting our wavelet space model and describing how we fit the model using a computationally efficient marginalized Metropolis-within-Gibbs sampler. In Section 4 we discuss prior selection issues for the variance components and the "smoothing" parameters for the nonparametric procedure. We apply this method to the MGMT data in Section 5, and conclude with a discussion of the strengths, weaknesses, and general applicability of our method in Section 6.

2. WAVELETS AND WAVELET REGRESSION

Wavelets are orthogonal families of basis functions that can be used to accurately and parsimoniously represent other functions. In $\mathrm{L}^2(\Re)$, for example, an orthogonal wavelet basis is obtained as translations and dilations of a *mother wavelet* ψ as $\psi^{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k)$ with j, k integers. A function g is then represented by a wavelet series as $g(x) = \sum_{j,k\in\Im} d^{j,k}\psi^{j,k}(x)$, with wavelet coefficients $d^{j,k} = \int g(x) \times \psi^{j,k}(x)\,dx$ describing features of the function g at spatial locations indicated by k and frequencies indicated by j. In usual wavelet notation, j and k are subscripts. In our notation, we use superscripts to indicate the wavelet levels and reserve subscripts to specify the hierarchical levels.

Let $\mathbf{Y} = (y_1, \ldots, y_n)$ be a sample of a function at n equally spaced points, with n assumed to be a power of 2. A fast algorithm, the *discrete wavelet transform* (DWT), exists for decomposing $\mathbf{Y}$ into a set of n wavelet coefficients (Mallat 1989) in only $O(n)$ operations. In matrix form, we can represent the

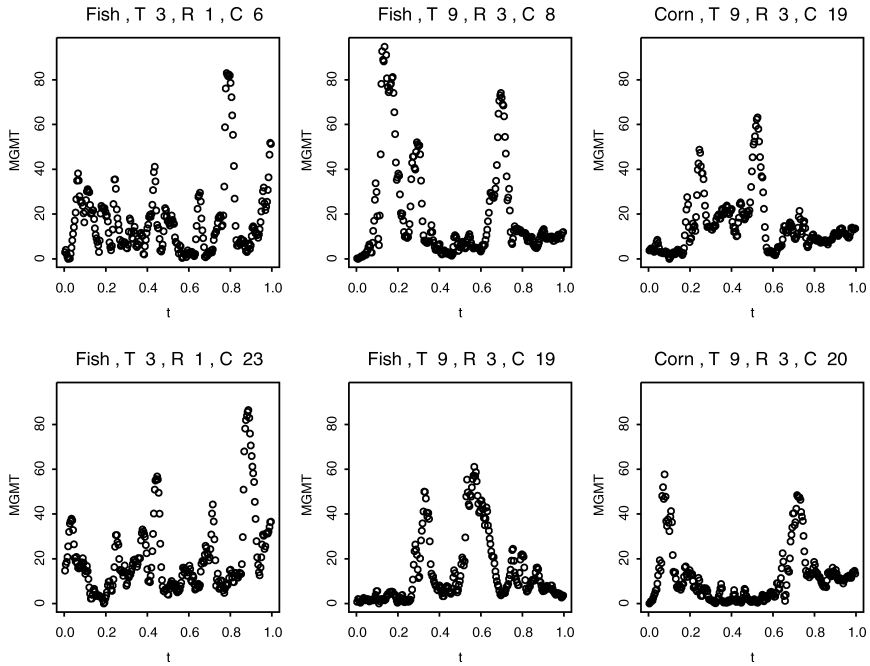

Figure 1. MGMT Levels as a Function of Relative Cell Position (t) for Some Crypts From Selected Rats. The plot headers indicate the diet and time group, rat number, and crypt number. For example, "fish T3 R1 C6" indicates that the plot is for crypt number 6 from rat 1 from the group fed fish oil diets and euthanized 3 hours after exposure to the carcinogen. The full dataset contains roughly 750 such crypts.

DWT through an orthogonal matrix $W = [W_1, W_2, \ldots, W_J, V_J]$ applied to the vector Y of observations as $d = WY$ that decomposes the data into sets of wavelet coefficients, $d = [d_1^T, d_2^T, \ldots, d_J^T, c_J^T]^T$, with $d_j = W_j^T \mathbf{Y}$ and $c_J = V_J^T \mathbf{Y}$. At scale 2^{j-1} or level j, the number of coefficients is given by

$$K_j = n/2^j. \tag{2}$$

An algorithm for the inverse reconstruction (the IDWT) also exists.

Wavelets can be used to perform nonparametric regression. Suppose that we have a response vector $\mathbf{y}$ of length n observed on an equally spaced grid $\mathbf{t}$, assumed without loss of generality to be on $(0, 1)$. The observed response $\mathbf{y}$ is assumed to be a function of t plus some additive white noise; that is, with $\boldsymbol{\epsilon} \sim MVN(0, \sigma_e^2 I_n)$,

$$\mathbf{y} = g(\mathbf{t}) + \boldsymbol{\epsilon}. \tag{3}$$

The aim of nonparametric regression is estimation of the function $g(t)$ without imposing any particular parametric structure on its form. There are numerous methods available to accomplish this, and all involve the choice of a regularization parameter, which determines the trade-off between bias and variance in the resulting estimator.

Wavelet regression follows three basic steps. First, the data are converted to the wavelet space using the DWT. The wavelet space model corresponding to (3) is given by

$$\mathbf{d} = \boldsymbol{\theta} + \boldsymbol{\epsilon}^*, \tag{4}$$

where $\mathbf{d} = \mathbf{W}\mathbf{y}$ are the empirical wavelet coefficients, $\boldsymbol{\theta} = \mathbf{W}g(\mathbf{t})$ are the true function's wavelet coefficients, and $\boldsymbol{\epsilon}^* = \mathbf{W}\boldsymbol{\epsilon} \sim MVN(0, \sigma_e^2 I_n)$ is the noise in the wavelet space.

Next, the true wavelet coefficients are estimated using a thresholding or shrinkage rule, whereby a coefficient may be set to 0 or shrunk toward 0, depending on the magnitude of its empirical coefficient. The thresholding or shrinkage can be done using frequentist rules (e.g., Donoho and Johnstone 1994) or Bayesian hierarchical models (e.g., Clyde, Parmigiani, and Vidakovic 1998). This estimation technique effectively removes the noise from the signal, and so after the third step, whereby the true wavelet coefficient estimates are converted back into the data space using the IDWT, we end up with a denoised, or regularized, nonparametric estimate of the function.

3. OVERVIEW OF METHOD

Implementation of our method involves three basic steps:

1. Perform a DWT on the response vector for each crypt, $\mathbf{y}_{abc}$, to obtain the corresponding empirical wavelet coefficients.
2. Fit a Bayesian multilevel hierarchical model to these empirical wavelet coefficients, yielding posterior samples of the true wavelet coefficients corresponding to the treatment, rat and crypt-level profiles, as well as the variance components.
3. Using inverse discrete wavelet transforms, obtain the posterior distribution of the mean diet/time profiles $g_a(t)$, as well as any rat- or crypt-level profiles on a fixed grid of t, which can then be used to perform various Bayesian inferences.

3.1 Wavelet Space Model

We specify a version of model (1) in the wavelet space. The data are the scaling coefficients, $c_{abc}^{J,k}$, and empirical wavelet coefficients, $d_{abc}^{j,k}$, for $a = 1, \ldots, A$, $b = 1, \ldots, B_a$, $c = 1, \ldots, C_{ab}$, obtained by performing a DWT on the vectors $\mathbf{y}_{abc}$, response vectors of length n, a power of 2. Recall that $j = 1, \ldots, J$ is the scale or wavelet resolution level, and $k = 1, \ldots, K_j = n/2^j$ specifies the location of the wavelet coefficient within level j.

Let σ_ϵ^2 and $\Omega = \{\Omega^j\} = (\sigma_{1,j}^2, \sigma_{2,j}^2)$ be the set of variance components in the model

$$d_{abc}^{j,k} = \text{Normal}(\Theta_{abc}^{j,k}, \sigma_\epsilon^2), \tag{5}$$

$$\Theta_{abc}^{j,k} = \text{Normal}(\Theta_{ab}^{j,k}, \sigma_{1,j}^2), \tag{6}$$

and

$$\Theta_{ab}^{j,k} = \text{Normal}(\Theta_{a}^{j,k}, \sigma_{2,j}^2). \tag{7}$$

An analogous model is defined for the observed scaling coefficients $c_{abc}^{J,k}$, with true mean coefficients $\omega_{abc}^{J,k}$, $\omega_{ab}^{J,k}$, and $\omega_a^{J,k}$ and variance components $\nu_{1,J}^2$ and $\nu_{2,J}^2$.

