Bayesian Hierarchical Spatially Correlated Functional Data Analysis with Application to Colon Carcinogenesis

Veerabhadran Baladandayuthapani[1,*], **Bani K. Mallick**[2], **Mee Young Hong**[3,4], **Joanne R. Lupton**[4], **Nancy D. Turner**[4], and **Raymond J. Carroll**[2]

[1]Department of Biostatistics, Box 447, The University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030-4009, U.S.A.

[2]Department of Statistics, Texas A&M University, TAMU 3143, College Station, Texas 77843-3143, U.S.A.

[3]Center for Human Nutrition, University of California, Los Angeles, California 90095, U.S.A.

[4]Department of Nutrition and Food Science, Texas A&M University, TAMU 2253, College Station, Texas 77843-2253, U.S.A.

Summary

In this article, we present new methods to analyze data from an experiment using rodent models to investigate the role of p27, an important cell-cycle mediator, in early colon carcinogenesis. The responses modeled here are essentially functions nested within a two-stage hierarchy. Standard functional data analysis literature focuses on a single stage of hierarchy and conditionally independent functions with near white noise. However, in our experiment, there is substantial biological motivation for the existence of spatial correlation *among* the functions, which arise from the locations of biological structures called colonic crypts: this possible functional correlation is a phenomenon we term *crypt signaling*. Thus, as a point of general methodology, we require an analysis that allows for functions to be correlated at the deepest level of the hierarchy. Our approach is fully Bayesian and uses Markov chain Monte Carlo methods for inference and estimation. Analysis of this data set gives new insights into the structure of p27 expression in early colon carcinogenesis and suggests the existence of significant crypt signaling. Our methodology uses regression splines, and because of the hierarchical nature of the data, dimension reduction of the covariance matrix of the spline coefficients is important: we suggest simple methods for overcoming this problem.

Keywords

Bayesian methods; Carcinogenesis; Functional data analysis; Hierarchical model; Markov chain Monte Carlo; Mixed models; Regression splines; Semiparametric methods; Spatial correlation

1. Introduction

Colon cancer is among the leading causes of death in the United States and affects men and women equally. Given the asymptomatic nature of the development of colon cancer and the limited efficacy of treatments in its advanced stage, prediction of early carcinogenesis is crucial in prevention of this deadly disease. One important task to this end is to understand the

biological mechanisms underlying colon carcinogenesis, which includes distinguishing risk factors and determining alterations in cell-cycle kinetics at various stages of the carcinogenic process.

Although it is well known that environmental factors, most notably diet, have a significant impact on the prevention of colon cancer (Cummings and Bingham, 1998), the underlying mechanisms still remain relatively unexplored. The aim of this article is to present a new statistical method to analyze data following an experiment conducted by nutrition researchers at Texas A&M University, to study the interplay between diet and colon cancer at a cellular level. A carcinogen-induced rat colon tumor system is used in order to investigate the mechanisms by which diet modulates colon tumor development. The rats are fed particular diets of interest for specific periods, exposed to a carcinogen that induces colon cancer and subsequently euthanized for sample collection. The colon is then resected from these rats and examined for responses of interest.

As a point of general statistical methodology, this article is concerned with hierarchical functional data, where functions at their deepest possible level are correlated, rather than being assumed independent as in the standard literature. The major source of novelty in our approach stems from the fact that unlike the standard functional data analysis (FDA) literature (Ramsay and Silverman, 1997) that focuses on the special case of one diet/treatment, one animal and conditional independent functions with near-white noise, our functions are not necessarily independent. In this article, we develop a new methodology to model and test such dependency.

There is a special architecture of the cells in the colon that is crucial to understanding and modeling the underlying biological mechanisms of colon cancer (see Morris et al., 2002, 2003, for extensive statistical and biological details). Briefly, in the colon the cells are arranged in patterns called crypts; finger-like invaginations down toward the muscle layers, the top of which opens to the luminal surface of the intestine. Apoptosis (programmed cell death), differentiation, proliferation, and p27 (a protein that inhibits the cell cycle), etc., can be determined for each cell by staining them with an appropriate chemical. Measuring the levels of these biomarkers from the histological sections of colonic tissues is a well-documented technique for determining alterations in cell-cycle kinetics at various stages of the carcinogenic process (Hong et al., 1997) and may be used to identify the underlying cellular mechanisms that lead to eventual tumor development.

A crypt is typically 25-30 cells deep and is assayed from the bottom to the top of the crypt. Progenitor (stem) cells are toward the bottom of the crypt, where daughter cells are generated (cell proliferation) before moving up the crypt as they mature. While moving up the crypt toward the lumen, cells lose their ability to divide, and are exfoliated into the lumen of the intestine upon reaching the surface (Roncucci et al., 2000). Thus cells at different depths within a crypt are at different stages of maturity, and the cell position with respect to its position within a crypt is an important variable to consider in any subsequent analysis. We define the relative cell position, X such that $X = 0$ at the bottom of each crypt and $X = 1$ at the topof each crypt, i.e., for C cells in a given crypt, the ith cell is given a relative cell position $(i - 1)/(C - 1)$.

Morris et al. (2002) previously conjectured the existence of a *coordinated response* at the crypt level: biological response in one crypt may affect the response in neighboring crypts. In the experiment they considered, the spatial locations of the crypts were not measured, which rendered it impossible to understand the existence or extent of coordinated response. In our experiment, we have measured the physical locations of the crypts, and thus obtained the mutual distances among all crypts. Figure 1a shows the location of all the crypts, ≈20 per rat counted for rats sacrificed at 24 hours after administration of a carcinogen. The circles represent the physical location of the crypt in the tissue: the first crypt assayed is given a nominal location

zero. The horizontal axis is the distance in microns. In this study, four groups of animals are formed by combinations of diet (corn oil or fish oil) and butyrate supplementation (no or yes).

There are two special aspects to the data resulting from this experiment. First, the responses are inherently functional in nature, as functions of cell position within each crypt, rather than as discrete measurements. Second, the data resulting from this experiment have a natural hierarchical structure: diet/treatment groups, rats within diet/treatment, crypts within rat, and cells within crypts. While many important biological questions can be answered using these data, for this article, we will focus on the p27 response. p27 is a protein that inhibits the cell cycle by acting on the cyclin-dependent kinases, and thus is thought to be predictive of apoptosis and cell proliferation. Our goal in this article is twofold; first, we would like to model the mean p27 expression profiles taking into account the nested hierarchy: diet, rat, and crypt levels, respectively. Second, and more importantly for our purposes, we wish to determine if there is a coordinated response for p27, namely, how the level of p27 in the cells in a given crypt is affected by neighboring crypts, as function of crypt distances. We call this phenomenon *crypt signaling*.

If the observed functions at the various levels of the hierarchy could be well modeled using simple parametric forms, then estimation could proceed using standard mixed model methodology (Laird and Ware, 1982) accounting for the between-curve correlation imposed by design (see Verbeke and Mohlenberghs, 2000). However, some of the p27 expression profiles in our case cannot be well represented by simple parametric forms, leading us to look for nonparametric/semiparametric alternatives. There are numerous related approaches in the literature that deal with nonparametric estimation of replicated functions (e.g., Shi, Weiss, and Taylor, 1996; Staniswalis and Lee, 1998; Wang, 1998; Fan and Zhang, 2000; Rice and Wu, 2001; Wu and Zhang, 2002; Liang, Wu, and Carroll, 2003; Wu and Liang, 2004) but limit their scope to a single level of hierarchy. Grambsch et al. (1995) employed FDA-based methods to model the crypt data structure similar to one we consider here, although they also considered only one level of hierarchy. Brumback and Rice (1998) present a flexible smoothing spline based method to model functional data from nested and crossed designs, but treat individual-specific curves as fixed instead of random effects. Guo (2002) proposed a spline-based functional mixed model accommodating a broad range of fixed and random effect structures. Morris et al. (2003) developed a wavelet-based methodology for modeling functional data occurring within a nested hierarchy. However, in their framework the functions at the lowest level of the hierarchy (crypts) are assumed independent. Our work extends their methodology, wherein we accommodate for more general between-curve covariance structures, although we work with splines.

We handle the problem of nonparametric/semiparametric modeling using regression splines at the diet, rat, and crypt levels, with a representation of the random effects that allows for parsimonious modeling of the smoothing parameters. We exploit the link between P-spline smoothing and mixed models as shown by Wand (2003). This allows us to use the mixed model technology already in place for fitting penalized splines (Ngo and Wand, 2004) and also in a Bayesian framework (Crainiceanu, Ruppert, and Wand, 2005). Also because of the hierarchical nature of the data, dimension reduction of the covariance matrix of the spline coefficients is important: we suggest simple methods for overcoming this problem. In addition, we allow for the functions within a subject to be correlated by a parsimonious parametric representation of the correlations among the functions, using a flexible parametric family of autocorrelation functions. Although we will phrase much of our discussion in terms of the carcinogenesis example, the methodology is applicable to model any data where the functional responses (e.g., longitudinal, times series profiles) are inherently correlated. The treatment is fully Bayesian and uses Markov chain Monte Carlo (MCMC) techniques for inference.

The remainder of the article is organized as follows. In Section 2, we introduce the Bayesian model for hierarchical spatially correlated functional data. Section 3 deals with the estimation of the parameters and random variables in the model. In Section 4, we show the application to the colon carcinogenesis data and discuss model justifications and finally discussion and conclusions are drawn in Section 5. All the technical details of the MCMC sampler and simulations can be accessed via the Supplementary Materials.

2. Bayesian Hierarchical Spatially Correlated Functional Model

In this section, we lay out the basic modeling scheme for a spatially correlated functional model for the colon carcinogenesis data and defer the specifics of estimation to the next section. As previously described, our data consist of a nested hierarchy of functions. We have as responses: functions for each crypt, as a function of the relative cell depth, and these functions are sampled on rats within diet groups.

Suppose $d = 1, \ldots, D$ denotes the diet/treatment group, $r = 1, \ldots, R_d$ the rat, $i = 1, \ldots, m_{dr}$ the crypt and $j = 1, \ldots, n_{dri}$ the cell within the crypt. Let the marker response (logarithm transformed p27 in our case) from a given cell observed at location $X = X_{drij}$ in a crypt be denoted by $Ydrij (X)$. Let the mean function within a crypt be $\Theta_{dri} (X)$, which is corrupted in practice by near white-noise/measurement error $_{dri} = _{dri} (X_{drij}$, such that

$$Y_{drij} = Y_{drij}\left(X_{drij}\right) = \Theta_{dri}\left(X_{drij}\right) + \epsilon_{drij}. \tag{1}$$

Within each crypt, the errors $\varepsilon_{dri} = (_{dri1} \cdots _{drin_{dri}})^{\mathrm{T}}$ are assumed (for simplicity) to have mean zero and covariance matrix $\sigma^2_\epsilon I_{n_{dri}}$. Here we set $\epsilon_{dri} = \mathrm{Normal}\left(0, \sigma^2_\epsilon I_{n_{dri}}\right)$.

We will decompose the function $\Theta_{dri} (\cdot)$ into functions at the group/diet level, the rat/individual level and the crypt level, and we will allow the crypt-level functions to be correlated, i.e., we allow for crypt signaling. The way we do this is to define possibly different basis functions at the three levels, and we model $\Theta_{dri} (\bullet)$ as

$$\Theta_{dri}\left(X_{dri}\right) = \tilde{W}_{dri}\eta_d + \tilde{X}_{dri}\zeta_{dr} + \tilde{Z}_{dri}\beta_{dri}, \tag{2}$$

where $\left(\tilde{W}_{dri}, \tilde{X}_{dri}, \tilde{Z}_{dri}\right)$ are any basis matrix (e.g., regression splines, B-splines, smoothing splines, wavelets), and $(\eta_d, \zeta_{dr}; \beta_{dri})$ are the diet, rat, and crypt effects, respectively. Note that $\dim\left(\tilde{W}_{dri}\right) = n_{dri} \times p_3$; $_{(\eta_d, \zeta_{dr}, \beta_{dri})}; \left(\tilde{X}_{dri}\right) = n_{dri} \times p_2$ and $\tilde{W}_{dri} (\eta_d)$ are each (p_3, p_2, p_1)-variate vectors, where p_i is dimension of the spline basis. One could use in principle any nonparametric/ semiparametric basis function to model the individual curves; we will lay out the theory here using a flexible semiparametric modeling approach that of penalized regression splines (Ruppert, Wand, and Carroll, 2003). Penalized regression splines are a flexible and easily implemented methodology for fitting complex nonparametric models. We will revisit this issue later in the article. Also, note here that we allow the different basis matrix (and hence different amounts of smoothing) for each level of the hierarchy: diet, rats, and crypts. Although not needed, this added flexibility will be assumed throughout this article.

With this formulation, the diet-level function is $\tilde{W}_{dri}\eta_d + \tilde{X}_{dri}\zeta_{dr}$ and the rat-level function is $\tilde{W}_{dri}\eta_d + \tilde{X}_{dri}\zeta_{dr}$. In essence, we specify a hierarchical multilevel random effects model at each level of the hierarchy. Using standard formulation, the random effects distributions are: $\beta_{dri} =$

Normal(0, Σ_1) and ζ_{dr} = Normal(0, Σ_2) both mutually independent. The diet-level effects η_d are assumed to be fixed effects and are given a prior η_d = Normal(0, Σ_3). Note here, with this construction Σ_i, i = 1, 2, 3 are of very high dimension, and left unstructured, we are left with the task of estimating a large number of parameters. Hence, as a practical and methodological point of view, it is imperative we reduce the dimensionality of these matrices and we suggest simple tools in the next section.

In standard analysis, the crypt-level functions, $\tilde{Z}_{dri}\beta_{dri}$ are assumed independent, i.e., the crypt-level random effects (β_{dri}, β_{drk}) for crypts (i, k), respectively, are uncorrelated. In our experiment, it is biologically plausible that the nearer the crypts (in spatial proximity), the higher the relationship of the overall p27 expression. This phenomenon, termed crypt signaling, translates as the functions within a rat being correlated across crypts. We model the correlation between the crypts (i, k) as corr(β_{dri}, β_{drk}) = $\rho\{\Delta_{dr}(i, k), \chi_d\}$, where $\rho(\bullet)$ is a family of autocorrelation functions with a parameter vector χ_d possibly depending on diet and Δ is the Euclidean distance between the crypts. Thus, we assume the correlation function between any two crypts is of a parametric form and is only a function of the distance between them.