We assume that σ_ϵ is known. Following Donoho and Johnstone (1994), in practice we plug in a robust estimator of the standard deviation of all wavelet coefficients at finest resolution level for σ_ϵ, because these coefficients are expected to consist almost entirely of noise. Specifically, we use the median absolute deviation over all crypts divided by .6745 (see Welsh 1996, p. 107).

When unknown, the $\sigma_{1,j}^2$ and $\sigma_{2,j}^2$ are assumed to have proper inverse gamma priors with hyperparameters $(\mathcal{A}_{1,j}, \mathcal{B}_{1,j})$ and $(\mathcal{A}_{2,j}, \mathcal{B}_{2,j})$. The inverse gamma parameterization that we use is detailed in the Appendix.

The model (5)–(7) is essentially a standard two-level random-effects model for each wavelet coefficient, with random effects at the crypt and rat levels. To obtain denoised or regularized estimates of the diet/time profiles, as is done in single-function wavelet regression, we place the following shrinkage prior on the wavelet coefficients $\Theta_a^{j,k}$:

$$\Theta_a^{j,k} = \text{Normal}(0, \gamma_a^{j,k}\tau_j^2) \tag{8}$$

and

$$\gamma_a^{j,k} = \text{Bernoulli}(p_j). \tag{9}$$

The hyperparameters τ_j^2 and p_j simply determine the amount of regularization done in estimating the mean profiles $g_a(t)$, and so are like smoothing parameters. More discussion regarding the choice and effect of these priors and how they result in regularized profile estimates is given in Section 4.2.

3.2 Evaluation of Relative Variability

We choose the trace to summarize the variability at the various hierarchical levels in the data space model (1), so $V_0 = \text{trace}(\sigma_e^2 I)$, $V_1 = \text{trace}(\Sigma_1)$, and $V_2 = \text{trace}(\Sigma_2)$ summarize the amount of variability within crypts, between crypts, and between rats in our rat colon carcinogenesis example. The relative variability at level i is then defined as $V_i/(V_0 + V_1 + V_2)$. This measure gives a rough assessment of the amount of variability at the different hierarchical levels, and estimates of these quantities can aid the design of future experiments, in that larger sample sizes can be used at the hierarchical levels with the most variability.

Because the DWT is an orthogonal transformation, these quantities are the same in the data space and the wavelet space. Using the variance components in (5)–(7), these quantities are easily calculated to be $V_0 = n\sigma_e^2$, $V_1 = \sum_j K_j\sigma_{1,j}^2 + K_J\nu_{1,J}^2$, and $V_2 = \sum_j K_j\sigma_{2,j}^2 + K_J\nu_{2,J}^2$, with K_j as defined in (2). The amount of variability within hierarchical level i that can be explained by wavelet resolution level j can be measured by $W_{i,j} = K_j\sigma_{i,j}^2/V_i$ and provides a frequency domain summary of the profiles at the various levels.

3.3 Fitting the Model When Variance Components Known

If the set of variance components, Ω, is known, then we have a closed-form solution for the posterior distribution of the diet/time level wavelet coefficients $\Theta_a^{j,k}$, from which we can compute the posterior distribution of the diet/time profiles $g_a(t)$. Let $\underline{d}_a^{j,k} = (d_{a11}^{j,k}, \ldots, d_{aB_aC_{ab}}^{j,k})^T$. For each a, j, and k independently, the posterior distribution $(\Theta_a^{j,k}|\underline{d}_a^{j,k}, \Omega^j)$ is a mixture of a point mass at 0 and a normal distribution with mean and variance given by (A.7) and (A.8) in the Appendix, with the mixture proportion for the normal given by (A.1) in the Appendix.

Samples from this mixture distribution are easily obtained using a Monte Carlo simulation. Applying the IDWT to the posterior sample vectors of wavelet coefficients for each treatment group a, $\mathbf{\Theta}_a = (\Theta_a^{1,1}, \ldots, \Theta_a^{J,K_J}, \omega_a^{J,1}, \ldots, \omega_a^{J,K_J})^T$, we obtain posterior samples of the treatment profiles $g_a(\mathbf{t})$ on the equally spaced grid $\mathbf{t}$.

3.4 Fitting the Model When Variance Components Unknown

In the more realistic situation where the variance components Ω are unknown and of interest, they must be estimated along with the mean parameters. This precludes closed-form expressions for our posterior distributions and requires the use of an iterative procedure. Here we propose a Metropolis-within-Gibbs MCMC sampler performed on a marginalized version of the model specified in (5)–(9), marginalized in the sense that the rat- and crypt-level random effects, $\Theta_{ab}^{j,k}$ and $\Theta_{abc}^{j,k}$, are integrated out.

Sampling is done as follows. Let $\mathbf{d}_a^{j,k} = (d_{a11}^{j,k}, \ldots, d_{aB_aC_{ab}}^{j,k})^T$, $\mathbf{d}^j = (d_{111}^{j,1}, \ldots, d_{AB_aC_{ab}}^{j,K_j})^T$, and $\mathbf{\Theta}_3^j = (\Theta_1^{j,1}, \ldots, \Theta_A^{j,K_j})^T$. Each iteration of the MCMC consists of the following two steps performed for each wavelet coefficient level $j = 1, \ldots, J$, with similar expressions for the scaling coefficients:

1. Generate $\Theta_a^{j,k}$ from $(\Theta_a^{j,k}|\mathbf{d}_a^{j,k}, \Omega^j)$ for each $a = 1, \ldots, A$ and $k = 1, \ldots, K_j$. (10)
2. Generate $\Omega^j = \{\sigma_{1,j}^2, \sigma_{2,j}^2\}$ from $(\Omega^j|\mathbf{d}^j, \mathbf{\Theta}_3^j)$. (11)

The distribution in (10) is simply the posterior distribution of $\Theta_a^{j,k}$ assuming the variance components are known, given in Section 3.3. The distribution in (11) is proportional to expression (A.10) in the Appendix, from which samples are drawn using a random-walk Metropolis step. These give us posterior samples of the $\Theta_a^{j,k}$ and the variance components $\sigma_{1,j}^2$ and $\sigma_{2,j}^2$. As described in Section 3.3, IDWTs can be subsequently used to obtain posterior samples of the diet/time profiles $g_a(\mathbf{t})$ on the fixed grid $\mathbf{t}$.

In principle, a simple Gibbs sampler without marginalization could be used to estimate these posterior distributions. Although much simpler analytically, this method had poor convergence properties, especially for the covariance parameters. Our marginalized Metropolis-within-Gibbs sampler resulted in better mixing.

3.5 Estimating Crypt- and Rat-Level Profiles

The procedure described in Section 3.4 for fitting our model gives only the posterior samples and estimates for the diet/time-level profiles and the variance components. If estimated profiles for individual rats and/or crypts are desired, then the following steps may be performed for each j and k using the samples from the MCMC in Section 3.4. Let $\mathbf{d}_{ab}^{j,k} = (d_{ab1}^{j,k}, \ldots, d_{abC_{ab}}^{j,k})^T$:

1. Generate $\Theta_{ab}^{j,k}$ from $(\Theta_{ab}^{j,k} | \mathbf{d}_{ab}^{j,k}, \Theta_a^{j,k}, \sigma_{2,j}^2)$ for each a, b. (12)
2. Generate $\Theta_{abc}^{j,k}$ from $(\Theta_{abc}^{j,k} | d_{abc}^{j,k}, \Theta_{ab}^{j,k}, \sigma_{1,j}^2)$ for each a, b, c. (13)

It can be shown using standard calculations that $(\Theta_{ab}^{j,k} | \underline{d}_{ab}^{j,k}, \Theta_a^{j,k}, \sigma_{2,j}^2)$ is normally distributed with mean $[C_{ab}\bar{d}_{ab}^{j,k}\sigma_{2,j}^2 + \Theta_a^{j,k}(\sigma_{1,j}^2 + \sigma_e^2)]/[C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_e^2]$ and variance $[\sigma_{2,j}^2 * (\sigma_{1,j}^2 + \sigma_e^2)]/[C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_e^2]$. Similarly, $(\Theta_{abc}^{j,k} | d_{abc}^{j,k}, \Theta_a b^{j,k}, \sigma_{1,j}^2)$ is normally distributed with mean $(d_{abc}^{j,k}\sigma_{1,j}^2 + \Theta_{ab}^{j,k}\sigma_e^2)/(\sigma_{1,j}^2 + \sigma_e^2)$ and variance $(\sigma_{1,j}^2\sigma_e^2)/(\sigma_{1,j}^2 + \sigma_e^2)$.

As previously described, applying the IDWT to the posterior samples of these wavelet coefficients yields posterior samples of the individual rat- and crypt-level profiles.