There are several choices available for the correlation function $\rho(\bullet)$ (see Stein, 1999 for an extensive overview). In this article, we work with a parametric family of autocorrelation functions, the Matérn family (Handcock and Stein, 1993; Stein, 1999). We will, however, follow an alternate parameterization as in Handcock and Wallis (1999) where the isotropic autocorrelation function has the general form

$$\begin{aligned} \rho(t, \alpha, \nu) &= 2^{1-\nu}\left(2t\nu^{1/2}/\alpha\right)^\nu K_\nu\left(2t\nu^{1/2}/\alpha\right)/\Gamma(\nu), \\ \chi &= (\alpha, \nu) > 0, \end{aligned}$$

where $\kappa_\nu(\bullet)$ is the modified Bessel function of order ν. The range parameter, α, controls the rate of decay of the correlation between observations as distance t increases. Large values of α indicate that sites that are relatively far from one another are moderately (positively) correlated. The parameter ν can be described as controlling the behavior of the autocorrelation function for observations that are separated by small distances. The attractive feature of the parameterization of Handcock and Wallis (1999) is that the interpretation of α is largely independent of ν and this aids in our Bayesian computations as it reduces the posterior correlation between the parameters.

Further, we also assume that the correlation function is stationary with respect to the distance (crypt locations). The assumption of stationarity in the underlying spatial process is a viable one in our case because the slice of the colon from where the crypts are assayed is only around 1-1.5 cm long.

Using the above formulation, let $\Sigma_{dr}(\chi_d>)$ be the correlation matrix formed from the terms $[\rho\{\Delta_{dr}(i, k), \chi_d\}]_{i,k=1}^{m_{dr}}$. Let $\mathcal{B}_{dr} = (\beta_{dr1}, \ldots, \beta_{drm_{dr}})$, obtained by arranging the β_{dri}'s "column-wise" with dim $(\mathcal{B}_{dr}) = p_1 \times m_{dr}$. Denote vec $(\mathcal{B}_{dr}) = \left(\beta_{dr1y}^T, \ldots, \beta_{drm_{dr}}^T\right)^T$, a $p_1 m_{dr} \times 1$ vector obtained by concatenating the columns of $\mathcal{B}_{dr}$. We assume $\mathcal{B}_{dr}$ are independent of one another and

$$\text{vec}(\mathcal{B}_{dr}) = \text{Normal}\left[0, \{\Sigma_{dr}(\chi_d) \otimes \Sigma_1\}\right], \tag{3}$$

where $\otimes$ is the Kronecker product. Thus we assume a separable covariance structure on our crypt-level coefficients.

3. Models and Estimation

Having laid out our model and distributional assumption in the previous section, this section describes our modeling approach, which involves the use of regression splines that are correlated at the crypt level.

3.1 Semiparametric Modeling of Hierarchical Functions

As mentioned before we will model the functions across rat and crypt level in a semiparametric framework using regression splines although one could, in principle, use any basis function, e.g., wavelets, B-splines, and smoothing splines. Regression splines are approximations to functions typically using a low-order number of basis functions. A particularly appealing class are low-order basis penalized regression splines, which achieve smoothness by penalizing the sum of squares or likelihood by a penalty parameter. The penalty parameter and the fit using penalized regression splines are easy to compute using mixed model technology (see Robinson, 1991; Coull, Ruppert, and Wand, 2001; Rice and Wu, 2001, among others). One nice feature of penalized regression splines is that because they are often cast within the class of mixed models methods, they are readily adapted to new problems.

For example, in a linear regression spline, the functional form of the crypt-level functions $\tilde{Z}_{dri}\beta_{dri}$ is taken as

$$\tilde{Z}_{dri}\beta_{dri} = \beta_{dri,I} + X\beta_{dri,L} + C(X)\beta_{dri,S},$$

where $(\beta_{dri})^T = (\beta_{dri,\,I}, \beta_{dri,\,L}, \beta_{dri,\,S})^T$ are the regression coefficients with subscripts (I, L, S) corresponding to intercept, linear, and spline parts, respectively. Here $C(x) = \{(x - \kappa_1)_+, \ldots, (x - \kappa_{K_1})_+\}$, the κ's are knots and the subscripted plus sign denotes the positive part of the argument. The regression coefficients $\beta_{dri}^T = \left(\beta_{dri,I}^T, \beta_{dri,L}^T, \beta_{dri,S}^T\right)$ are now distributed as Normal(0, Σ_1). Extension to higher-order polynomials are trivial: we use quadratic regression splines in our data analysis. Ruppert et al. (2003) give a detailed exposition for the number and placement of knots and their corresponding penalization. For penalized regression splines, the placement and number of knots is generally not crucial because the penalty takes care of overfitting (Ruppert, 2002). In our application, the underlying functions at each stage of the hierarchy are smooth, and hence a small number of knots suffice to capture all the local features of the data. We take knots at equally spaced quantiles of the data.

The same construction is also used for rat- and diet-level functions. With the above construction, we are left to estimating the covariance matrices at the diet, rat, and crypt levels Σ_i, $i = 1, 2, 3$. Left unstructured, each of these matrices has $p(p+1)/2$ unique parameters, respectively, where p is the dimension of matrices. Because in principle the number of knots and hence p can be relatively large, there is an obvious need for dimension reduction of these covariance matrices, a topic we take upin the next section.

3.2 Dimension Reduction of Covariance Matrices

In the implementation of our methodology for our particular example, the number p of the basis functions is relatively small. At least in principle then we can allow the covariance matrices (Σ_i) to be general. However, from both a practical and methodological point of view it is crucial to lower the dimensionality of (Σ_i). There are a variety of approaches available to this end. For example, Shi et al. (1996) achieve parsimony using a principal component decomposition of the covariance matrix of random effects. In a different context, Daniels and Pourahmadi (2002) provide a Bayesian method based on Cholesky decomposition.

In our case, for our implementation we use truncated power series basis functions. Motivated by a standard mixed-model representation of these basis functions for nonhierarchical settings, see for example Ruppert et al. (2003, p. 108), our dimension reduction has a natural form. The essential idea is to take the coefficients at the knots to be independent, while allowing the polynomial part to have an unstructured covariance matrix. Thus, if p_d is the degree of the regression splines, then we take $\Sigma_1 = \text{diag}\left(\Sigma_a, \sigma_\beta^2 I_{K_1}\right)$, where $\Sigma\alpha$ is an unstructured $p_d \times p_d$ matrix and K_1 are the number of knots. A similar formulation is assumed for $\Sigma_2 = \text{diag}\left(\Sigma_b, \sigma_\zeta^2 I_{K_2}\right)$ and $\Sigma_3 = \text{diag}\left(\Sigma_c, \sigma_\eta^2 I_{K_3}\right)$, where K_2 and K_3 are the number of knots at the rat and diet levels, respectively. Note that with this construction, the diet effects are considered mixed effects, with the spline part of the function being random effects with covariance matrix $\sigma_\eta^2 I_{K_3}$ while the polynomial part is given a prior Σ_c. We set $\Sigma_c = cI$, with c being set to a large number (say 100), thus serving as a noninformative vague prior on our fixed effect (polynomial) coefficients of the diet-level functions.