4. PRIOR SELECTION ISSUES

4.1 Choosing a Prior for the Variance Components

We avoid using improper priors for the variance components, because it is well known that this may lead to improper posteriors in this setting. However, it is not practical to elicit informative proper priors in this setting, because a priori it is difficult to know what the variation of wavelet coefficients at various hierarchical and resolution levels will be. Our approach is to use priors that are proper but contain little information relative to the data, and here we present a method based loosely on the data for choosing the inverse gamma hyperparameters.

Rather than directly specifying the hyperparameters $\mathcal{A}_{i,j}$ and $\mathcal{B}_{i,j}$, we specify $\mu_{i,j}$ and $\Delta_{i,j}$, a one-to-one reparameterization of $\mathcal{A}_{i,j}$ and $\mathcal{B}_{i,j}$ given the sample sizes of the study design. Here $\mu_{i,j}$ is the prior density's mode (i.e., the value of $\sigma_{i,j}^2$ considered most likely a priori) whereas $\Delta_{i,j}$ generally indicates the weight of the prior relative to the current dataset. A choice of $\Delta = 1$ indicates a prior whose influence on the posterior is roughly equal to the data, and $\Delta \ll 1$ and $\Delta \gg 1$ correspond to very vague and very informative priors. Based on these quantities, the inverse gamma parameters are set to $\mathcal{A}_{i,j} = \Delta_{i,j} N_{i,j}/2$ and $\mathcal{B}_{i,j} = \mu_{i,j}(\mathcal{A}_{i,j} + 1)$, where $N_{i,j}$ is the effective number of data points used in estimating the variance component at the ith hierarchical level and jth wavelet resolution level. In our example, $N_{1,j} = K_j \sum_{a=1}^{A} \sum_{b=1}^{B_a} C_{ab}$ and $N_{2,j} = K_j \sum_{a=1}^{A} B_a$, with K_j defined in (2).

For the prior modes, $\mu_{i,j}$, we plug in easily computed estimators of the $\sigma_{i,j}^2$ in an empirical Bayes-like fashion. We use ANOVA-type method-of-moment estimators computed under the assumption of no thresholding, that is, $p_j = 1$ and $\tau_j^2 \to \infty$ for all j. In our experience, these yield estimates that are close enough to the estimates from our model to be useful. For example, assuming equal sample sizes ($B_a = B$ and $C_{ab} = C$) and σ_e^2 known, our estimators are $\hat{\sigma}_{1,j}^2 = \mathrm{MSW}^j - \sigma_e^2$ and $\hat{\sigma}_{2,j}^2 = (\mathrm{MSB}^j - \mathrm{MSW}^j)/C$, with $\mathrm{MSW}^j = \mathrm{SSW}^j/\{K_j AB(C-1)\}$, $\mathrm{MSB}^j = \mathrm{SSB}^j/\{K_j A(B-1)\}$, $\mathrm{SSW}^j = \sum_a \sum_b \sum_c \sum_k (d_{abc}^{j,k} - \bar{d}_{ab}^{j,k})^2$, and $\mathrm{SSB}^j = \sum_a \sum_b \sum_k C_{ab}(\bar{d}_{ab}^{j,k} - \bar{d}_a^{j,k})^2$. A negative estimate of $\sigma_{i,j}^2$ can be replaced by a very small positive value ϵ.

4.2 Choosing the Regularization Parameters

The hyperparameters p_j and τ_j^2 regulate the shrinkage of the wavelet coefficients, and thus the smoothing or regularization done in nonparametrically estimating the $g_a(t)$. We now discuss the interpretation of these parameters and give some guidelines for choosing them.

To investigate these parameters' effect on regularization, we focus on the posterior means of the $\Theta_a^{j,k}$ in the known variance components case given by (A.7), with the understanding that the effect is similar when the variance components must be estimated. This posterior mean can be rewritten as

$$\widehat{\Theta}_a^{j,k} = \mathrm{E}(\Theta_a^{j,k} | \underline{d}_a^{j,k}, \Omega^j) = \widehat{\Theta}_{a,NS}^{j,k} \times h(T_j^2, p_j, Z_a^{j,k}) \quad (14)$$

and

$$h\left(T_j^2, p_j, Z_a^{j,k}\right) = \left(\frac{T_j^2}{T_j^2 + 1}\right)\left(\frac{O}{O+1}\right), \quad (15)$$

where $T_j^2 = \tau_j^2/\mathrm{var}(\widehat{\Theta}_{a,NS}^{j,k})$, $O = [p_j/(1-p_j)] \times (1+T_j^2)^{-1/2} \exp[(Z_a^{j,k})^2/2(1+T_j^{-2})^{-1}]$, and $Z_a^{j,k} = (\widehat{\Theta}_{a,NS}^{j,k})/\sqrt{\mathrm{var}(\widehat{\Theta}_{a,NS}^{j,k})}$.

$\widehat{\Theta}_{a,NS}^{j,k}$ is the maximum likelihood estimator for $\Theta_a^{j,k}$ assuming that the variance components are known and no thresholding is done (e.g., if $p_j = 1$), and is defined in Section A.1 in the Appendix along with $\mathrm{var}(\widehat{\Theta}_{a,NS}^{j,k})$. Our posterior mean estimators $\widehat{\Theta}_a^{j,k}$ in (14) are the product of this unshrunken estimator and the shrinkage factor h, which is a function of the regularization parameters and the data through the quantities T_j^2, p_j, and $Z_a^{j,k}$. This shrinkage factor has two components, one linear and the other nonlinear. The linear shrinkage component $T_j^2/(T_j^2+1)$ performs the usual Bayesian shrinkage toward the mean and acts equally on all coefficients at the jth resolution level. The nonlinear shrinkage component $O/(O+1)$ effectively performs Bayesian model averaging over models with and without $\Theta_a^{j,k}$ (i.e., $\gamma_a^{j,k}$ being 1 or 0), and operates differentially depending on the magnitude of $\widehat{\Theta}_{a,NS}^{j,k}$. The smaller the $\widehat{\Theta}_{a,NS}^{j,k}$, the more it is shrunk toward 0, with very large $\widehat{\Theta}_{a,NS}^{j,k}$ left largely unaffected by the nonlinear shrinkage. This differential shrinkage is the key to denoising the signal and thus regularizing the function estimates while retaining important features of the functions, because the small coefficients expected to consist of mostly noise are shrunk the most, whereas the larger coefficients consisting of mostly signal are left alone. The dependence on $\widehat{\Theta}_{a,NS}^{j,k}$ is through the quantity $Z_a^{j,k}$, which measures how many standard deviation units $\widehat{\Theta}_{a,NS}^{j,k}$ is from 0.

In single-function wavelet regression, it is useful to evaluate the shrinkage properties of a particular estimator by plotting the shrunken estimators of Θ versus the empirical coefficients d (e.g., Vidakovic 1999, chap. 6). In cases where no shrinkage

is done, these curves simply follow a 45-degree line. We introduce a similar curve for the hierarchical function case, which we call a *shrinkage curve*. A shrinkage curve plots the quantity Z against the shrinkage factor h for a particular choice of T_j^2 and p_j. The range of these curves is from 0 to 1, with 0 representing a completely thresholded coefficient and 1 representing a coefficient with no shrinkage done. Figure 2 contains the shrinkage curves for various choices of p_j and T_j^2, and graphically illustrates many of the points made in this section.

Our prior structure in (8)–(9) implies that when $\gamma_a^{j,k} = 0$, the wavelet coefficient $\Theta_a^{j,k}$ is considered negligible and effectively removed (thresholded) from the model. Thus the parameter p_j represents the expected proportion of wavelet coefficients at the jth resolution level that are nonnegligible. A smaller p_j results in more smoothing of any features of the function corresponding to that resolution level or frequency. Generally, it is expected that p_j should increase in j. In the single-function case, Abramovich, Sapatinas, and Silverman (1998) suggested a parameterization in which p_j increases exponentially in j.

One must take special care in choosing the parameters τ_j^2, which represent the expected variation of the nonnegligible wavelet coefficients. Very small τ_j^2 will result in much linear shrinkage, which affects even the large coefficients at level j, leading to more bias and possible distortion in estimating the function. Making τ_j^2 very large will make the linear shrinkage component negligible, but making if too large can lead to undesirable nonlinear shrinkage as a result of Lindley's paradox (see Lindley 1957; Kass and Raftery 1995). This corresponds to shrinkage curves whose shape is too steep (see Fig. 2). Because τ_j^2 enters into the shrinkage factor via T_j^2, its ratio with $\mathrm{var}(\widehat{\Theta}_{a,NS}^{j,k})$, any evaluation of the size of τ_j^2 must be made relative to $\mathrm{var}(\widehat{\Theta}_{a,NS}^{j,k})$, a function of the possibly unknown variance components and the sample sizes. When the variance components are unknown, the method-of-moments estimators discussed in Section 4.1 can be used as rough estimates. Thus one easy way to choose τ_j^2 is to first estimate $\mathrm{var}(\widehat{\Theta}_{a,NS}^{j,k})$ for each j, then elicit T_j^2, and let $\tau_j^2 = T_j^2 * \widehat{\mathrm{var}}(\widehat{\Theta}_{a,NS}^{j,k})$. We have found that moderate values of T_j^2 between 10 and 100 seem practical, avoiding the problems at either extreme. It is typical in wavelet regression to not perform any shrinkage at the smoothest resolution levels; for example, choosing $p \sim 1$ and $\tau^2 = 1/\epsilon$ for some $\epsilon \approx 0$ at the desired levels.