3.3 Estimation of Autocorrelation Function

The Matérn class of autocorrelation family is indexed by parameters $\chi = (\alpha, \nu)$, which control the range and the rate of decay (smoothness) of the correlation as a function of distance, respectively. In this article, we will be estimating both the spatial parameters of interest, as it is crucial in capturing the spatial correlation between the crypt-level functions. The choice of prior distribution for the autocorrelation parameters is important and can lead to improper posterior distributions (Stein, 1999). Possible choices of prior distribution include a uniform prior on $(0, d_{\max})$ for α, where $d_{\max}$ is the maximum distance between observed crypt locations. A similar uniform prior limiting ν to be in $(0, c)$ can be taken. The upper limit c is set so that a wide variety of behaviors is possible. There is little difference in the autocorrelation function for large values of ν. In fact, for large values of ν, it is diffcult to distinguish between autocorrelation functions; thus limiting the magnitude of ν does not greatly influence the behavior near the origin. However, none of these priors ensure any posterior conjugacy, so we will be using a Metropolis-Hastings (MH) step within a Gibbs sampler to estimate these parameters.

With this setup the set of model parameters and random variables to be estimated are $\mathcal{M} = \left(\eta_d, \zeta_{dr}, \beta_{dri}, \Sigma_a, \Sigma_b, \sigma_\beta^2, \sigma_\zeta^2, \sigma_\eta^2, \sigma_\epsilon^2, \chi_d\right)$. To complete the model specifications, the covariance matrices (Σ_a, Σ_b) are given Inverse Wishart priors and $\sigma_\beta^2, \sigma_\zeta^2, \sigma_\eta^2, \sigma_\epsilon^2$ are given Inverse Gamma priors. The full conditionals for all the model parameters and random variables are in proper form and a Gibbs sampler can be used to sample them, except for χ_d for which we have to resort to a MH algorithm. The full conditionals for the MCMC sampling scheme are given in the Supplementary Materials.

4. Application to Colon Carcinogenesis Data

4.1 Implementation of Method

We applied our method to the colon carcinogenesis experiment described in Section 1. In this study, we have four diet groups formed by combinations of diet (corn oil or fish oil) and butyrate supplementation (no or yes). There are three rats per diet group, on average around 20 crypts per rat and average 26 cells per crypt, for a total of 6389 observations. The data are assayed at four time points: 0, 12, 24, and 48 hours. We focus on the data assayed at the 24-hour time point for the exposition of our methodology. The p27 responses (logarithm transformed) are standardized to have overall mean 0 and variance 1. Recall that in our experiment we have the exact physical mutual distances between the crypts measured as shown in Figure 1a. The circles

represent the physical locations of the crypt in the tissue: the first crypt assayed is given a nominal location of zero. The horizontal axis is the distance in microns. On the right-hand side (Figure 1b) are shown the histograms of the mutual distance $\Delta_{dr}(i, k)$ between all pairs of crypts less than 1000 microns apart. Our primary interest is to estimate the correlation function between 5 and 200 microns.

After some initial data analysis, we assumed the following functional forms for the various stages of hierarchy, namely, a quadratic penalized regression spline with 3 knots at the crypt level, and quadratic penalized regression splines with 5 knots at the rat level and diet level. Thus, our basis functions are of the form,

$$W_{drij}^T = \begin{bmatrix} 1 & m_1\left(X_{drij}\right) & m_2\left(X_{drij}\right) & C_\eta^T\left(X_{drij}\right) \end{bmatrix};$$
$$X_{drij}^T = \begin{bmatrix} 1 & m_1\left(X_{drij}\right) & m_2\left(X_{drij}\right) & C_\zeta^T\left(X_{drij}\right) \end{bmatrix};$$
$$Z_{drij}^T = \begin{bmatrix} 1 & m_1\left(X_{drij}\right) & m_2\left(X_{drij}\right) & C_\beta^T\left(X_{drij}\right) \end{bmatrix};$$

corresponding to diet level, rat level, and crypt level, respectively, and where $m(\bullet)$ and $C(\bullet)$ are the polynomial and spline parts, respectively. In fitting the polynomials, we scaled the polynomials to be approximately orthogonal and to have approximate variance 1, i.e.,

$$
\begin{aligned}
m_0(x) &= 1 \\
m_1(x) &= \sqrt{12}\,(x - 0.5); \\
m_2(x) &= \sqrt{180}\left\{(x - 0.5)^2 - 1/12\right\}.
\end{aligned}
$$

With this model we are essentially treating diet-level effects as mixed effects and the rat- and crypt-level effects as random. We use the dimension reduction mentioned in Section 3.3, leaving the covariance matrix of the polynomial terms as unstructured 3×3 matrices with an Inverse Wishart prior and a structured diagonal matrix for the spline part. We set the degrees of freedom for the Inverse Wishart prior to be $p + 1$ in order for it to suffciently diffuse around the prior mean. We use a method of moments estimate as the prior mean for the Inverse Wishart distribution. We also implemented our method without dimension reduction, i.e., using unstructured covariance matrices with an Inverse Wishart prior and obtained very similar answers. The hyperparameters for all Inverse Gamma priors are $(1, 1)$, for them to be suffciently noninformative. We used a single MCMC chain of 60,000 iterations, keeping every 10th sample with a burn-in of 10,000 iterations. We ran parallel MCMC chains with diverse starting values; these converged to the same range of values for each parameter. Algorithmic and computational details of the MCMC implementation can be accessed via the Supplementary Materials.

4.2 Results

Figure 2 shows the correlation functions as a function of crypt distance with the corresponding 95% credible interval. Our major interest is in the correlation between the crypt functions between 25 and 200 microns. We see an interesting degree of correlation between the functions; the correlation at 25 microns is 0.57 with 95% credible interval (0.44, 0.68). This observation strongly supports our biological hypothesis of the existence of crypt signaling: the p27 expression in the crypts that are in closer proximity tend to have similar expression levels and hence are highly correlated. The estimated (posterior mean) of the Matérn order, ν, is 0.11 with 95% credible interval (0.06, 0.17), significantly different from the classic autoregressive model which corresponds to $\nu = 0.5$.

Figure 3a shows the posterior mean diet-level functions for the four diet groups: CO is corn oil and FO is fish oil with or without (±) butyrate supplement. There seem to be some diet differences especially between the CO + B diet and the rest of the diets. To investigate this further, we plot the posterior mean along with 90% credible intervals of the pointwise difference between the diet functions. Figure 3b shows the posterior pointwise differences between the diet functions as a function of relative cell depth between two pairs of diets. We find that the CO + B diet is significantly different from the others. As can also be seen from Figure 3a that the p27 expression tends to be lower in the middle of the crypt. It has been reported in the biological literature (Lloyd et al., 1999; Sgambato et al., 2000) that p27 happens to be an inhibitor of cell proliferation, an increase in the number of cells as a result of cell growth and cell division. The middle of the crypt is the proliferating zone and thus p27 expression tends to be lowest in this zone.

We also investigated the existence of an interaction between diet (fish oil/corn oil) versus butyrate supplement (yes/no). We first computed the interaction function between diet and butyrate supplement as a function of relative cell position, namely, $f_{C+B} - f_{C-B} - f_{F+B} + f_{F-B}$, where for example the subscript "C + B" indicates corn oil with(+) butyrate and "C - B" indicates corn oil without(-) butyrate. To gain strength by sharing of information, we then averaged the interaction function within the middle tertile of the crypt, and constructed a 99% credible interval for this average, the interval being from (0.08, 1.29). This suggests an interaction in p27 expression between diet and butyrate supplementation in the middle tertile of the crypt. We repeated these calculations without taking into account the functional correlations, and found no such interaction. At least in this instance, taking into account the functional correlation appears to lead to a somewhat different finding than when the correlation is ignored. The actual plot is given in Figure 4, where it is seen that the interaction is confined to the middle tertile.

4.3 Model Justifications

In response to some important concerns of the reviewers, we discuss in this section a few issues about the model and the methodology. First, we justify some of the assumptions made in constructing and implementing the model and perform simple checks if the model actually fits.

To understand the nature of the fit at the diet level, we did a marginal analysis, i.e., lumped all the data together and fit the spline as if the data were independent. The resulting fits have somewhat the same shape as what we have obtained, i.e., a smooth quadratic type surface. Of course, as expected, the levels are different, because there are only three rats per diet. Also, as a part of our exploratory analysis, we looked at many hundreds of crypts, and there is no visual evidence that the functions vary rapidly in any one part of the crypt, i.e., there is no visual evidence of the need for spatial adaptation. The functions were relatively smooth within the crypts and a penalized spline with small number of knots sufficed to capture the variations of the functions. In principle, however, it would be useful to adjust our method to allow for spatially adaptive penalties. An example of this in the univariate smoothing context is given by Baladandayuthapani, Mallick, and Carroll (2005), but in the present context this extra flexibility does not appear to be needed.

In order to provide at least a partial check on whether the MCMC methodology is driven too strongly by starting values, priors, run length, etc., we performed the following simple frequentist pseudolikelihood analysis. For each rat, we fit a marginal cubic regression and formed the residuals from this cubic regression: visually and with simple spline fits, none of the rats exhibited vast departures from the cubic regression. Within each crypt, we selected 10 such residuals, equally spaced based on cell position. Thus, for example, if there were 28 cells in the crypt, we selected the cell numbers (1, 4, 7, 10, 13, 16, 19, 22, 25, 28). These residuals are linearly independent and, assuming a cubic regression at the rat level, their distribution is

jointly normally distributed with mean zero and a covariance matrix depending solely upon the error variance σ_ϵ^2, the Matérn shape and range parameters ν and α, along with the six parameters in the crypt-level covariance matrix $\Sigma_1 = \text{diag}\left(\Sigma_a, \sigma_\beta^2 I\right)$. To parameterize Σ_a, we used the construction of Daniels and Pourahmadi (2002), namely, that

$$
\begin{aligned}
\Sigma_a &= \mathcal{S}^{-1} \text{diag}\left(\sigma_I^2, \sigma_L^2, \sigma_Q^2\right)\left(\mathcal{S}^{\mathrm{T}}\right)^{-1}; \\
\mathcal{S} &= \begin{bmatrix} 1 & 0 & 0 \\ \zeta_1 & 1 & 0 \\ \zeta_2 & \zeta_3 & 1 \end{bmatrix}.
\end{aligned}
$$

In this construction, $\left(\sigma_I^2, \sigma_L^2, \sigma_Q^2\right) \geq 0$ while $(\zeta_1, \zeta_2, \zeta_3)$ are unconstrained. We then maximized the pseudolikelihood in these parameters. We obtained parameter estimates $\widehat{\nu}=0.13$ and $\widehat{\alpha}=0.66$, very similar to our posterior mean estimates of 0.11 and 0.96, respectively, with estimated correlation at 25 microns of 0.63. Thus this analysis shows no evidence that our MCMC methods are being stuck far from reasonable values.

In addition, we redid our model fits for the data below the middle of the crypt and above the middle of the crypt. Of course, the answers changed, but there was no major evidence that the correlation functions were vastly different, so that stationarity of the correlation surface seems a reasonable assumption in our context. As part of that process, we looked at the residuals of cubic fits at the crypt level, via q - q plots, and they appear to be roughly normally distributed.

One of the referees raised a concern that our separable covariance model requires the Matérn functions to be independent of cell depth. To test whether this is reasonable, we repeated the pseudolikelihood method for the bottom third, middle third, and top third of the crypts, using 6 residuals from each crypt rather than 10. The pseudolikelihood estimates for ν were all between 0.09 and 0.16, with none being significantly different via a likelihood ratio test from values in this range. The estimated functions are given in Figure 5. The maximum difference in the correlations in the range of 25-200 microns is approximately 0.15, and when the calculations are done at the same values of $\nu = 0.15$, the maximum difference is 0.10. Thus, there appears to be no substantial evidence that the correlation surface of the functions depends upon cell depth.

5. Discussion and Conclusions

Motivated by an application in colon carcinogenesis, we have proposed a Bayesian method to model the spatial correlation in hierarchical functions. The individual functions at each stage of the hierarchy are modeled semiparametrically using regression splines, although of course one could use any nonparametric/semiparametric basis functions. The data we model here consist of profiles of p27, an important cell-cycle mediator that changes in early carcinogenesis, nested within a two-stage hierarchy. Unlike the standard literature, our functions at the lowest level of the hierarchy (crypts) are not conditionally independent. Thus, as a point of general statistical methodology we require an analysis that allows for the functions to be correlated at the deepest level of the hierarchy. We model the spatial correlation between the functions parametrically using a flexible family of the autocorrelation functions, namely, the Matérn family. Our analysis gives new insights into the structure of p27 expression in early colon carcinogenesis. Our results show considerable correlation between the colonic crypt functions and help establish the existence of the biological phenomenon that we call crypt signaling.

With many knots, fitting an unstructured model for the crypt-level covariance matrix Σ_1 clearly violates the idea of penalization, and could reasonably be seen as the wrong way to attack the

problem: it is trying to model how the spline regression coefficients vary in a high-dimensional nonparametric way. Our low-dimensional approximation method enforces penalization, at the risk of potential model misspecification, and is clearly not the correct approach at the most general level. This being said, we do feel that our method has great value in practice, and the formulation of the penalization that we use fits naturally into the literature of penalized spline smoothing.

In general, a compromise is needed, although in our data example both extremes gave similar fits. The compromise that we think will work is, in one way or another, to regularize the estimation of Σ_1: our method is one simple and direct form of this. The regularization that allows for penalization while being flexible seems to us best placed (in the Bayesian framework) on variable selection methods for covariance matrix estimation, as in for example Wong, Carter, and Kohn (2003). Implementing such covariance selection methods in the complex context of our example is an interesting and challenging problem.