5. APPLICATION

5.1 Implementation of Method

We applied our method to the MGMT data described in Section 1.2. First, we used linear interpolation to obtain estimates for all crypt curves on a common grid $\mathbf{t}$, equally spaced of length $n = 2^8$. The interpolation should not have a substantial effect on the results, given the high sampling frequency and high signal-to-noise ratio present in these data.

We used the S-PLUS Wavelets package to perform all DWT and IDWT, and fit the hierarchical model using a C++ program written by us. We chose the Daubechies wavelet basis with four vanishing moments (Daubechies 1992). We allowed the variance components to differ across diets, and chose priors on them to be diffuse, more specifically inverse gammas with method-of-moments estimates as mode and weight .001

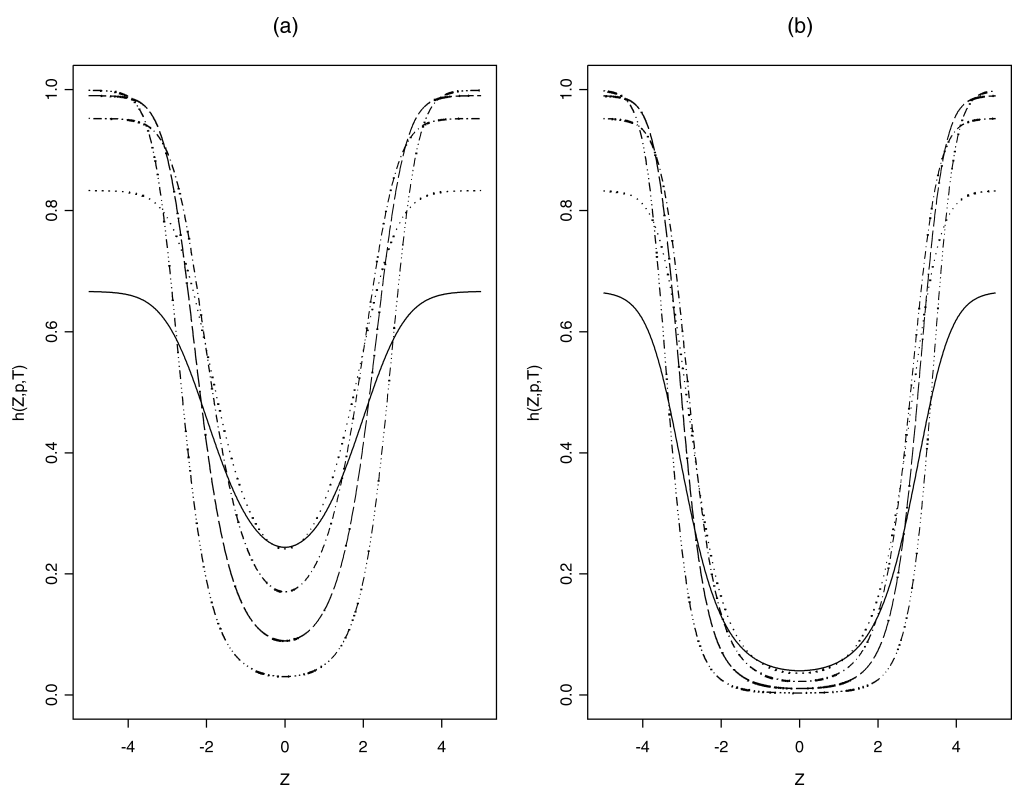

Figure 2. Shrinkage Functions Versus Z for Various Choices of T^2, for (a) $p = .50$ and (b) $p = .10$. (———) $T^2 = 2$; (·····) $T^2 = 5$; (·—·—·) $T^2 = 20$; (— —) $T^2 = 100$; (–···–) $T^2 = 1000$. Note how small choices of T^2 yield much shrinkage, even for large coefficients. Also note how the curve steepens as T^2 increases, corresponding to more nonlinear shrinkage for moderately large empirical coefficients.

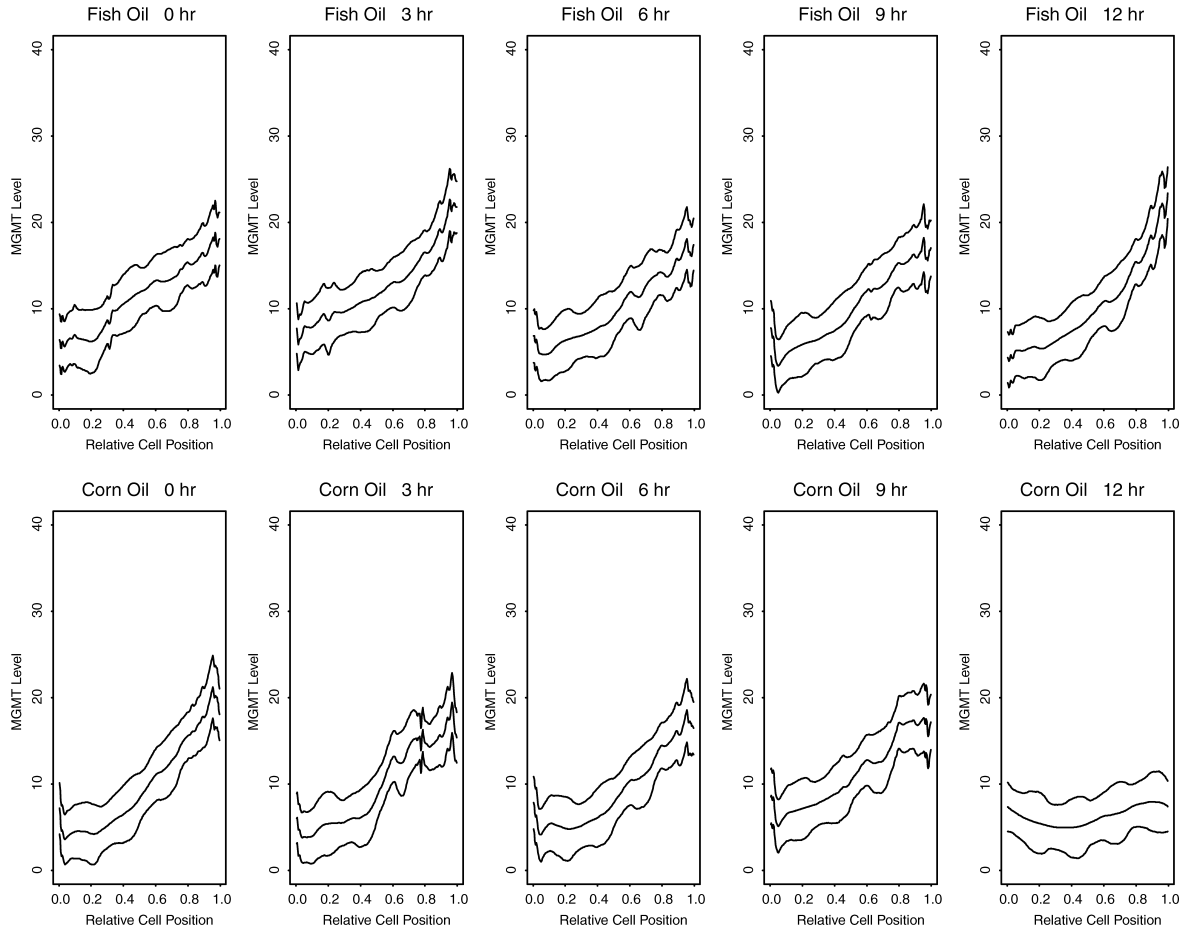

Figure 3. Posterior Mean MGMT Functions for Diet/Time Groups, Distal Colon, With 90% Posterior Pointwise Bounds.

relative to the data ($\Delta_{i,j} = .001$ for all i, j). No shrinkage was done on the scaling function coefficients. For the wavelet coefficients, we chose $\mathbf{p} = \{.005, .01, .015, .03, .07, .20\}$. We chose the τ_j^2 as described in Section 4.2, using $T_j^2 \equiv 20$. We fit the models using the marginalized Metropolis-within-Gibbs sampler described in Section 3.4. A sensitivity analysis on these prior choices is discussed in Section 5.3.