As we mentioned before, our aim in the article is twofold. First, model the mean functions at each level of the hierarchy: diet, rat, and crypt level in a flexible manner. Second, model the correlation between the random functions at the deepest level of the hierarchy, i.e., crypt level, in a spatial manner. We believe our computationally intensive Bayesian treatment of the problem has many advantages. First, our Bayesian hierarchical modeling exercise allows us to model both the mean functions and correlation function in a unified framework, such that the uncertainty in the estimation process is accounted for and propagated through, at each level of the hierarchal model. Secondly, having run the MCMC chain, a number of inferential questions can be answered in a coherent manner using the posterior distribution (samples) such as exact confidence statements regarding the mean and correlation function via Bayesian credible intervals. Although our treatment of the problem is Bayesian, we certainly agree that there are simpler devices like the marginal pseudolikelihood analysis presented in Section 4.3 available to answer the questions we pose here. For the application we consider here, these devices might have to be tailored for the level of detail of inference we achieve here. For example, modeling the hierarchical functions flexibly using splines, accounting for possibly different number of replicates at each level of the hierarchy and treating the functions as random effects rather than fixed effects. Taking into account these observations, these modified devices are no simpler than the approach we follow here. We also note that we did a small simulation study to evaluate the operating characteristics of our proposed methodology. We found that the method performed well on simulated data and the estimated correlation function had reasonable frequentist properties. The details of the simulation study can be accessed via the Supplementary Materials.

We have also developed a MATLAB software suite for implementing the methods we describe in the article. The code will be available for download through the first author's website at http://odin.mdacc.tmc.edu/~veera/. The directory contains stepby stepdetails about implementing the MCMC computations we adopt here. The code contains default settings needing minimal input from an average user like starting values generated using method of moments approach. The easy to use code also allows an expert user to change the settings like number of knots for the spline, starting values and priors and hyperparameters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

Baladandayuthapani V, Mallick BK, Carroll RJ. Spatially adaptive Bayesian penalized regression splines (P-splines). Journal of Computational and Graphical Statistics 2005;14:378–394.

Brumback BA, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). Journal of the American Statistical Association 1998;93:961–976.

Coull BA, Ruppert D, Wand MP. Simple incorporation of interactions into additive models. Biometrics 2001;57:539–545. [PubMed: 11414581]

Crainiceanu C, Ruppert D, Wand M. Bayesian analysis for penalized spline regression using WinBUGS. Journal of Statistical Software 2005;14(14)

Cummings JH, Bingham SA. Diet and prevention of cancer. British Medical Journal 1998;317:1636–1640. [PubMed: 9848907]

Daniels MJ, Pourahmadi M. Dynamic models and Bayesian analysis of covariance matrices in longitudinal data. Biometrika 2002;89:553–566.

Fan J, Zhang JT. Two-step estimation of functional linear models with applications to longitudinal data. Journal of the Royal Statistical Society, Series B 2000;62:303–322.

Grambsch PM, Randall BL, Bostick RM, Potter JD, Louis TA. Modeling the labeling index distribution: An application of functional data analysis. Journal of the American Statistical Association 1995;90:813–821.

Guo W. Functional mixed effects models. Biometrics 2002;58:121–128. [PubMed: 11890306]

Handcock MS, Stein ML. A Bayesian analysis of kriging. Technometrics 1993;35:403–410.

Handcock MS, Wallis JR. An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). Journal of the American Statistical Association 1993;89:368–378.

Hong MY, Chang WL, Chapkin RS, Lupton JR. Relationship among coloncyte proliferation, differentiation, and apoptosis as a function of diet and carcinogen. Nutrition and Cancer 1997;28:20–29. [PubMed: 9200146]

Laird N, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–974. [PubMed: 7168798]

Liang H, Wu H, Carroll RJ. The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coeeffcient models with measurement error. Biostatistics 2003;4:297–312. [PubMed: 12925523]

Lloyd RV, Erickson LA, Jin L, Kulig E, Qian X, Cheville JC, Scheithauer BW. p27Kip1: A multifunctional cyclin-dependent kinase inhibitor with prognostic significance in human cancers. American Journal of Pathology 1999;154:313–323. [PubMed: 10027389]

Morris JS, Wang N, Lupton JR, Chapkin RS, Turner ND, Hong MY, Carroll RJ. A Bayesian analysis involving colonic crypt structure and coordinated response to carcinogens incorporating missing crypts. Biostatistics 2002;3:529–546. [PubMed: 12933596]

Morris JS, Vannucci M, Brown PJ, Carroll RJ. Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). Journal of the American Statistical Association 2003;98:573–583.

Ngo L, Wand M. Smoothing with mixed model software. Journal of Statistical Software 2004;9(1)

Ramsay, JO.; Silverman, BW. Functional Data Analysis. Springer; New York: 1997.

Rice JA, Wu C. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 2001;57:253–269. [PubMed: 11252607]

Robinson GK. That BLUP is a good thing: The estimation of random effeects (with discussion). Statistical Science 1991;6:15–51.

Roncucci L, Pedroni M, Vaccina F, Benatti P, Marzona L, De Pol A. Abberrant crypt foci in colorectal carcinogenesis: Cell and crypt dynamics. Cell Proliferation 2000;33:1–18. [PubMed: 10741640]

Ruppert D. Selecting the number of knots for penalized splines. Journal of Computational and Graphical Statistics 2002;11:735–757.

Ruppert, D.; Wand, MP.; Carroll, RJ. Semi-parametric Regression. Cambridge University Press; New York: 2003.

Sgambato A, Cittadini A, Faraglia B, Weinstein IB. Multiple functions of p27kip1 and its alterations in tumor cells: A review. Journal of Cell Biology 2000;183:18–27.

Shi M, Weiss RE, Taylor JMG. An analysis of pediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. Applied Statistics 1996;45:151–163.

Staniswalis JG, Lee JJ. Nonparametric regression analysis of longitudinal data. Journal of the American Statistical Association 1998;93:1403–1418.

Stein, ML. Interpolation of Spatial Data. Springer; New York: 1999.

Verbeke, G.; Molenberghs, G. Linear Mixed Models for Longitudinal Data. Springer; New York: 2000.

Wand M. Smoothing and mixed models. Computational Statistics 2003;18:223–249.

Wang Y. Mixed effects smoothing spline analysis of variance. Journal of the Royal Statistical Society, Series B 1998;60:159–174.

Wong F, Carter CK, Kohn R. Efficient estimation of covariance selection models. Biometrika 2003;90:809–830.

Wu H, Zhang JT. Local polynomial mixed-effects models for longitudinal data. Journal of the American Statistical Association 2002;97:883–897.

Wu H, Liang H. Backfitting random varying-coefficient models with time-dependent smoothing covariates. Scandinavian Journal of Statistics 2004;31:3–20.

(a) Spatial location of the crypts

(b) Histogram of crypt distances
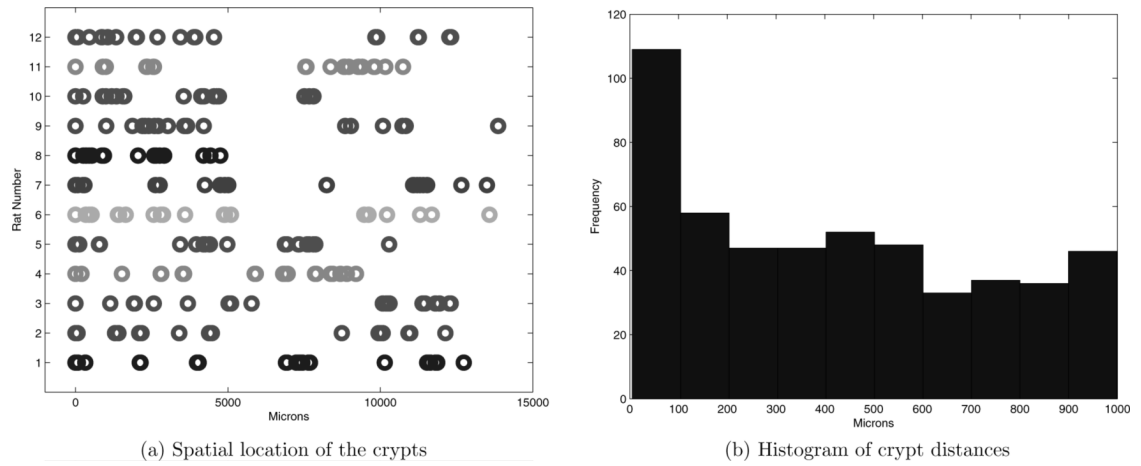
Figure 1.
(a) The vertical axes are the individual rats and the horizontal axes are the distances in microns and circle represent the physical location of the crypts for all rats assayed at 24-hour time point. (b) Histogram of the mutual crypt distances ($|\Delta|$) for all rats assayed at 24-hour time point. Plotted are the distances less than 1000 microns.
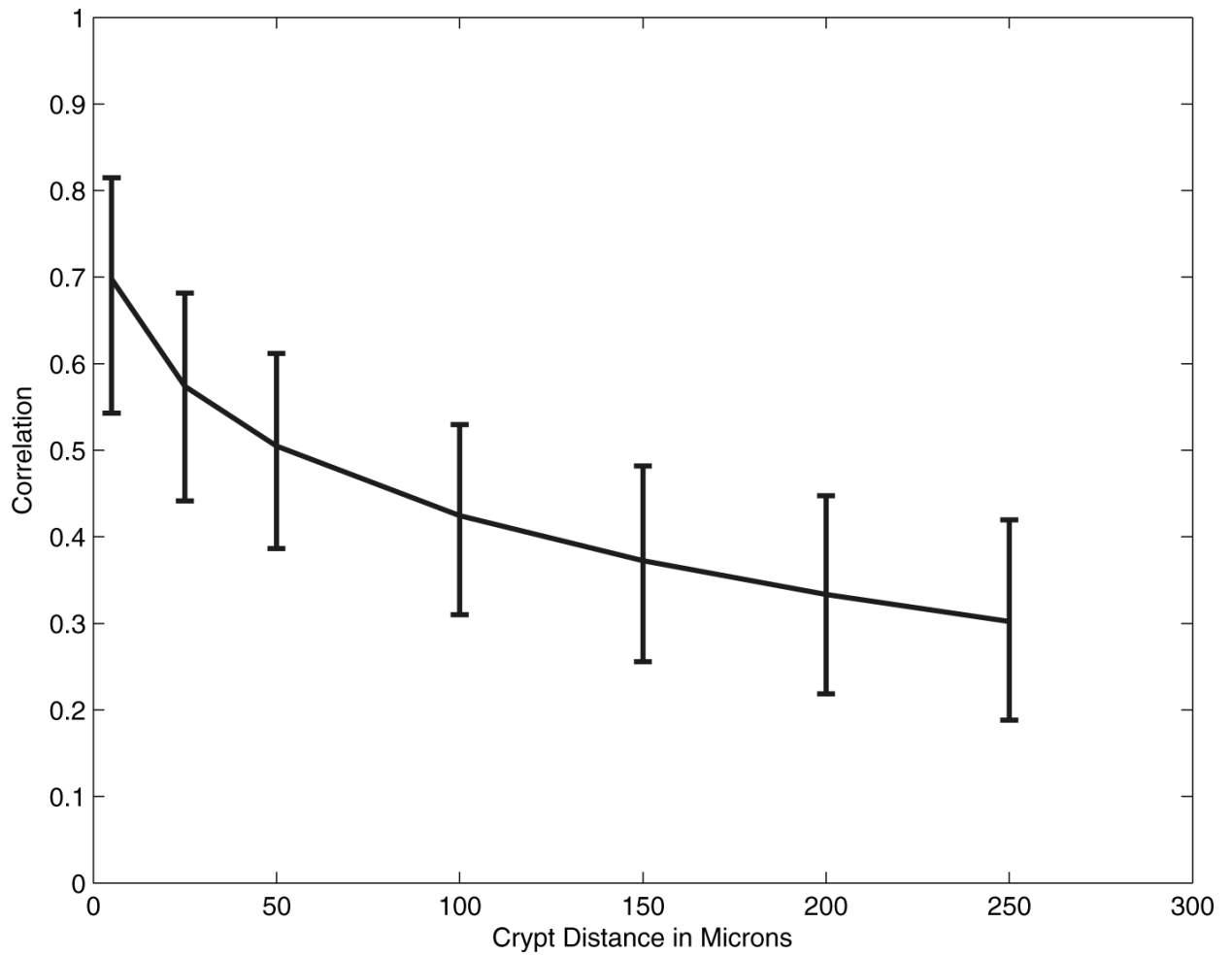
Figure 2.
Posterior correlations as a function of crypt distance with 95% error bars. The vertical axis is the correlation and the horizontal axis is the distance between the crypts (Δ).
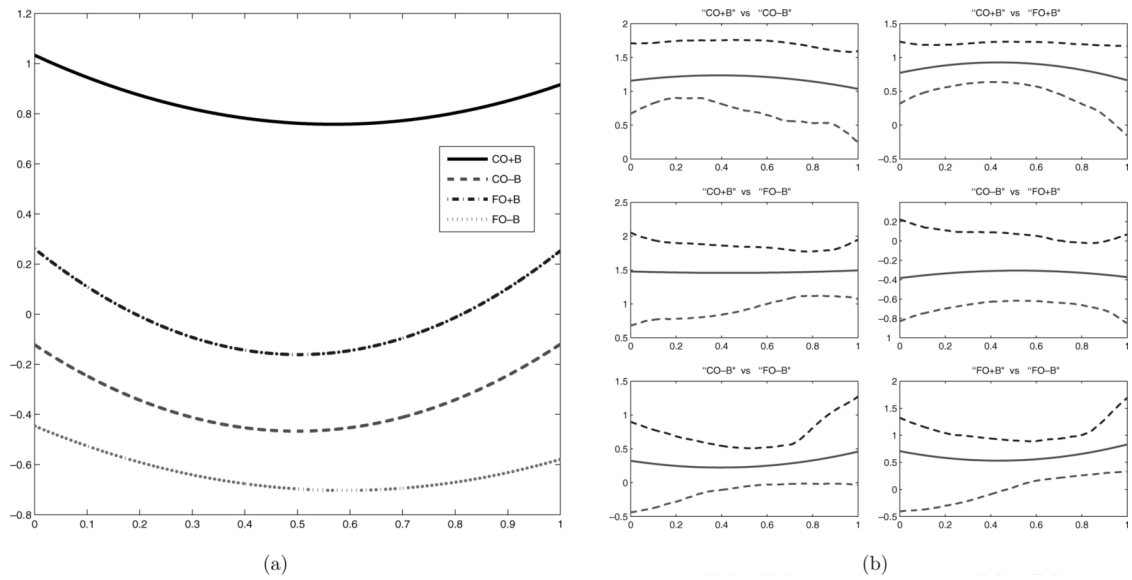
Figure 3.
(a) Shown here are the posterior marginal mean functions for the four diet groups. The horizontal axis is the relative cell depth. CO is corn oil and FO is fish oil with or without (±) butyrate supplement. (b) Posterior mean along with 90% credible intervals of the pointwise difference between the diet functions.