We ran parallel MCMC chains with diverse starting values; these converged to the same range of values for each parameter. Given the extremely large number of parameters in this problem, we did not do formal tests of MCMC convergence, but instead used iteration plots to assess convergence. All model parameters appeared to mix satisfactorily. The Metropolis acceptance probabilities were all between .16 and .47. The results presented here are from a single chain of 50,000 iterations, keeping every tenth one, after a burn-in of 5,000.

5.2 Results

Figure 3 presents the estimated MGMT mean profiles in the distal colon for each diet-by-time group, along with the corresponding 90% pointwise posterior credible intervals. First, we see that there is more MGMT expression toward the lumenal surface of the distal colon ($t \sim 1$) than toward the base of the crypts ($t \sim 0$). The posterior probability that $g_a(1) > g_a(0)$ is equal to 1.00 for all diet/time groups except corn oil (time 12 hours), where it is .73. Targeted apoptosis, another major mechanism for combatting the carcinogen-induced DNA damage, has been shown to be stronger at the bases of the crypts (Hong et al. 1999; Lantham, Lund, and Johnson 1999). This raises the possibility that apoptotic removal of damaged cells is the major mechanism of action at the bases of distal crypts, whereas DNA repair by MGMT may be a major factor at the lumenal surfaces.

We see little evidence of any diet or time effects from the time points 0–9 hours, but at time 12 hours the data suggest an interesting difference between the fish oil and corn oil diets at the tops of the crypts. Twelve hours after exposure to the carcinogen AOM, the fish oil–fed rats tend to have significantly higher MGMT expression levels than the corn oil–fed rats near the lumenal surface of their distal crypts (posterior probability, 1.00). This effect is especially interesting given that colon tumors form at the lumenal surface. One possible explanation for this effect is that after a period of time, the fish oil–fed rats' DNA repair mechanism at the lumenal surface continues to repair damaged cells, whereas somehow the corn oil–fed rats' repair response drops off. Future experiments will be designed to include the 12-hour and later time points to see whether this holds true.

To confirm that this diet effect is not driven by a single outlying rat, we estimated the rat-level profiles for all rats sacrificed at 12 hours (not shown), revealing that all three rats in the fish oil group still had notably higher MGMT levels at the tops of their crypts, whereas all three corn oil rats had less MGMT than

any of the fish oil rats at the top of their crypts. This result was not driven by an outlier but reflects a pattern seen in all three rats in the two diet groups.

The within-crypt error variance is $\sigma_e^2 = .284$ for the fish and $\sigma_e^2 = .249$ for the corn oil diet groups. The posterior mean and standard error estimates of the crypt-level variance components for the wavelet and scaling coefficients from finer to coarser resolutions for the fish oil and corn oil diet groups are $\sigma^2_{1,\text{FISH}} = \{.14 \pm .002, 4.01 \pm .03, 23.43 \pm .23, 111 \pm 1.5, 292 \pm 5.5, 491 \pm 13, 1312 \pm 35\}$ and $\sigma^2_{1,\text{CORN}} = \{.13 \pm .002, 3.97 \pm .03, 21.03 \pm .20, 105 \pm 1.4, 233 \pm 4.4, 454 \pm 12, 930 \pm 25\}$. For the rat-level variance components, the estimates and standard errors for the fish oil and corn oil diet groups are $\sigma^2_{2,\text{FISH}} = \{.029 \pm .001, .33 \pm .02, .71 \pm .16, 6.68 \pm 1.17, .016 \pm .030, 85 \pm 22, 457 \pm 117\}$ and $\sigma^2_{2,\text{CORN}} = \{.020 \pm .001, .26 \pm .02, .54 \pm .12, 2.85 \pm .79, 11.7 \pm 2.8, 42 \pm 12, 786 \pm 188\}$.

Note that for both diets, the crypt-to-crypt variation dominates both the rat-to-rat and residual error variation for distal and proximal MGMT. Based on the traces of the covariance matrices for the fish oil group (see Sec. 3.2), 83.8% (with standard error ±2.7%) of the distal MGMT variation is at the crypt level, whereas only 15.7% ± 2.7% and .5% can be explained by the rat and within-crypt levels. For the corn oil group, we see slightly less variation at the crypt level, but this variability still comprised most of the total. The percentages of variation were 74.0% ± 4.0%, 25.5% ± 4.0%, and .5% at the between-crypt, between-rat, and within-crypt hierarchical levels. This suggests that in future experiments, it will be important to sample a large number of crypts from each rat for both diets, because the crypt-to-crypt variation makes up such a large fraction of the total.

The multiresolution decomposition inherent in our method allows us to use the frequency domain to deduce features of the profiles at the various hierarchical levels. We see that the crypt and rat levels exhibit different patterns of variation at the various wavelet levels, which implies that the crypt- and rat-level profiles have different degrees of smoothness. Based on the trace, 99.0% of the distal variability at the rat level can be explained using just the scaling coefficients and coarsest wavelet coefficients, whereas at the crypt level, these levels account for only 57% of the variability. One must include all but the finest wavelet resolution levels before accounting for 99% of the variability at the crypt level. This suggests that the underlying profiles at the rat level are inherently smoother than the spike-laden crypt-level profiles. This result is not surprising given that the rat-level profiles are effectively obtained by averaging over crypt-level curves, which tends to have a smoothing effect.

Note that the crypt-level variance components at wavelet levels 3, 4, and 5 are orders of magnitudes larger than the corresponding rat-level components. Our chosen wavelet bases at these levels have support roughly between 28/256 and 112/256, and major peak of width 4/256–16/256, suggesting that the crypt-level profiles contain features at these frequencies that do not tend to be present in the smoother rat-level curves. Inspection of the raw data reveal that these features are likely the isolated peaks of width 10–20 pixels that characterize the crypt-level profiles. These peak widths match up with the expected width of individual cells, given that there are roughly 300 pixels and 15–30 cells per crypt, suggesting that the peaks may correspond to individual cells with high levels of MGMT activity. This cannot be confirmed using these data because a limitation in immunohistochemical staining prevents us from visually making out the actual cell boundaries, which is the reason why a pixel-based measuring scheme was necessary in lieu of a cell-based scheme. If true, however, this isolated peak phenomenon suggests that the MGMT DNA repair process may operate on a cell-by-cell basis, with little signaling between neighboring cells. With more intercellular signaling, one would expect more smoothness in the crypt-level curves, or at least similar heights of peaks corresponding to adjacent cells. This insight is potentially important in understanding how MGMT operates to repair carcinogen-induced DNA damage in colonic crypts.

5.3 Sensitivity Analysis

We performed a sensitivity analysis to assess the robustness of our biological results to the choice of prior in our Bayesian method. Because our priors on the variance components were chosen to be essentially uninformative relative to the data, we focus on the hyperparameters T_j and p_j of the shrinkage prior (see Sec. 4.2), which we argue are simply "smoothing" parameters that determine the degree of regularization in the nonparametric estimation.

We fit our model to the MGMT data under the following choices for T_j^2 and p_j:

(a) $T_j^2 = 20$, $\mathbf{p} = \{.005, .01, .015, .03, .07, .20\}$
(b) $T_j^2 = 20$, $p_j = .999999$ for all $j = 1, \ldots, 6$.
(c) $T_j^2 = 20$, $p_j = .0001$ for all $j = 1, \ldots, 6$.
(d) $T_j^2 = 20$, $\mathbf{p} = \{.015625, .03125, .0625, .125, .250, .500\}$.
(e) $T_j^2 = 10{,}000$, $\mathbf{p} = \{.005, .01, .015, .03, .07, .20\}$
(f) $T_j^2 = 1$, $\mathbf{p} = \{.005, .01, .015, .03, .07, .20\}$.

Case (a) contains the settings that we used in our application. Cases (b) and (c) correspond to the two extreme cases in which there is no shrinkage or total shrinkage in the wavelet coefficients. In case (d), the p_j decay exponentially by powers of two. Cases (e) and (f) have very large or very smaller values for T_j^2 but the same $\mathbf{p}$ as our application. Note that some of these hyperprior choices are extreme and would never be used in practice, but are included to assess the robustness of our results to even the most extreme priors.

As expected, changing these shrinkage hyperparameters affected the degree of regularization in estimating the diet/time profiles, but did not affect any of the major substantive results of our case study. For all cases, we still found higher MGMT levels at the lumenal surface of the crypts, and we observed a diet difference at the lumenal surface at 12 hours. The estimates from (a), (b), (d), and (e) were virtually the same, differing only in the amount of regularization. In case (b), no shrinkage of the wavelet coefficients is done, so the estimates are more noisy in appearance, as expected. In cases (c) and (f), the nondiscriminant shrinkage of all wavelet coefficients toward 0 results in distortion in the estimates near the crypt bases for fish oil–fed rats, again as expected from our discussion in Section 4.2. It is clear that the substantive results of our case study have not been driven by the prior.