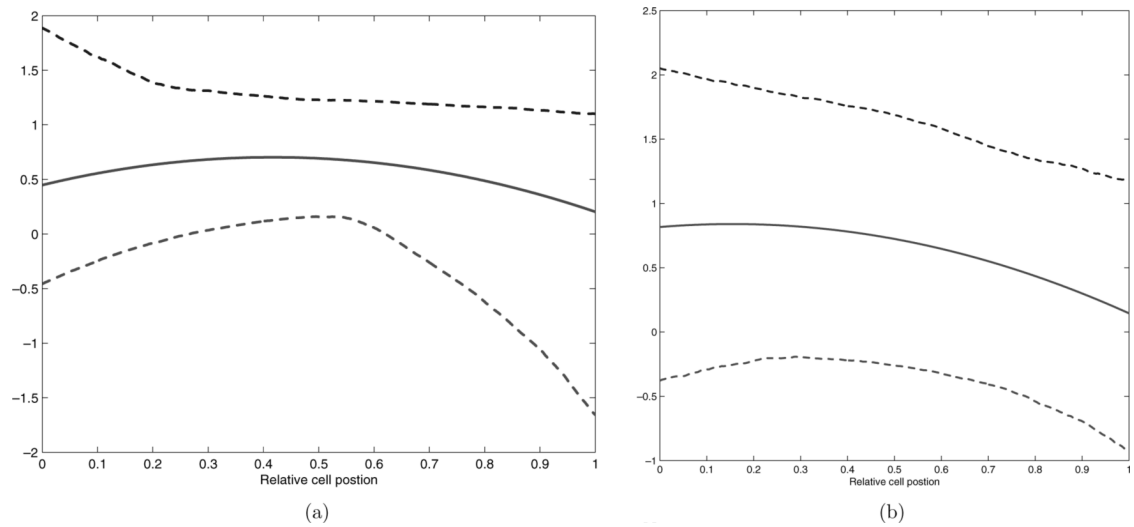**Figure 4.**
Posterior interaction function between diet and butyrate as a function of relative cell depth using (log)p27 response (a) accounting for correlation and (b) assuming independence between crypts. Also shown are the 95% Bayesian credible intervals.
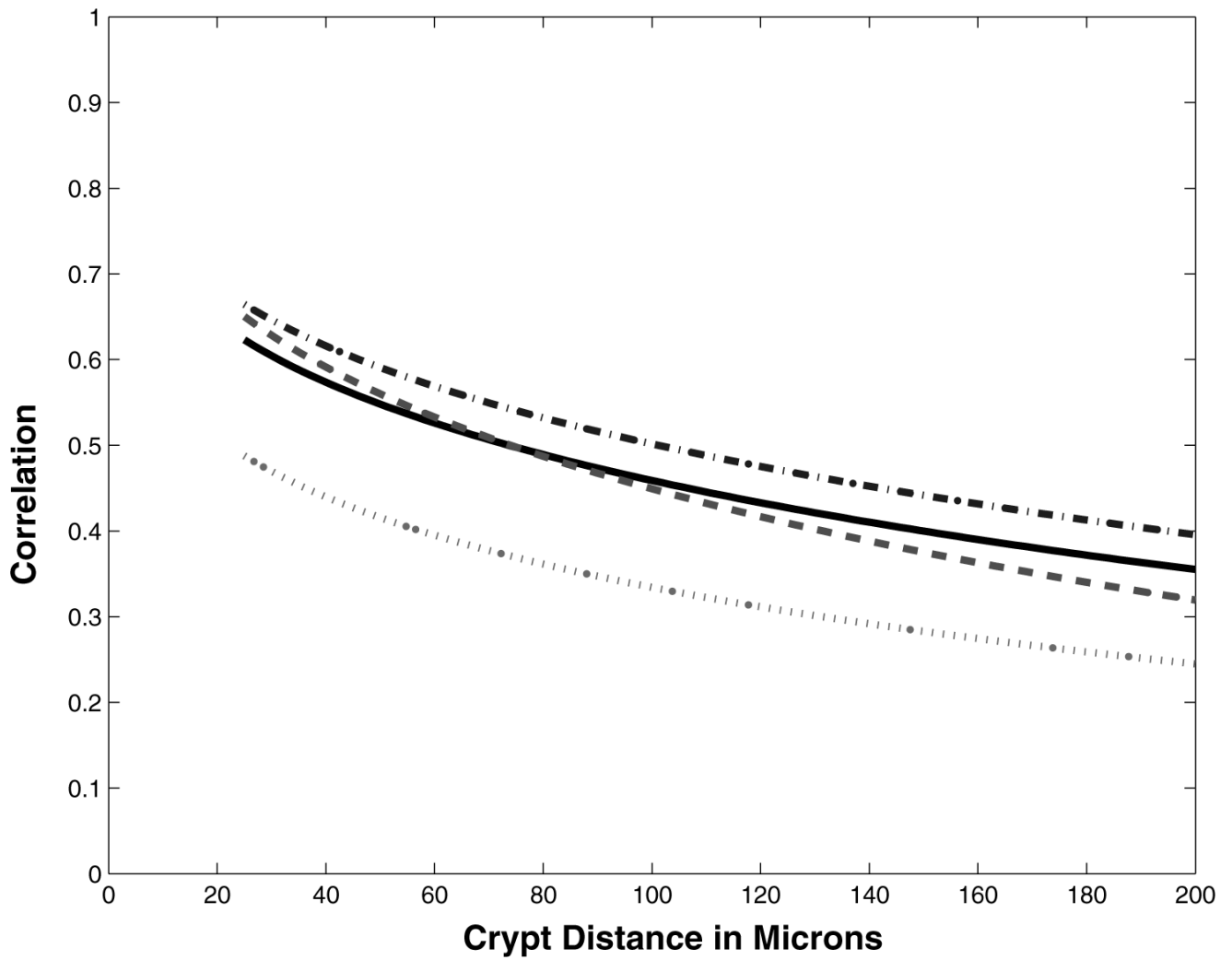
Figure 5.
Plot of correlation function estimates using pseudolikelihood analysis. All the data (solid line), using only the toptertile of the crypts (dashed line), the middle tertile (dash dotted) and bottom tertile (dotted line).