6. DISCUSSION

From this case study, we found that MGMT expression tends to be greater near the lumenal surface of distal crypts than deeper within the crypts, suggesting the possibility of differential mechanisms of response to carcinogen-induced DNA damage depending on cell depth. Also, we found that rats fed the protective fish oil diet had higher MGMT levels at their lumenal surface at the latest time point included in the study, suggesting a possible mechanism by which the fish oil confers protection. From rat-level profile estimates, we confirmed that this result was not caused by a single outlying rat, but rather was a pattern seen across rats. Our analysis of covariance parameters indicated that a vast majority of the variability was at the crypt level, suggesting the importance of sampling a large number of crypts. Our multiresolution analysis pointed to the presence of peaks in the crypt-level data at frequencies corresponding to the typical width of individual cells, suggesting that the MGMT process may operate on a largely cell-by-cell basis.

We obtained these results by fitting a single unified model in the wavelet space. The method that we have introduced and applied to the MGMT data provides a unified framework for modeling hierarchical functional data without prespecifying a parametric form on the functions. As presented, it models one or two hierarchical levels but is easily generalized to more complex hierarchies. Our method is Bayesian, but our prior choices are uninformative except for the shrinkage hyperparameters, which, as we have demonstrated in our sensitivity analysis, are essentially smoothing parameters. By using wavelets, our method can handle spatially heterogeneous data, because in estimating the profiles, it smooths in such a way as to preserve local features. Note that our wavelet regression-based estimators are not as smooth in appearance as kernel or spline-based estimators, which is a trade-off for the procedure's strength in preserving local features of the profiles. Our method makes full use of the data and yields estimates and posterior samples for nearly any quantity that may be of interest.

We first obtain nonparametric estimates of the population profiles. Having obtained posterior samples for these profiles, we also have standard errors and credible intervals and can perform any Bayesian inferences of interest. We also obtain nonparametric estimates and posterior samples for the individual (rat)- and subsample (crypt)-level profiles, along with variance component estimates and posterior samples at the various hierarchical and wavelet levels. These can provide insight into the characteristics of the profiles at the different hierarchical levels and aid in designing future experiments. We can also garner insights about the data from the frequency domain information contained in the various wavelet resolution levels.

The reader will naturally wonder whether our approach is necessary for either our particular application or more general problems, and whether simpler approaches might give the same information. This seeming complexity occurs in other contexts of function estimation in hierarchical data. For example, if the crypt-level functions had been smooth instead of spikey as in Figure 2, then we could have used the smoothing-spline methods of, for example, Brumback and Rice (1998). These authors' approach in that context is not noticeably simpler than ours, either computationally or notationally, and their use of fixed effects for many crypt-level parameters is in contrast to our use of random effects throughout.

We believe that the power of our approach is that having run the MCMC, many questions can be answered in a straightforward fashion while still taking into account the intrinsic biological nature of the data. Among the questions that we have answered in Figure 3 and Section 5.2 are the following:

(a) Model diet-level functions with confidence statements via Bayesian credible intervals.
(b) Make probability statements about the diet functions at different colonic crypt depths.
(c) Make probability statements about different diet-level functions.
(d) Obtain crypt-level, rat-level, and diet-level variance components, allowing estimation, inference, comparison across diets, and so on.
(e) Obtain individual rat-level functions along with confidence statements.
(f) Obtain individual crypt-level functions along with confidence statements.

There is no doubt that other, seemingly simpler methods can be constructed to shed light on all of (a)–(f), although important subtleties of inference remain. Consider, for example, (a), the estimation of and inference about diet-level functions. We focus on this issue, although similar remarks can be made for (b)–(f). One modeling method would be as follows. For each cell depth t, average the responses in the data space across all of the crypts for each rat, average over all rats in the diet–time combination, and then finally fit some sort of nonparametric regression across t. We have applied this method to our data using a smoothing spline as the function fit, and obtained function estimates that are essentially the same as the solid lines in Figure 4. What our methods yield naturally that this method does not is confidence and probability statements of the type (b)–(c), discussed in Section 5.2. To obtain such statements in our problem (and, importantly, for other applications), the simple method would have to be adjusted to account for possibly different number of crypts within rats, for possibly different numbers of rats within diet–time combinations, to allow the rats and crypts within rats to be random and not fixed, and so on. At the level of detail of *inference* as opposed to estimation, this "simple" approach does not seem much simpler than our method.

Now consider (e). Naive estimates of the rat profiles can be obtained by averaging together all of the observed crypt profiles within that rat, but these estimates are noisy compared with the regularized estimates resulting from our method. Alternatively, smooth estimates could be obtained by fitting a smoothing procedure such as loss separately to the data for each rat. The advantage of our method is that we obtain regularized profile estimates that borrow strength from other rats and give standard error estimates that appropriately take into account the correlation structure imposed by the hierarchy.

An alternative to our procedure is to fit a multivariate hierarchical model to the data-space values. In our problem the responses are 256-dimensional, precluding the use of a general covariance structure. Whether the results of (a)–(f), and

especially likelihood inferences, are sensitive to the choice of the covariance structure would remain to be seen. Alternatively, one could choose a coarser grid of cell depths, say $t = 0, .05, .10, \ldots, 1.00$, and fit such multivariate hierarchical model to these data-space values. The general covariance structure is still quite large even in this case. It would appear to be worth future investigation to see whether the loss of information inherent in using only some of the data would be substantial enough to affect any of (a)–(f).

We believe that the method we have introduced here and applied to the MGMT dataset is a valuable tool for researchers working with hierarchical functional data. The method also shows promising potential for future extensions to other settings where replicated functional data are encountered with even more complex covariance structures, and the ideas contained herein may also be useful in building methods for hierarchical functional analysis using basis functions other than wavelets, such as Fourier series or regression splines.

APPENDIX: DERIVATION OF DISTRIBUTIONS FOR MCMC

A.1 Sampling for the γ's

We show that

$$\Pr(\gamma_a^{j,k} = 1 | \underline{d}_a^{j,k}, \Omega) = O/(1+O), \tag{A.1}$$

where $O = \{p_j/(1-p_j)\} * BF$ is a posterior odds ratio, and

$$BF = \left\{1 + \tau_j^2 / \operatorname{var}\left(\widehat{\Theta}_{a,NS}^{j,k}\right)\right\}^{-1/2} \times \exp\left[\left(\widehat{\Theta}_{a,NS}^{j,k}\right)^2 / 2\operatorname{var}\left(\widehat{\Theta}_{a,NS}^{j,k}\right)\left\{1 + \operatorname{var}\left(\widehat{\Theta}_{a,NS}^{j,k}\right)/\tau_j^2\right\}^{-1}\right] \tag{A.2}$$

is a Bayes factor comparing models with and without $\Theta_a^{j,k}$, given the variance components Ω, where $\widehat{\Theta}_{a,NS}^{j,k} = \sum_{b=1}^{B_a}\{\bar{d}_{ab}^{j,k}/\operatorname{var}(\bar{d}_{ab})\} / \sum_{b=1}^{B_a}\{\operatorname{var}(\bar{d}_{ab})\}^{-1}$ is a nonthresholded estimator of $\Theta_a^{j,k}$, with $\bar{d}_{ab} = C_{ab}^{-1}\sum_{c=1}^{C_{ab}} d_{abc}^{j,k}$, $\operatorname{var}(\bar{d}_{ab}) = \sigma_{2,j}^2 + (\sigma_{1,j}^2 + \sigma_\epsilon^2)/C_{ab}$, and $\operatorname{var}(\widehat{\Theta}_{a,NS}^{j,k}) = [\sum_{b=1}^{B_a}\{\operatorname{var}(\bar{d}_{ab})\}^{-1}]^{-1}$.

We now begin our proof of (A.1). For any set of matrices G_m for $m = 1, \ldots, M$, we define $\operatorname{diag}_M(G_m) = \operatorname{diag}(G_1, \ldots, G_M)$. Let $|A|$ represent the determinant of matrix A. Let J_n be the $n \times n$ matrix of 1s, let 1_n be a $n \times 1$ vector of 1s, and let I_n be the $n \times n$ identity matrix. If J or I has no subscript, then it is assumed to match the dimensions of the argument. Let $N = N_a = \sum_b C_{ab}$. We first show that

$$BF = \left[|\Sigma_1|^{1/2}\exp\left\{-(1/2)\left(\underline{d}_a^{j,k}\right)^T\Sigma_1^{-1}\left(\underline{d}_a^{j,k}\right)\right\}\right] \Big/ \left[|\Sigma_0|^{1/2}\exp\left\{-(1/2)\left(\underline{d}_a^{j,k}\right)^T\Sigma_0^{-1}\left(\underline{d}_a^{j,k}\right)\right\}\right], \tag{A.3}$$

where $\Sigma_0 = \sigma_{2,j}^2 \operatorname{diag}_{B_a}(J_{C_{ab}}) + (\sigma_{1,j}^2 + \sigma_\epsilon^2)\operatorname{diag}_{B_a}(I_{C_{ab}})$ and $\Sigma_1 = \Sigma_0 + \tau_j^2 J_N$.

To prove (A.3), first note by a trivial calculation that $BF = f(\underline{d}_a^{j,k}|\gamma_a^{j,k} = 1, \Omega)/f(\underline{d}_a^{j,k}|\gamma_a^{j,k} = 0, \Omega)$. Making the definitions $\underline{\Theta}_1^{j,k} = (\Theta_{a11}^{j,k}, \ldots, \Theta_{aB_aC_{ab}}^{j,k})^T$ and $\underline{\Theta}_2^{j,k} = (\Theta_{a1}^{j,k}, \ldots, \Theta_{aB_a}^{j,k})^T$, it is easily shown that $(\underline{d}_a^{j,k} - \underline{\Theta}_1^{j,k}|\underline{\Theta}_1^{j,k}) = \operatorname{normal}(0, \sigma_\epsilon^2 I_N)$, $(\underline{\Theta}_1^{j,k} - \underline{\Theta}_2^{j,k} \otimes 1_{C_{ab}}|\underline{\Theta}_2^{j,k}) = \operatorname{normal}(0, \sigma_{1,j}^2 I_N)$, $(\underline{\Theta}_2^{j,k} \otimes 1_{C_{ab}} - \Theta_a^{j,k} \otimes 1_N|\Theta_a^{j,k}) = \operatorname{normal}\{0, \sigma_{2,j}^2 \operatorname{diag}_{B_a}(J_{C_{ab}})\}$, $(\Theta_a^{j,k} \otimes 1_N|\gamma_a^{j,k} = 1) = \operatorname{normal}(0, \tau_j^2 J_N)$, and $(\Theta_a^{j,k} \otimes 1_N|\gamma_a^{j,k} = 0) = 0$. It follows immediately by summation that $(\underline{d}_a^{j,k}|\gamma_a^{j,k} = 0, \Omega) = \operatorname{normal}(0, \Sigma_0)$ and $(\underline{d}_a^{j,k}|\gamma_a^{j,k} = 1, \Omega) = \operatorname{normal}(0, \Sigma_1)$. This proves (A.3).

We now prove (A.1). First, note that $|\Sigma_1|/|\Sigma_0| = |I_N + \tau_j^2 J_N \Sigma_0^{-1}| = |\{1 + \tau_j^2(1_N^T\Sigma_0^{-1}1_N)\}|$. Because Σ_0 is a block-diagonal matrix with B_a diagonal blocks $\Sigma_{0(bb)}$, say, for $b = 1, \ldots, B_a$, it follows that $1_N^T\Sigma_0^{-1}1_N = \sum_{b=1}^{B_a} 1_{C_{ab}}^T\Sigma_{0(bb)}^{-1}1_{C_{ab}}$. Note that each diagonal block $\Sigma_{0(bb)}$ is an intraclass correlation matrix, with the eigenvectors $C_{ab}^{-1/2}1_{C_{ab}}$, a unit vector, and $C_{ab} - 1$, contrast vectors. This means that if $\lambda_{1(bb)}$ is the largest eigenvalue of $\Sigma_{0(bb)}$, corresponding to the unit eigenvector, then $1_N^T\Sigma_0^{-1}1_N = \sum_b^{B_a} C_{ab}^{-1}1_{C_{ab}}^T 1_{C_{ab}}\lambda_{1(bb)}^{-1}1_{C_{ab}}^T \times 1_{C_{ab}} = \sum_b^{B_a} C_{ab}\lambda_{1(bb)}^{-1}$. Because $\lambda_{1(bb)} = C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2$, it follows that

$$1_N^T\Sigma_0^{-1}1_N = \sum_b^{B_a} C_{ab}\left(C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2\right)^{-1} = \left\{\operatorname{var}\left(\widehat{\Theta}_{a,NS}^{j,k}\right)\right\}^{-1}, \tag{A.4}$$

and hence $\{|\Sigma_1|/|\Sigma_0|\}^{-1/2} = (1 + \tau_j^2\{\operatorname{var}(\widehat{\Theta}_{a,NS}^{j,k})\}^{-1})^{-1/2}$. Note that this is the first term in (A.2).

We next turn to the consideration of $\Sigma_1^{-1} - \Sigma_0^{-1}$. If $\Sigma_1 = \Sigma_0 + uv^T$, then the usual inverse calculation states that $\Sigma_1^{-1} - \Sigma_0^{-1} = -\Sigma_0^{-1}uv^T\Sigma_0^{-1}/(1 + v^T\Sigma_0^{-1}u)$. In our case, $u = \tau_j^2 1_N$ and $v = 1_N$, thus showing that $\Sigma_1^{-1} - \Sigma_0^{-1} = -\{(1_N^T\Sigma_0^{-1})^T(1_N^T\Sigma_0^{-1})\}/\{(\tau_j^2)^{-1} + 1_N^T\Sigma_0^{-1}1_N\}$. We now refer to (A.4), so that

$$\Sigma_1^{-1} - \Sigma_0^{-1} = -\left\{(1_N^T\Sigma_0^{-1})^T(1_N^T\Sigma_0^{-1})\right\} \Big/ \left\{(\tau_j^2)^{-1} + \left\{\operatorname{var}\left(\widehat{\Theta}_{a,NS}^{j,k}\right)\right\}^{-1}\right\}. \tag{A.5}$$

We now focus on the numerator of (A.5). We know that from the eigenvalue calculations given earlier, $1_N^T\Sigma_0^{-1} = (1_{C_{a1}}^T\Sigma_{0(11)}^{-1}, 1_{C_{a2}}^T\Sigma_{0(22)}^{-1}, \ldots, 1_{C_{aB_a}}^T\Sigma_{0(B_aB_a)}^{-1})$ and $1_{C_{ab}}^T\Sigma_{0(bb)}^{-1} = C_{ab}^{-1}(1_{C_{ab}}^T 1_{C_{ab}}) \times \lambda_{1(bb)}^{-1}1_{C_{ab}}^T = \lambda_{1(bb)}^{-1}1_{C_{ab}}^T$. Thus the numerator of (A.5) equals the $N \times N$ matrix

$$V = \left[(\lambda_{1(ss)}\lambda_{1(pp)})^{-1}1_{C_{as}}1_{C_{ap}}^T\right]_{s,p}. \tag{A.6}$$

Combining (A.5) and (A.6), we find that $\exp\{-(1/2)(\underline{d}_a^{j,k})^T(\Sigma_1^{-1} - \Sigma_0^{-1})\underline{d}_a^{j,k}\} = \exp((1/2)[(\tau_j^2)^{-1} + \{\operatorname{var}(\widehat{\Theta}_a^{j,k})\}^{-1}]^{-1}(\underline{d}_a^{j,k})^T V\underline{d}_a^{j,k})$. Now recall that $\underline{d}_a^{j,k} = (d_{a11}^{j,k}, \ldots, d_{aB_aC_{ab}}^{j,k})^T$. Additionally, let $\underline{d}_{ab}^{j,k} = (d_{ab1}^{j,k}, \ldots, d_{abC_{ab}}^{j,k})^T$. It follows that $(\underline{d}_a^{j,k})^T V\underline{d}_a^{j,k} = \{\sum_{b=1}^{B_a}\lambda_{1(bb)}^{-1} \times (\underline{d}_{ab}^{j,k})^T 1_{C_{ab}}\}^2 = \{\sum_{b=1}^{B_a} C_{ab}(C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)\bar{d}_{ab}\}^2 = \{\sum_{b=1}^{B_a}\bar{d}_{ab}/\operatorname{var}(\bar{d}_{ab})\}^2 = (\widehat{\Theta}_{a,NS}^{j,k})^2\{\operatorname{var}(\widehat{\Theta}_{a,NS}^{j,k})\}^{-2}$. This proves (A.2) and hence (A.1).

A.2 Sampling $\Theta_a^{j,k}$

Obviously, if $\gamma_a^{j,k} = 0$, then $\Theta_a^{j,k} = 0$. We now show that the distribution of $\Theta_a^{j,k}$ given $\gamma_a^{j,k} = 1$, $\underline{d}_a^{j,k}$, and Ω^j is normally distributed with mean and variance given by

$$\mu = \widehat{\Theta}_{a,NS}^{j,k}\left[\tau_j^2/\left\{\tau_j^2 + \operatorname{var}\left(\widehat{\Theta}_{a,NS}^{j,k}\right)\right\}\right] \tag{A.7}$$

and

$$\sigma^2 = \operatorname{var}\left(\widehat{\Theta}_{a,NS}^{j,k}\right)\left[\tau_j^2/\left\{\tau_j^2 + \operatorname{var}\left(\widehat{\Theta}_{a,NS}^{j,k}\right)\right\}\right]. \tag{A.8}$$

Note that the required conditional density is proportional to

$$f(\Theta_a^{j,k}|\gamma_a^{j,k} = 1, \Omega^j)\prod_{b=1}^{B_a} f(\underline{d}_a^{j,k}|\Theta_a^{j,k}\gamma_a^{j,k} = 1, \Omega^j). \tag{A.9}$$

The first term in (A.9) is proportional to $\exp\{-(1/2)(\Theta_a^{j,k})^2/(\tau_j^2)\}$, and the second term is proportional to $\exp\{-(1/2)\sum_{b=1}^{B_a}(\Theta_a^{j,k}1_{C_{ab}} -$

$\underline{d}_{ab}^{j,k})^T \Sigma_0^{-1}(\Theta_a^{j,k} 1_{C_{ab}} - \underline{d}_{ab}^{j,k})\}$. However, $\Sigma_0^{-1} = (\sigma_{1,j}^2 + \sigma_\epsilon^2)^{-1} \times [I_{C_{ab}} - \sigma_{2,j}^2/\{(\sigma_{1,j}^2 + \sigma_\epsilon^2)(\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)\} J_{C_{ab}}]$. After detailed algebra, this means that the second term in (A.9) is proportional to $\exp[-(1/2)\sum_{b=1}^{B_a} C_{ab}\{(\Theta_a^{j,k})^2 - 2\Theta_a^{j,k}\bar{d}_{ab}^{j,k}\}(C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)^{-1}]$. Combining terms, we see that the posterior mean and variance are $\mu = \{\sum_{b=1}^{B_a} C_{ab}\bar{d}_{ab}^{j,k}/(C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)\}/\{(\tau_j^2)^{-1} + \sum_{b=1}^{B_a} C_{ab}/\ (C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)\}$ and $\sigma^2 = \{(\tau_j^2)^{-1} + \sum_{b=1}^{B_a} C_{ab}/(C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)\}^{-1}$. Further simplifications yield (A.7)–(A.8).

A.3 Sampling $\Omega^j = \{\sigma_{1,j}^2, \sigma_{2,j}^2\}$

We now verify that the distribution of Ω^j given $\underline{d}^j = (d_{111}^{j,1}, \ldots, d_{AB_aC_{ab}}^{j,K_j})^T$ and $\underline{\Theta}_3^j = (\Theta_1^{j,1}, \ldots, \Theta_A^{j,K_j})^T$ is proportional to the expression $f(\Omega^j|\underline{d}^j, \underline{\Theta}_3^j)$ given by

$$
\begin{aligned}
&f(\Omega^j|\underline{d}^j, \underline{\Theta}_3^j) \\
&= (\sigma_{1,j}^2 + \sigma_\epsilon^2)^{-K_j \sum_{a=1}^{A}\sum_{b=1}^{B_a}(C_{ab}-1)/2} \\
&\quad \times \prod_{a=1}^{A}\prod_{b=1}^{B_a}(C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)^{-K_j/2} \\
&\quad \times \prod_{i=1}^{2}(\sigma_{i,j}^2)^{-(\mathcal{A}_{i,j}+1)} \times \exp\left[-\sum_{i=1}^{2}\mathcal{B}_{i,j}(\sigma_{i,j}^2)^{-1}\right] \\
&\quad \times \exp\left[-1/2\sum_{a=1}^{A}\sum_{k=1}^{K_j}\sum_{b=1}^{B_a}\sum_{c=1}^{C_{ab}}(d_{abc}^{j,k} - \Theta_a^{j,k})^2(\sigma_{1,j}^2 + \sigma_\epsilon^2)^{-1}\right] \\
&\quad \times \exp\Bigg[1/2\sum_{a=1}^{A}\sum_{k=1}^{K_j}\sum_{b=1}^{B_a} C_{ab}^2(\bar{d}_{ab}^{j,k} - \Theta_a^{j,k})^2\big[\sigma_{2,j}^2/\big\{(\sigma_{1,j}^2 + \sigma_\epsilon^2) \\
&\qquad \times (C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)\big\}\big]\Bigg]. \qquad \text{(A.10)}
\end{aligned}
$$

First, note that the conditional distribution of interest is proportional to

$$
f(\sigma_{1,j}^2, \sigma_{2,j}^2)\prod_{a=1}^{A}\prod_{k=1}^{K_j} f(\underline{d}_a^{j,k}|\Omega_j, \Theta_a^{j,k}). \qquad \text{(A.11)}
$$

Assuming independent inverse gammas with parameters $\mathcal{A}_{i,j}$ and $\mathcal{B}_{i,j}$ for $\sigma_{i,j}^2$, $i = 1, 2$, the first part of (A.11) has kernel $(\sigma_{1,j}^2)^{-(\mathcal{A}_{1,j}+1)}(\sigma_{2,j}^2)^{-(\mathcal{A}_{2,j}+1)}\exp\{-\mathcal{B}_{1,j}/\sigma_{1,j}^2 - \mathcal{B}_{2,j}/\sigma_{2,j}^2\}$, giving the second and third terms in (A.10).

With Σ_0 and N defined as in (A.3), the kernel of $f(\underline{d}_a^{j,k}|\Omega_j, \Theta_a^{j,k})$ in (A.11) is

$$
|\Sigma_0|^{-1/2}\exp\{-1/2(\underline{d}_a^{j,k} - \Theta_a^{j,k}1_N)^T\Sigma_0^{-1}(\underline{d}_a^{j,k} - \Theta_a^{j,k}1_N)\}. \qquad \text{(A.12)}
$$

First, focus on $|\Sigma_0|$. Recall that Σ_0 is a block-diagonal matrix whose blocks, $\Sigma_{0(bb)}$, are intraclass correlation matrices, with largest eigenvalue $C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2$ and remaining $C_{ab} - 1$ eigenvalues $\sigma_{1,j}^2 + \sigma_\epsilon^2$. Thus $|\Sigma_0| = \prod_{b=1}^{B_a}(\sigma_{1,j}^2 + \sigma_\epsilon^2)^{C_{ab}-1}(C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_e^2)$. Substituting into (A.12) and (A.11), with simplification yields the first term of (A.10). Further, the inverse of the diagonal blocks of Σ_0 can be directly verified to be $-[\{\sigma_{2,j}^2/(\sigma_{1,j}^2 + \sigma_\epsilon^2)(C_{ab}\sigma_{2,j}^2 + \sigma_{1,j}^2 + \sigma_\epsilon^2)\}]J_{C_{ab}} + (\sigma_{1,j}^2 + \sigma_\epsilon^2)^{-1}I_{C_{ab}}$. After substituting for Σ_0^{-1} in (A.12) and (A.11) and doing some algebraic simplifications, we arrive at the remaining terms of (A.10).

[Received ????. Revised ????.]

REFERENCES

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998), "Wavelet Thresholding via a Bayesian Approach," *Journal of the Royal Statistical Society*, Ser. B, 60, 725–749.

Boyle, P., Zaridze, D. G., and Smans, M. (1985), "Descriptive Epidemiology of Colorectal Cancer," *International Journal of Cancer*, 36, 9–18.

Brown, P. J., Fearn, T., and Vannucci, M. (2001), "Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem," *Journal of the American Statistical Association*, 96, 398–408.

Brumback, B. S., 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., Parmigiani, G., and Vidakovic, B. (1998), "Multiple Shrinkage and Subset Selection in Wavelets," *Biometrika*, 85, 391–401.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Philadelphia: Society for Industrial and Applied Mathematics.

Donoho, D., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation via Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Fan, J., and Zhang, J. T. (2000), "Two-Step Estimation of Functional Linear Models With Applications to Longitudinal Data," *Journal of the Royal Statistical Society*, Ser. B, 62, 303–322.

Giovannucci, E., and Willett, W. C. (1994), "Dietary Risk Factors and Risk of Colon Cancer," *Annals of Medicine*, 26, 443–452.

Hong, M. Y., Chapkin, R. S., Morris, J. S., Wang, N., Carroll, R. J., Turner, N. D., Chang, W. C. L., Davidson, F. A., and Lupton, J. R. (2001), "Anatomical Site-Specific Response to DNA Damage Is Related to Later Tumor Development in the Rat AOM Colon Carcinogenesis Model," *Carcinogenesis*, 22, 1831–1835.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Lantham, P., Lund, E. K., and Johnson, I. T. (1999), "Dietary n-3 PUFA Increases the Apoptotic Response to 1,2-Dimthylhydrazine, Reduces Mitosis and Supresses the Induction of Carcinogenesis in the Rat Colon," *Carcinogenesis*, 20, 645–650.

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192.

Mallat, S. G. (1989), "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–693.

Rice, J. A., and Wu, C. O. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253–259.

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, New York: Wiley.

Welsh, A. H. (1996), *Aspects of Statistical Inference*, New York: Wiley.